The Application of Genetic, *In Silico* and *In Vitro* Tools to Elucidate the Biology of a Functional Amyloid Fiber

by

Daniel Ray Smith

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Molecular, Cellular, and Developmental Biology)
in The University of Michigan
2010

Doctoral Committee:

Associate Professor Matthew R. Chapman, Chair
Emeritus Professor Robert A. Bender
Professor Michele S. Swanson
Associate Professor Kenneth M. Cadigan

To my wife, my late father and family

Acknowledgements

I would like to thank my wife Jennifer for her support during my graduate career at the University of Michigan. I have greatly enjoyed living in Ann Arbor and the great state of Michigan. The University of Michigan has proven to be an excellent place to learn and grow as a scientist due in no small part to the collegial atmosphere within the Molecular, Cellular, and Developmental Biology Department. In particular, the scientists within the Chapman and Bender labs have been instrumental to my success here. I would also like to thank the staff for all their help especially Mary Carr, Jesse Skinner, and Ed Grant.

My time as a member in the Chapman-Bender Empire has been a pleasant and rewarding experience. Past members including Ryan, Neal, Xuan, Elisabeth, and Michelle have been awesome to work with both on a professional and personal basis. Master students and undergraduates I have trained and who have contributed to some of the projects presented here include Soon Gang, Chris, Justin, Lo Hua, and Abhishek. As the last of the first bolus of graduate students in the Chapman lab, I am happy to see how the new members have changed both the scientific inquiries and personality of the lab while maintaining the rigorous thought and harmonious atmosphere.

I would like to thank my committee members for their many great ideas and encouragement. Bob Bender has been an exceptional co-mentor. I have enjoyed bouncing ideas off him and discussing bacterial cell physiology, the history of microbiology, and many intriguing tangents too numerous to lists. Through probing questions, Ken Cadigan has helped me to think about the impact and importance of my work. Michele Swanson has been instrumental in pushing me to develop the protein economy story towards fruition and increased exposure within the microbial science community. I could not have hoped for a better committee.

Lastly, I would like to thank my graduate mentor Matthew Chapman. Matt has had a positive impact on me as he has many others who have passed through his lab. Through some mysterious force Matt has pushed me as he has Neal and Xuan to expand how we think about and do science. I expect the new members will make great strides in becoming productive well-trained scientists.

Table of Contents

List of Figures

List of Tables

List of Appendices

ABSTRACT

The Application of Genetic, *In Silico* and *In Vitro* Tools

to Elucidate the Biology of a Functional Amyloid Fiber

by

Daniel Ray Smith

Chair: Matthew R. Chapman

Curli are thin aggregative fimbriae produced by many Enterobacteriaceae as a structural component of biofilms. Curli share many biochemical and biophysical properties with amyloid fibers which are often associated with human neurological diseases including Alzheimer's, mad cow, and Parkinson's. However, curli are the product of a dedicated assembly system that consists of a complex gene regulatory network featuring CsgD; a secretion system that includes CsgG, CsgE, and CsgF; and the major and minor fiber subunits CsgA and CsgB. As a model system, many aspects of curli formation have been explored including subunit secretion, regulation, biological function, and amyloidgenesis. My work focused on the genetics of curli formation in *Escherichia coli*, the economic constraints on the evolution of CsgA and other extracellular proteins, and

the *in vitro* amyloidgenesis of CsgA-His. I screened the Keio collection of single gene deletions to discover new genes that affect curli production. Over 300 genes modulate curli production including the sodium antiporter *nhaA*, a regulator of the glycine cleavage system *gcvA,* multiple LPS biosynthesis genes, and genes involved in many fundamental cellular processes. This analysis suggests that curli production is part of a highly regulated and complex developmental pathway. The regulation of *glyA* by CsgD and the curli phenotype of *gcvA* focused my attention on the amino acid composition of CsgA. CsgA is incredibly rich in glycine and serine. As simple amino acids, both are inexpensive to synthesize. Consequently, CsgA is relatively cheap to produce on a per unit basis. Strikingly, other extracellular proteins including those in *Escherichia coli, Pseudomonas syringae, Mycobacterium tuberculosis, Saccharomyces cerevisiae*, and other microbes are also inexpensive relative to intracellular proteins. Since extracellular proteins are often lost to the environment, evolution has in turn selected them for increased economy to counteract lost resources. Finally, we studied the *in vitro* amyloid formation of CsgA-His. Like disease-associated amyloids, CsgA-His bound Thioflavin T upon polymerizing into fibers, reacted with an amyloid specific antibody, self seeded, and displayed other aspects of amyloid formation. Collectively, this work sheds new light on the biology of the functional amyloid fiber curli and hopefully will beget novel directions of inquiry.

Chapter I

Introduction

Curli are thin aggregative fimbriae produced by many Enterobacteriaceae (Fig. 1.1A) (23, 140, 193). As an integral structural component of biofilms (41, 87, 145, 152, 174), curli help bacteria persist in the face of environmental insults such as desiccation and oxidation (144, 167, 179, 181, 182). Curli have been implicated in pathogenesis and aid in adherence, internalization, and persistence (60, 61, 80, 89, 94, 95, 100, 146, 161, 168, 169, 179). Curli fibers mediate binding to a variety of host factors including fibronectin and MHC class I molecules (16, 38, 72, 120-123, 131, 151) and are potent inducers of the host inflammatory response (18, 20, 121, 164-166, 179). As a public health and food safety issue, curliated bacteria better adhere to plant tissues (10, 11, 26, 79, 88, 99, 163) and a variety of inert surfaces used in the food processing industry including glass, plastics, and steel (7, 15, 23, 30, 41, 128, 133, 162, 167).

Curli fibers are the product of a dedicated and highly regulated assembly system encoded by curli specific genes[1] (12), which are arranged in two divergent operons *csgDEFG* and *csgBAC* (63, 141). Found primarily in the genomes of Enterobacteriaceae, *csg* genes are also present in different arrangements and combinations in other Gammaproteobacteria families including Pseudomonadaceae, Shewanellaceae, and Vibrionaceae as well as in a few distantly related bacteria in the Bradyrhizobiaceae, Burkholderiaceae, and Flavobacteriaceae families (78, 160) (See Fig 2.11).

Much of the complex control of curli formation involves regulation of the *csgDEFG* operon (12); CsgD is a transcriptional regulator that is required for transcription of the *csgBAC* operon (63, 142, 192). Transcription of the *csgDEFG* operon is under the control

[1] Originally *csg* referred to curli subunit genes (6).

of RpoS (5, 120), Crl (6, 24, 132), MlrA (29), H-NS (5, 56, 120), IHF (56), and a number of bacterial two component systems (53, 75, 77, 117, 118, 141, 173, 174). All the curli specific genes except CsgD have a Sec secretion signal, which is necessary for their translocation across the inner membrane into the periplasmic space (12, 83). The accessory proteins CsgE, CsgF, and CsgG are all required for proper curli formation (Fig 1.2B) (31). CsgG is an outer membrane lipoprotein, that is thought to form a pore by which the major and minor curli subunits CsgA and CsgB are secreted (31, 50, 105, 136). The functions of CsgE, CsgF, and CsgC are currently not as well characterized (12, 31, 58). Both CsgE and CsgF interact with CsgG and may aid in its proper function (50, 136). CsgF is also important for the association of CsgB with the bacterial outer membrane (66, 116) where CsgB nucleates CsgA into the fiber known as curli (19, 64, 66). CsgC is dispensable for curli formation (See Fig 1.2B); however, it may influence curli fiber morphology (58, 63). See Fig. 1.2A for a schematic model of curli biogenesis.

Intriguingly, curli fibers have many of the properties of amyloids which are often associated with neurological disorders including Alzheimer's, mad cow, and Parkinson's diseases (31, 66, 177, 178, 180). Curli fibers produced by *Escherichia coli* K-12 strains appear as a mesh like matrix of thin fibers when viewed by transmission electron microscopy (TEM) (Fig. 1.1A). When purified the major curli subunit CsgA forms similar fiber structures (Fig. 1.1B). These CsgA-His fibers fluoresce when bound to Thioflavin T (ThT) (Fig. 1.3A) (85, 93, 101-103), have circular dichroism (CD) spectra indicative of high beta sheet content (Fig. 1.3B), and produce a red shift when bound to the amyloid specific dye Congo Red (CR) (31, 86, 91, 92, 143, 158, 159, 180). Using ThT fluorescence we have shown that CsgA-His polymerizes with a defined lag, growth and stationary phase, and that CsgA-His is capable of self seeding with preformed fibers (Fig 1.4) (180). Both properties are indicative of nucleation dependent polymerization. Like many amyloid fibers, curli are also highly stable (39, 40). Curli fibers resist proteinase K digestion, are insoluble in SDS, and require harsh denaturants to separate and mobilize into SDS polyacrylamide gels (Fig. 1.5C) (39, 81, 122).

Under denaturing conditions, many proteins form amyloids, suggesting that the amyloid fold results from backbone interactions and may be an intrinsic property of proteins (9, 34, 125). However, in many cases, amyloid formation leads to disease, the result of an aberrantly folded protein owing to either destabilizing point mutations or defects in the cell's folding, secretion, or turnover machinery (31, 51, 65, 66, 125, 149, 177, 178, 180). However, unlike these disease-associated amyloids, curli are not the result of an aberrant fold. Instead, curli have evolved to fulfill a functional role and require a complex assembly system (see above). As such, curli are part of a growing list of functional amyloids which are found in many organisms where they perform cellular tasks including regulation, storage, and structural roles (2, 9, 21, 31, 35, 36, 42-44, 49, 54, 55, 67, 108, 109, 119, 124, 125, 139, 150, 177, 183-185). Many of these functional amyloids share little sequence similarity, suggesting the amyloid fold has evolved multiple times to fulfill important physiological functions.

Curli as a model system

Though often imperfect (25), model systems are useful in biology to generate new hypotheses. These hypotheses often regard how components interact with one another and function together as a whole. By selecting model systems that are easily manipulated and well understood, one can generate concepts that apply to similar systems and improve understanding of complex biological processes. The key to a good model is the ease of study and the breadth of the science a model allows you to explore.

In both respects, Curli are an excellent model system, being amenable to study and applicable to diverse fields of research. Several properties of curli including the amyloid formation of CsgA and the genetic tractability of *Escherichia coli* allow for the use of multitude tools including genetic, *in vitro* and *in silico* techniques. Curli have been used to explore biofilm formation, pathogenesis, the assembly and secretion of fimbriae, and amyloidogenesis. This work focuses on the genetics (Chapter II), molecular evolution (Chapter III), and *in vitro* amyloid formation (Chapter IV) of curli fibers.

Genetics

For any curli defective mutant there are several techniques to discover how curli assembly is affected including Congo Red (CR) binding, inter-bacterial complementation, Western blotting, biofilm and pellicle formation, TEM, and transcriptional assays. Of these techniques CR binding is by far the most important. When curli producing bacteria are grown on agar plates supplemented with Congo Red they turn bright red (Fig. 1.5B), while strains that lack curli remain white (Fig. 1.2B and 1.5A). The Congo Red phenotype can vary from white to dark red and is fairly indicative of the amount of CsgA produced by different strains of bacteria (Compare Fig 1.5A and 1.5D). Thus CR allows for complementation and genetic screens for accessory factors, greatly facilitating the study of curli producing bacteria.

Utilizing this genetic tool, I screened the Keio collection of single mutants (8) for new genes that affect curli biogenesis. More than 300 genes were found to affect curli production in *E. coli* K-12. These genes include transport proteins, lipopolysaccharide (LPS) and purine biosynthetic genes, regulatory proteins, and many additional genes involved in fundamental cellular processes. Many of these genes affect RpoS levels or induce the sigma E stress response which promote and inhibit curli production, respectively. Few of the genes identified are involved in motility, underlying that motility and biofilm formation are distinct microbial lifestyles. In chapter III, I further explore how these results indicate curli production is part of a highly regulated and complex developmental pathway in *Escherichia coli*.

In silico

Curli are unique from a bioinformatics perspective due to their unique composition and extracellular location. I noticed the connection between these factors while pondering the *gcvA* mutant, which produces less curli (Fig 1.6B). GcvA regulates the glycine cleavage system which degrades glycine into ammonia and carbon dioxide (57, 69, 82, 154-157, 186-189). CsgD, the master regulator of curli, is known to increase *glyA* levels (32, 33); GlyA is serine hydroxymethyltransferase which interconverts glycine

and serine (Fig. 1.6C) (83). The regulation of *glyA* by CsgD and the *gcvA* CR phenotype indicated these amino acids were import for curli production. Strikingly, CsgA is incredibly glycine and serine rich, and the regulation of *glyA* by CsgD may help balance these amino acids for efficient curli production.

More importantly, as simple amino acids, glycine and serine are relatively inexpensive to synthesize (1, 45, 70). Consequently, on a per unit basis CsgA is an inexpensive protein to produce. I extended this analysis of protein economy to other extracellular proteins. Many extracellular proteins including those in *Escherichia coli* (Fig. 1.7)*, Pseudomonas syringae, Mycobacterium tuberculosis, Saccharomyces cerevisiae*, and multiple other microbes were found to contain, on average, less expensive amino acids. Furthermore, this effect seems to be independent of the abundance (Fig. 1.8), structure, function, or length of extracellular proteins. In Chapter III, I explore how economy is a selective pressure on the composition of extracellular proteins and discuss possible biological reasons for this selection.

In vitro

Curli were identified as an amyloid fiber by Matthew Chapman (31). One of the original projects in the lab was purifying the major curli subunit CsgA. Following this development, Xuan and I studied the *in vitro* polymerization of CsgA-His through Thioflavin T fluorescence, circular dichroism, TEM, and A11 antibody binding. The A11 antibody recognizes an intermediate in the polymerization of many disease associated amyloids (84). These techniques allowed us to study the kinetics, seeding, and factors favoring curli fiber formation. In Chapter IV, I describe aspects of the *in vitro* polymerization of CsgA-His and contrast them with disease-associated amyloids.

Economics plays an important role in evolution

The economic selection on extracellular proteins is part of a broader concept in molecular evolution. That is the evolution of proteins is not driven purely by selection for increased function. Another factor driving the composition and regulation of

proteins is the relative cost of their components. This concept of cost selection grew out of the heated debate over neutral evolution. Neutral evolution was originally postulated by Kimura (1968) and gained wider recognition following the publication of "Non-Darwinian Evolution" by King and Jukes in 1969 (90). The theory of neutral evolution states that many if not most of mutational changes in organisms do not affect function, that is mutations are frequently neutral to natural section and generally reflect genetic drift instead of positive selection for a particular trait. An early reply to "Non-Darwinian Evolution" by R.C. Richmond (135) stated that many neutral mutations are not necessarily so; one quotation of this critique is relevant here as it was elsewhere (1).

> Even if a substituted amino acid were truly neutral in a functional sense, it is highly unlikely that factors involved in the synthesis of the protein would render the substitution neutral in the broader sense of the organism's integrated functioning. The substituted amino acid must be present within the cell in equivalent quantity compared with the original amino acid and, indeed, its synthesis or derivation from other molecules and transport into the cell require an equivalent amount of energy output.
>
> R.C. Richmond "Non-Darwinian Evolution: a critique" Nature 1970

There are now numerous instances where protein composition and regulation reflects limited environmental resources. Much of the economical bias present in biological molecules appears to reflect global or overall reductions and substitutions of rare or expensive components. For example, Akashi and Gojobori (1) showed that many abundant proteins have amino acids that cost less for the cell to produce. In ocean environments, phytoplankton can overcome phosphorous limitation by substitution with sulfur and nitrogen containing lipids (171). Similarly, species of *Prochlorococcus* isolated at deeper depths have more nitrogen and sulfur in their proteomes than surface species, reflecting the content of these elements in the water column (106). By reducing the use of scarce environmental resources, organisms can decrease the cost of their metabolism and replication, thus increasing their relative fitness during times of resource limitation.

Another form of economic selection is more focused and reflects what happens during the sudden loss of a needed metabolite. When an organism is suddenly without

specific metabolite, it often responds by producing the proteins that transport or synthesize this metabolite. However, if proteins that transport or produce a metabolite contain a higher composition of the corresponding metabolite, the cell finds itself in a classic Catch-22 situation. Such transient starvation should exert a strong and specific pressure to reduce the composition of the scarce metabolite in the proteins that transport or synthesize it. Organisms that have increased economy of these 'stress' proteins should be better able to respond to transient starvation and thus be more fit. One of the best descriptions this kind of optimization was proposed in the economy of amino acid biosynthetic enzymes. Many amino acid biosynthetic proteins have reduced contents of the amino acid they ultimately produce (3, 130). Thus when a specific amino acid is needed the cell is better able to increase production of the proteins needed for an efficient response. Other instances of this form of economy include regulation of alternative enzymes during starvation that have less of the limiting resource.

Cost values for cellular components

Before delving into how economy affects evolution, one must consider the different cost values for proteins and other cellular components and their associated strengths and weaknesses. The amino acids of proteins have diverse physicochemical characteristics which affect protein folding and function. The diversity of amino acid chemistry requires different biochemical pathways and chemical precursors. These pathways and precursors result in some amino acids being more expensive to synthesize than others. The relative cost of amino acid synthesis can be calculated in many ways including the ATPs required to make an amino acid, the mass of an amino acid, the number of carbons, nitrogens, or sulfurs in an amino acid, or the Gibbs free energy of formation (4).

Craig and Weber (45) first calculated the relative synthetic costs of amino acids of *E. coli*. Using metabolic information, each amino acid cost was calculated by adding the ATPs equivalents consumed in synthesis with the energy lost from diverting precursor from central metabolism. Akashi and Gojobori (1) updated these values by

averaging out the slight differences in cost incurred during growth on different carbon sources. This bacterial cost has been extended to other chemoheterotrophic bacteria because there is conservation in the central metabolism of many aerobic bacteria (1, 70). Similar calculations have been done for yeast under both respiratory and fermentative conditions (28, 176), resulting in numbers that differ slightly but are highly correlated with the bacterial values (Fig. 1.9E).

Mass has also been proposed as a complementary approach to calculate relative costs of amino acid (148). The synthetic cost, mass, and Gibbs free energy (4) of amino acids can be used to compare the relative cost of proteins because they convert protein components to a common currency. All three cost schemes correlate well with one another; however, each emphasizes certain functional groups over others (Fig. 1.9A, 1.9F, 1.10D). Differences in the atomic composition and chemical nature of amino acid precursors result in a slight shuffling in the order of amino acid cost depending on which system is used. For example, the acidic and nitrogen rich amino acids tend to more massive relative to their synthetic cost (Fig. 1.9A).

A second type of cost is based on the atomic makeup of the amino acids and includes the number of carbons, nitrogens, or sulfurs in each amino acid. Oxygen is excluded from economic consideration. Oxygen is usually fixed in amino acids through water, and there is little evidence, even in anaerobes, of oxygen limitation affecting amino acid composition (175). The remaining elements are often scarce in the environment and their economy may be as important if not more so than energy concerns. Of these elements, only carbon correlates well with synthetic cost, mass, or Gibbs free energy. The vast majority of synthetic cost is due the number of reduced carbons diverted from central metabolism (1, 45, 176).

There are many other costs for proteins and other cellular components. Trace metals substitutions in enzymes can be an important cost during metal starvation. The amount of DNA an organisms has increases its need for phosphorous, a limiting element in the surface of the ocean (37). Additionally, each GC pair in DNA requires one more

nitrogen and one less carbon than an AT pair, and GC pairs are more expensive to produce (137). Lipids also display compositional biases that reflect environmental limitation (171, 172). Unfortunately, there is no single best way to quantify the relative cost of cellular components due to the diversity of metabolism and environmental conditions (13). Additionally, what is expensive for one organism or group of proteins may not necessarily be so for another (3, 48, 70, 130). However, the use of appropriate costs in a relevant biological context can elucidate how economy has affected cellular composition.

Synthetic cost and carbon

Abundant proteins were first shown by Akashi and Gojobori to use less expensive amino acids in their synthesis in *E. coli* and *B. subtilis* (1). More extensive studies in other bacteria found similar results whether using synthetic cost or amino acid mass (70, 148). Eukaryotes such as yeast also have less expensive and massive amino acids in their abundant proteins (104, 134, 148). Abundant proteins have also been shown to have less carbon, nitrogen, and sulfur contents (104). There are notable exceptions to this rule for abundant proteins. Ribosomal proteins in particular do not fit well (Fig. 1.8), and their removal improves these comparisons; their preference for basic amino acids results in high nitrogen contents (104).

Carbon is intrinsically linked to synthetic cost as much of this cost is due to the number of reduced carbons. Therefore, it is not surprising that inexpensive proteins have lower carbon contents (See Fig. 3.12). Attempts to separate carbon content and synthetic cost have suggested carbon limitation is the driving force in the economy of abundant proteins; however, these methods involved removing nitrogen and sulfur amino acids including tryptophan which can significantly affect protein cost (104). Reductions in carbon content are also present in proteins that assimilate carbon, which are preferentially expressed during carbon starvation (14, 28).

Nitrogen, sulfur, and phosphorous

Nitrogen has been implicated in the economical selection of DNA and proteins. *Pelagibacter ubique* is found near the ocean surface where nitrogen is limited. Despite the retention of DNA repair enzymes, the genome of *P. ubique* is quite AT rich, which reduces nitrogen requirements relative to GC rich DNA (59). The limiting nitrogen content in the ocean was also suggested to pressure the proteomes of *Prochlococcus* species. *Prochlococcus* isolates obtained nearer to the ocean surface had lower nitrogen and sulfur contents (106). Plants also have less nitrogen in their abundant proteins; however, animals do not have a consistent pattern between nitrogen content and protein abundance (48)

Sulfur limitation in a sulfur assimilation protein was noted prior to the theory of neutral evolution (127). Reductions in sulfur content in other sulfur metabolism genes were subsequently seen in bacteria and yeast (14), and this depletion is also present in abundant proteins (104, 111). During sulfur starvation, sulfur depleted enzymes are preferentially expressed in bacteria and yeast (22, 46, 52). However, sulfur reduction can also have a strong functional basis which can be conflated with economic selection. Both cysteine and methionine are readily oxidized (68, 153, 190); subsequently, compartmentalization (138) and reduction of these amino acids occurs to prevent loss of protein function (27, 175).

Phosphorous is primarily used in lipids and nucleotides and is quite limiting in the ocean (37, 114). Consequently, the lipids and nucleotides of some ocean microbes appear to reflect phosphorous economy. For example, *Prochlorococcus* and *P. ubique* have greatly reduced genomes (37, 59), which reduces the overall phosphorous requirements. *P. ubique* currently has the smallest genome among free living bacteria, and the average intergenic region between its genes is only three bases (59). However, despite the genome reduction found in *Prochlorococcus* strains, nucleotides still account for about half the cellular phosphorous (17, 37). Ocean microbes can also overcome phosphorous limitation by using lipids that contain sulfur and nitrogen (171, 172).

Metals

Ocean microbes also have evolved ways to deal with limiting trace metals needed for metabolic enzymes. Many microbes can replace scarce metals such as Zinc with Cobalt or Cadmium (97, 114, 115). Several enzymes have been identified that explain this exchange of metal cofactors including carbonic anhydrase (96) and superoxide mutase (126). *Thalassiosira weissflogii* expresses an alternative cadmium containing carbonic anhydrase when zinc is limiting (96, 98, 191). Similarly, zinc limitation increases expression of a zinc independent form of GTP cyclohydrolase, which can uses other divalent cations (147). Iron limitation leads to higher manganese content in superoxide dimutase (129). Species of *Azotobacter* express a vanadium containing nitrogenase under molybdenum deficiency (47, 62). In several of the instances above the limitation of one metal is linked to production of isozymes that uses alternative cofactors which helps the cell overcome a metabolic block. Metal limitation can also change basic metabolism to reduce the need for a scarce resource. In *Euryarchaeon halobacterium,* low iron changes respiratory pathways and reduces the number of iron sulfur clusters (73).

There are likely many additional examples of economic selection of cellular components or regulation of alternative isozymes during starvation. Several such examples are listed above along with descriptions of general trends in protein economy. Collectively, I believe these examples demonstrate an invaluable biological concept, that cost is an important pressure on the composition of life. The exact nature and strength of the selection for economy in many of these examples requires further study. Either positive selection for increased economy or negative selection of expensive substitutions would produce similar results. An early paper suggest protein economy result more from the latter; that is selection is against the substitutions of expensive amino acids (76). However, either selection allows life to flourish despite sparse resources.

Figure Legends

Fig. 1.1. TEM of curli fibers produced by *Escherichia coli* K-12 and purified CsgA-His which has polymerized into amyloid fibers. (A) A BW25113 strain of *E. coli* K-12 produced curli fibers after being grown on YESCA plates for two days at 26C. Magnification is 15,500X. (B) Purified CsgA-His allowed to polymerize at room temperature for 7 days. Bar indicates 500nm. Published image (180). A Philips CM12 Scanning Transmission Electron Microscope was used to visualize the fiber aggregates. Samples (10 μl) were placed on formvar coated copper grids (Ernest F. Fullam, Inc.), washed with deionized water, and negatively stained with 2% uranyl acetate.

Fig. 1.2. Model of curli biogenesis and *in vivo* phenotypes of *csg* single gene deletions. (A) Curli is the product of the transcription of two operons *csgDEFG* and *csgBAC*. CsgD is required for transcription of the *csgBAC* operon. Phosphorylation of the Asp 59 of CsgD reduces its ability to bind to the *csgBAC* promoter region (192). All the curli specific genes except CsgD have a Sec secretion signal resulting in the translocation across the inner membrane into the periplasmic space. CsgG is transported by the LOL system to the outer membrane where it is thought to form a pore by which the major curli subunit CsgA and the minor subunit CsgB are secreted. CsgE aids in the function of CsgG, but may have other functions. CsgF helps tether CsgB to the outer membrane where CsgB nucleates CsgA transforming it into the majority of the amyloid fiber known as curli. (B) Congo red binding and (C) western blots for CsgG and CsgA of WT (BW25113) *E. coli* K-12 and single mutant from the Keio collection (8). Strains were grown on YESCA plates for two days at 26C.

Fig. 1.3. Fibers formed from CsgA-His display amyloid properties including increased ThT fluorescence, CD spectra indicative of high β-sheet content, and a red shift when bound by Congo Red. (A) CsgA-his or Bovine Serum Albumin were buffered in 50mM KPi pH 7.2. Thioflavin T (ThT) was added to 2 μM. Fluorescence emission was read on a Spectramax M2 plate reader (Molecular Devices) with excitation set to 438 nm and the

cutoff frequency to 465 nm. Chemical structure of Thioflavin T is inset. (B) Circular dichroism (CD) spectra of purified CsgA-His (10 μM in 50mM KPi) incubated at room temperature for the times indicated. Spectra were performed in a Jasco J-810 spectropolarimeter from 190 to 250nm in a quartz cell with 1-mm path length at 25°C. (C) Absorbance of Buffer (50mM KPi), polymerized CsgA-His, and Bovine Serum Albumin (BSA) with the addition of Congo Red (CR). (D) Difference spectra of BSA and polymerized CsgA-His where the absorbance of Buffer + CR was subtracted from the respective protein + CR spectra. The maxima of the red shift for CsgA-His with CR was 544 nm.

Fig. 1.4. CsgA polymerization kinetics and self seeding using ThT fluorescence. (A) Thioflavin T time course of CsgA-His polymerization showing the typical lag, log and stationary phase (after green arrow) of amyloid fiber growth. CsgA-His was purified as described (180) mixed with ThT and allowed to polymerize in a 96 well plate (Costar #3605) and be read by a Spectramax M2 plate reader (Molecular Devices) which was set to 438 nm excitation 495 nm Emission and 475 nm Cutoff frequencies. The lag time (blue arrow) is set as the time between starting the reaction and the point at which the two slopes indicated intersect on the X axis (time). The half time (t ½) (black arrow), which is affected by both the lag phase and polymerization rate, is the time point half way through the growth phase of polymerization. (B) Addition of preformed CsgA-His seeds (sonicated) reduces the lag phase of unpolymerized CsgA-His.

Fig. 1.5. Curli fibers bind Congo Red (CR) and are SDS insoluble. (A) Congo Red binding phenotypes of different strains of *E. coli* K-12 from the Keio collection (8). Phenotypes range from white to light pink and correlate well with western blots for CsgG and CsgA (See Fig. 1.5D). (B) Chemical structure for Congo Red. (C) Curli fibers from a BW25113 wild type strain of *E. coli* K-12 are SDS insoluble unless harsh denaturants such as Formic Acid (FA) or Hexafluoroisoproponol (HFIP). (D) Western blot of whole cells of the strains seen in Fig. 1.5A. Blots in Fig. 1.5C,D were probed antibodies to CsgA and CsgG.

Fig. 1.6. Use of Congo Red binding in the Keio collection screen found many mutants including *gcvA* were defective for curli formation leading to an exploration of glycine composition in curli specific genes including CsgA. (A) Typical Congo Red (CR) YESCA plate from using a plate replicator to transfer Keio collection strains to screen for new genes that affect curli production in *E. coli* K-12. Strains were grown two days at 26C. (B) Comparison of CR phenotype of WT (BW25113), *csgA,* and *gcvA* strains of *E. coli* K-12. Strains were grown two days at 26C on CR YESCA plates. (C) Diagram of serine and glycine interconversion by GlyA and glycine breakdown by the glycine cleavage system. (D) Histogram of 4033 proteins in *E. coli* MG1655 showing their percentage of glycine plus serine. The mean % of glycine plus serine is 12.9 %; CsgA is 28.5 % glycine plus serine.

Fig. 1.7. Economy of extracellular proteins in *Escherichia coli* strain MG1655. Extracellular proteins contain on average less expensive amino acids than cellular proteins. For each protein in *E. coli* strain MG1655, the average synthetic cost or ASC (∑(synthetic cost of amino acid in ATPS)/number of amino acids) was calculated using the values for chemoheterotrophic bacteria (1, 70). Using the average synthetic cost and the location data from Echobase's Echolocation website (113) available at (http://www.york.ac.uk/res/thomas/echolocadv.cfm) each protein was graphed accordingly; within each location the data were spread vertically using protein length (0.2×LOG(Length)). The dotted black line represents the mean ASC of all the proteins used (23.55 ATPs/aa). The thick solid black line indicates the mean ASC of extracellular proteins (20.63 ATP/aa). FliC and CsgA are indicated (black arrows). The difference in ASC between extracellular and cellular proteins is significant (Mann Whitney U-test, $p=1.24\times10^{-10}$). A few locations were changed based on the literature: FlgM (74), FliK (112), and Eno (170) to extracellular, FlgJ to periplasmic (107), and Ag43 was divided into its secreted soluble alpha domain and outer membrane beta domain (71) (both α-Ag43 and β-Ag43 are shown in blue and indicated with arrows).

Fig. 1.8. Comparison of the effects of location and abundance on protein ASC. Abundance does not appear to be a good indication of protein economy in *E. coli* K-12. Proteins from *E. coli* MG1655 were plotted according to their average synthetic cost (ASC) and either their abundance if known (darker blue) (110) or their total length (lighter blue). Both Y scales are logarithmic. Extracellular proteins are indicated in green; ribosomal proteins (Rpl, Rpm, Rps, and Sra) are indicated in red. The dotted black line represents the mean ASC of all the proteins used. See Fig. 1.7 legend for changes to location and Table 3.13 in Chapter III for Notes on the merging of Echobase sequences (113) with the abundance data of Matsuda et al 2009 (110).

Fig. 1.9. Comparison of the synthetic cost of amino acids in bacteria with their atomic mass; number of carbons, nitrogens, or sulfurs; yeast synthetic costs, and Gibbs free energy of formation. The synthetic cost of for bacteria is for chemoheterotrophs (1, 70). Yeast respiratory and fermentative growth synthetic cost values are from Wagner (176). Values for Gibbs free energy of formation of amino acids in sea water (18°C) and hydrothermal solutions (100°C) are from Amend and Shock (4). (A) Comparison of bacterial synthetic cost and mean amino acid mass. Comparison of bacterial amino acid synthetic cost and the numbers of Carbons (B), Nitrogens (C), or Sulfurs (D) per amino acid. (E) Comparison of amino acid synthetic cost in bacteria (aerobic growth) with yeast in respiratory (green) and fermentative (blue) conditions. (F) Comparison of bacteria amino acid synthetic cost with the Gibbs free energy of formation at 18°C in seawater (orange) and 100°C in hydrothermal solution (purple).

Fig. 1.10. Comparison of mass of amino acids to their number of carbons, nitrogens, or sulfurs and Gibbs free energy of formation. Values for Gibbs free energy of formation of amino acids are from Amend and Shock (4). Comparison of amino acid mass with their numbers of Carbons (A), Nitrogens (B), or Sulfurs (C). (D) Comparison of amino acid mass with their Gibbs free energy of formation at 18°C in seawater (orange) and 100°C in hydrothermal solution (purple).

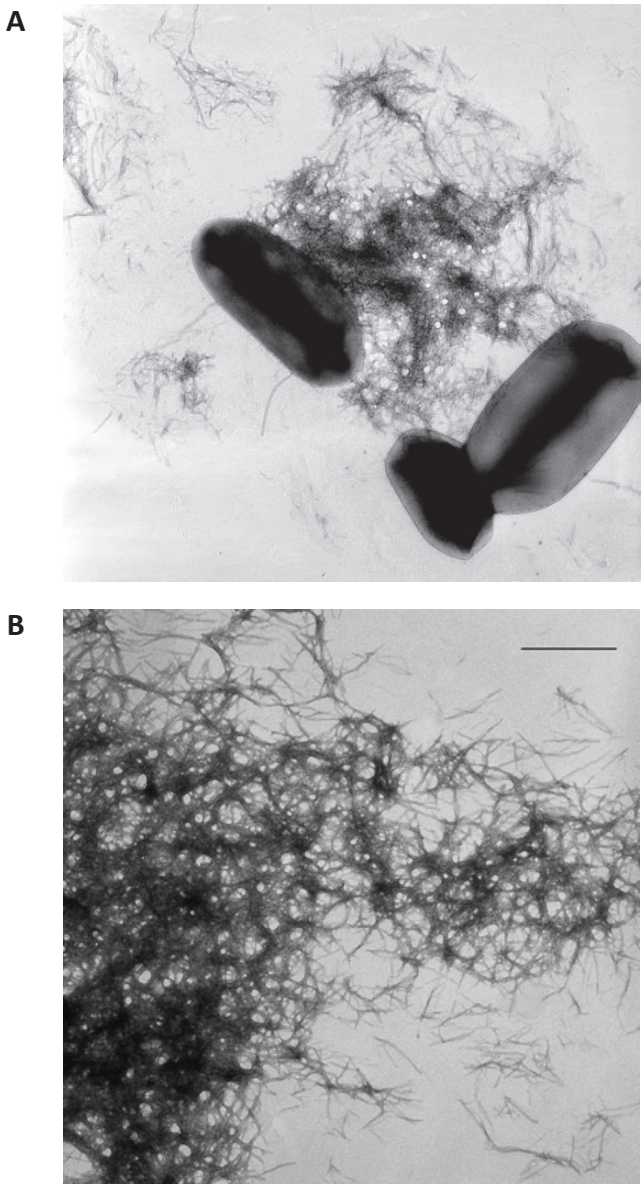


Figure 1.1. TEM of curli fibers produced by Escherichia coli K-12 and purified CsgA-His which has polymerized into amyloid fibers.

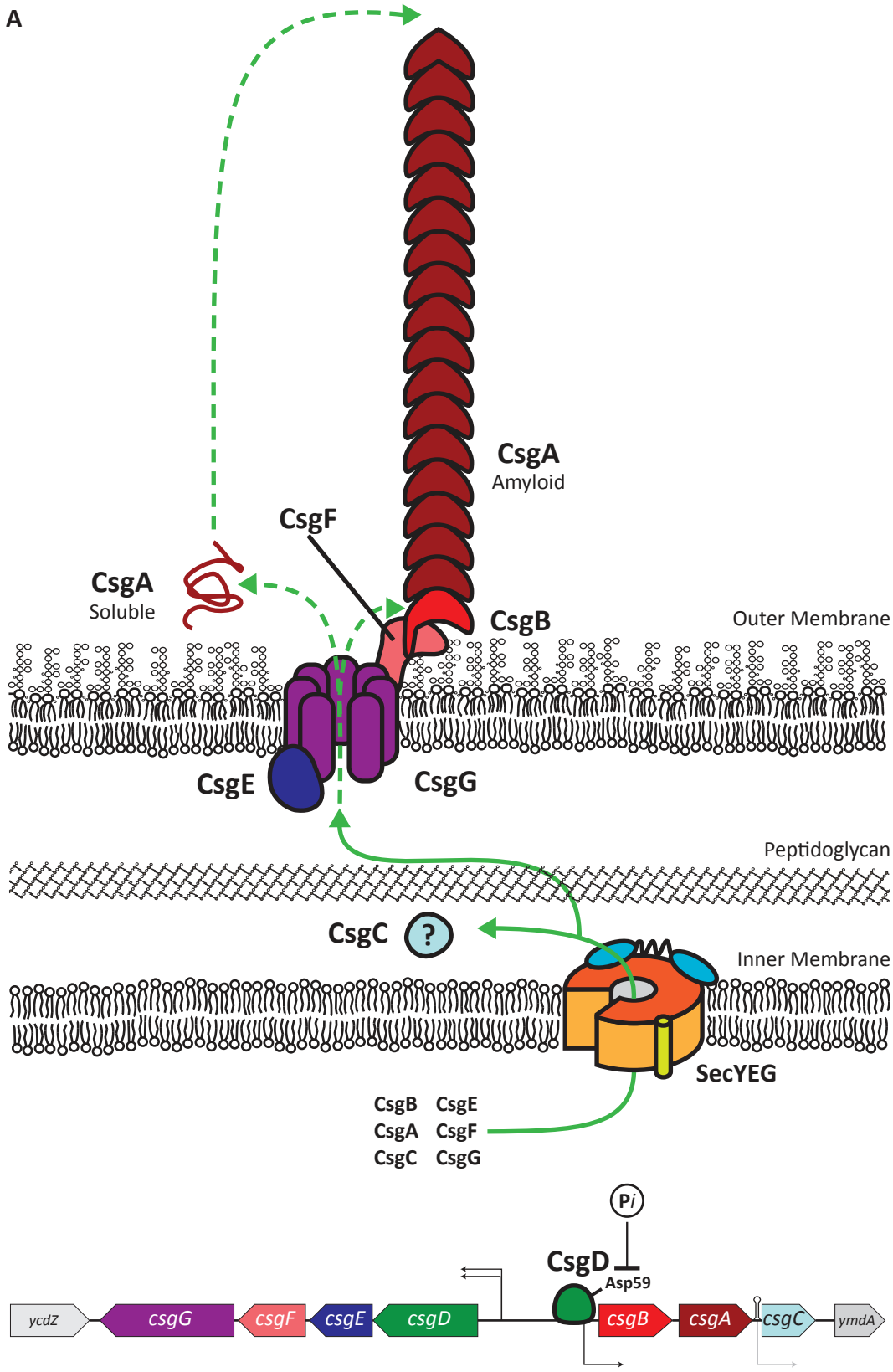
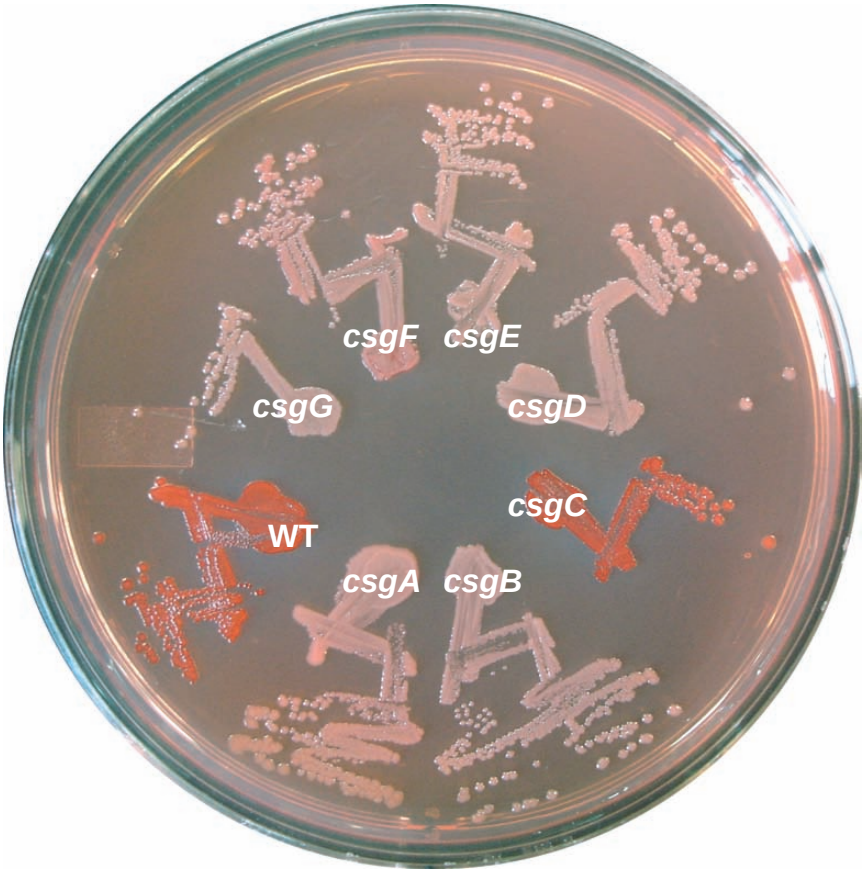


Figure 1.2. Model of Curli biogenesis and *in vivo* phenotypes of *csg* single gene deletions.

B



C

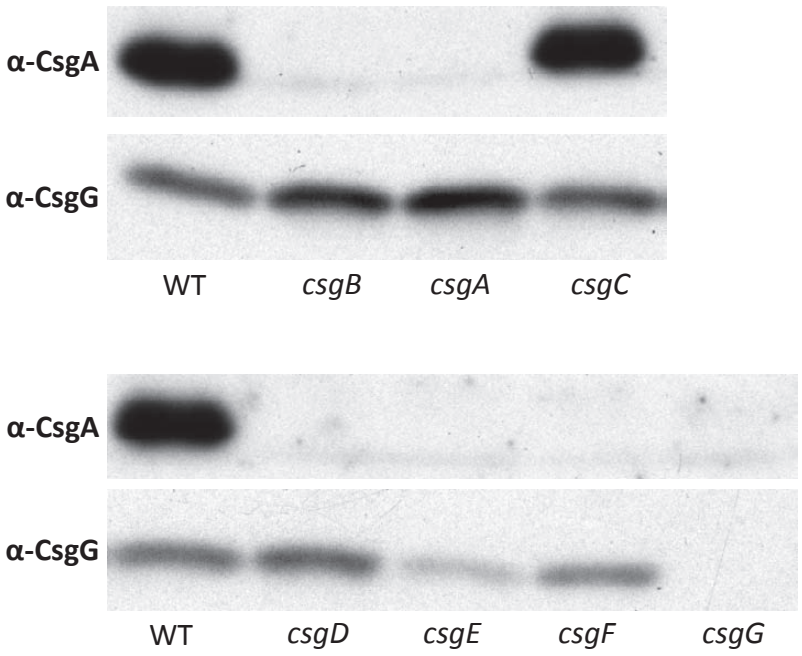


Figure 1.2 (continued). Model of Curli biogenesis and *in vivo* phenotypes of *csg* single gene deletions.

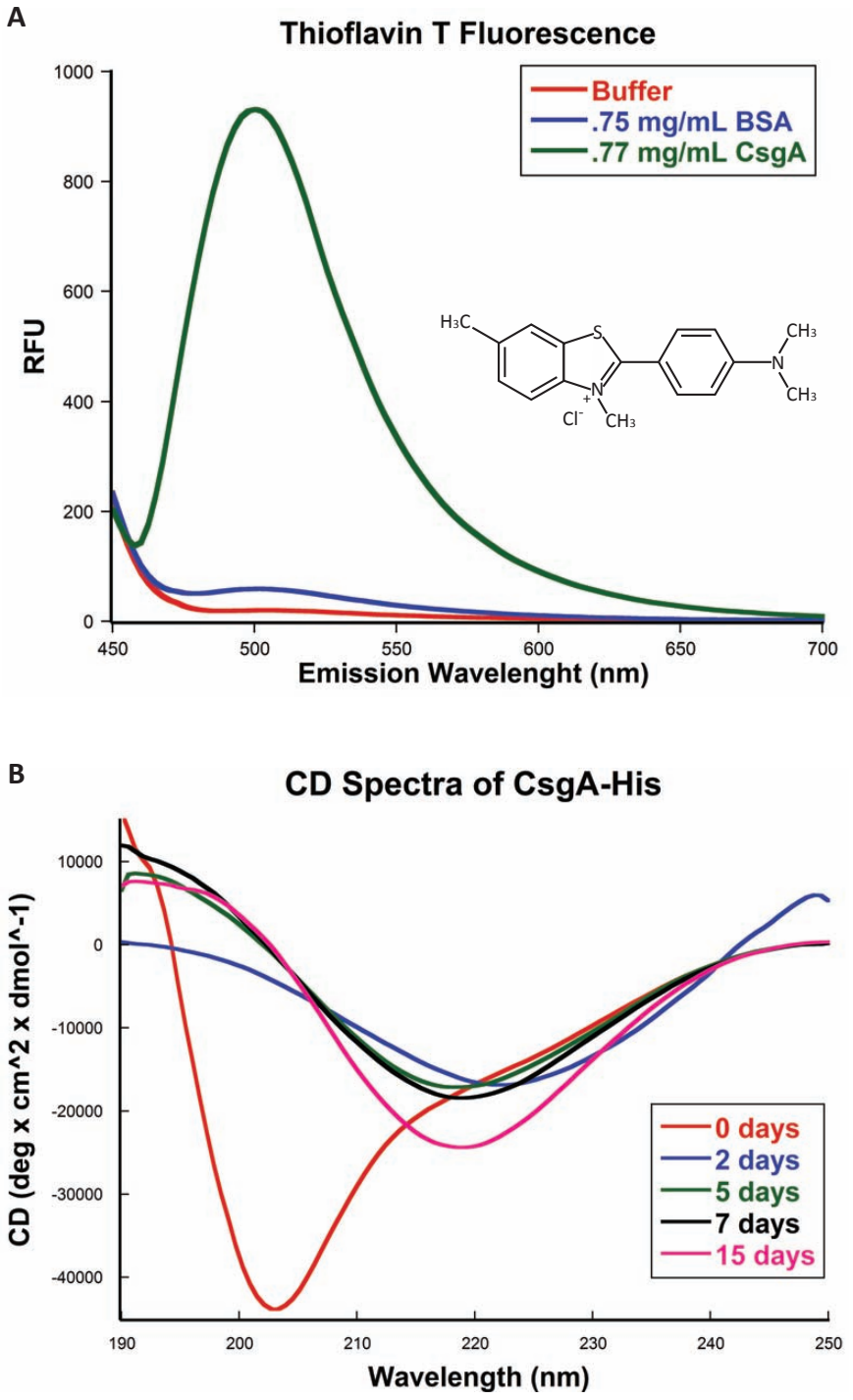


Figure 1.3. Fibers formed from CsgA-His display amyloid properties including increased ThT fluorescence, CD spectra indicative of high β-sheet content, and a red shift when bound by Congo Red.

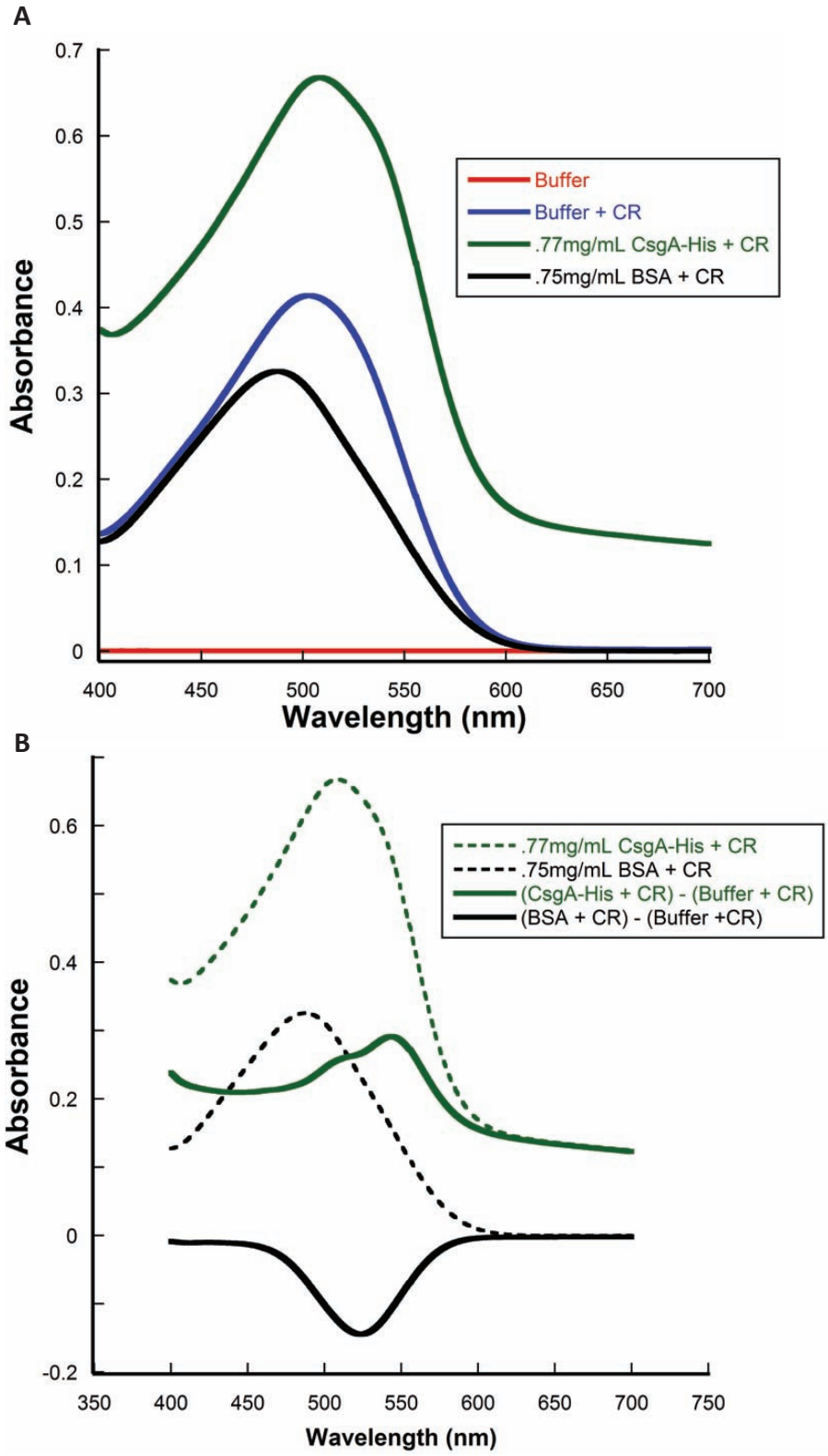


Figure 1.3 (continued). Fibers formed from CsgA-His display amyloid properties including increased ThT fluorescence, CD spectra indicative of high β-sheet content, and a red shift when bound by Congo Red.

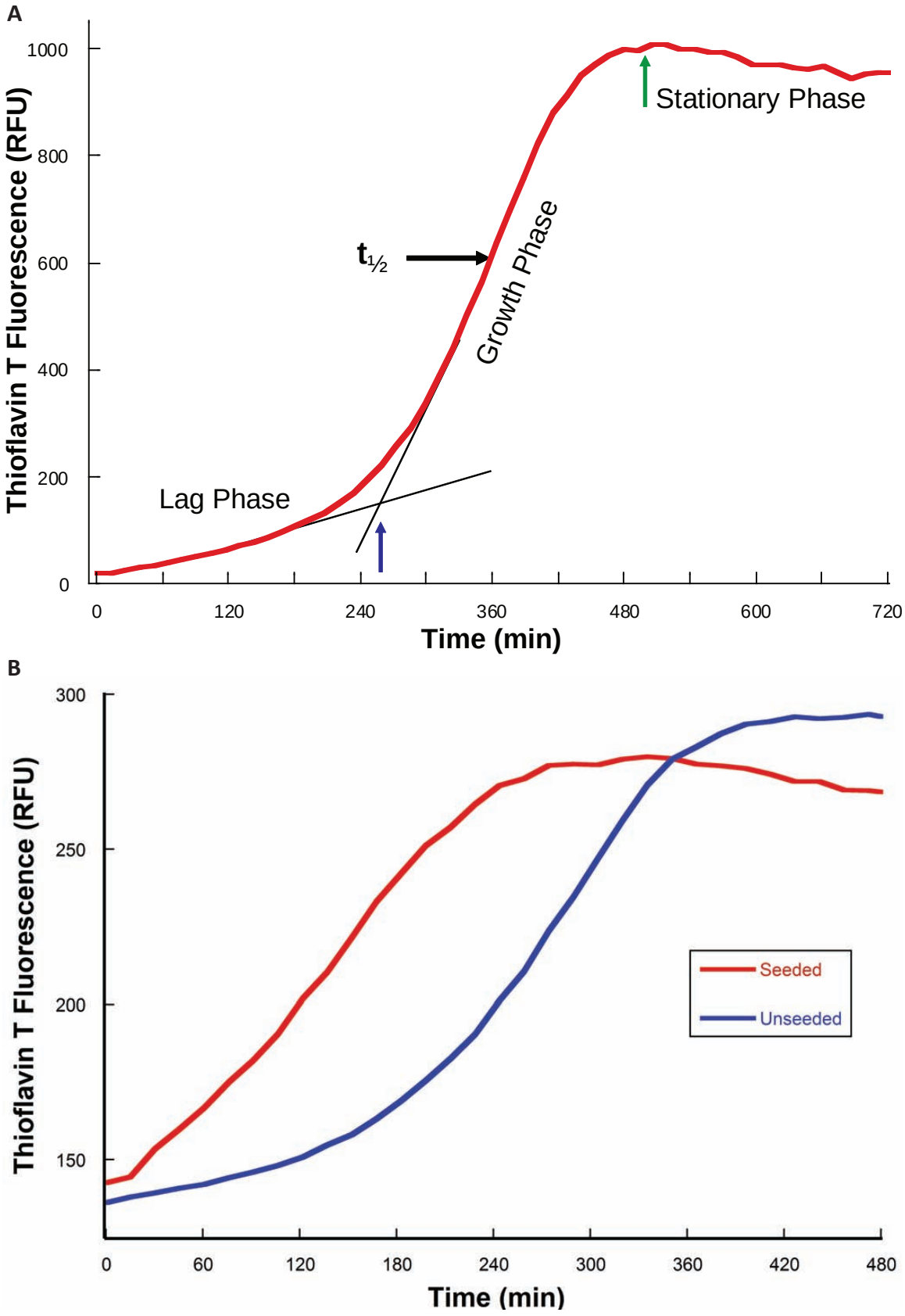


Figure 1.4. CsgA polymerization kinetics and self seeding using Thioflavin T fluorescence.

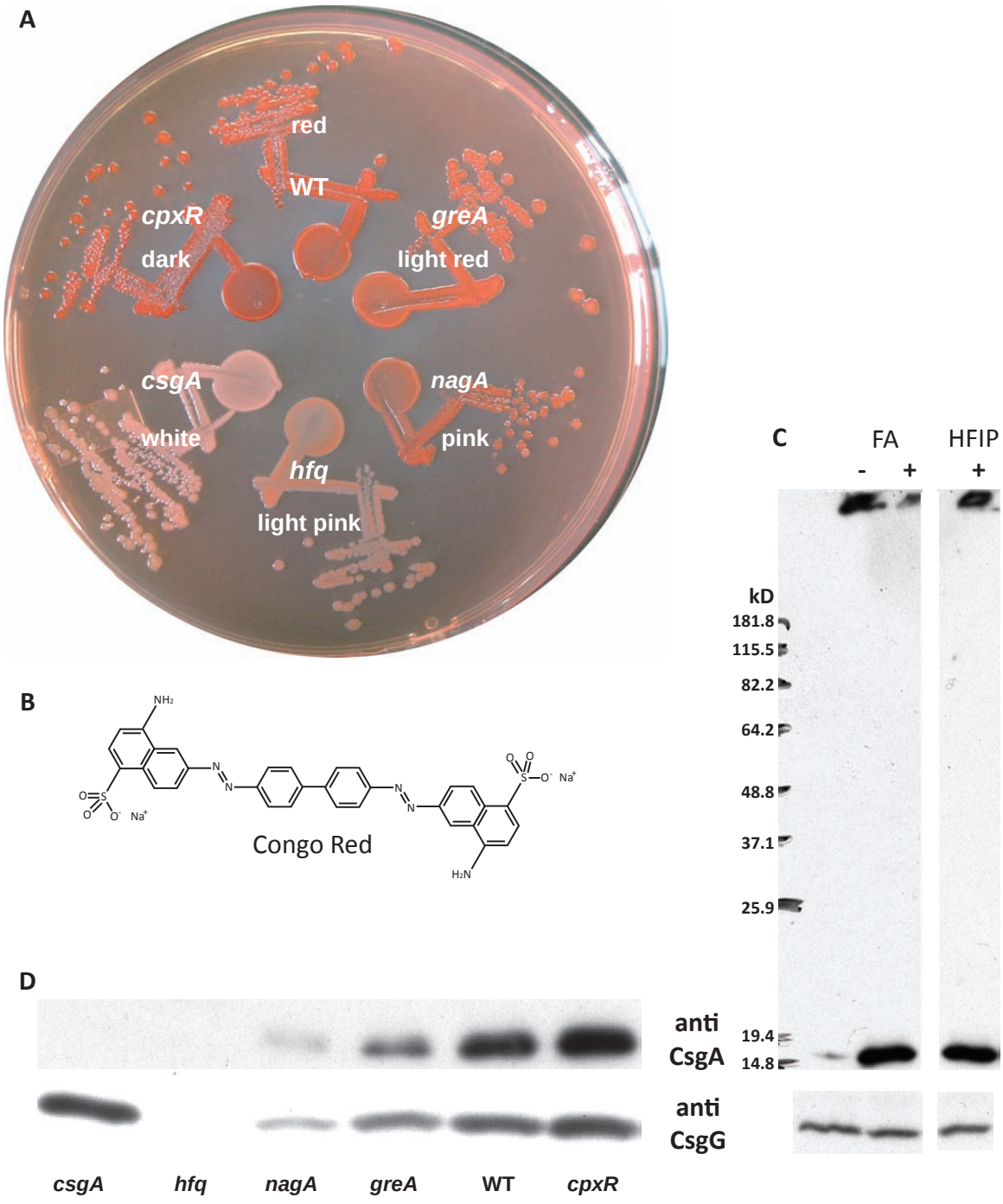


Figure 1.5. Curli fibers bind Congo Red (CR) and are SDS insoluble.

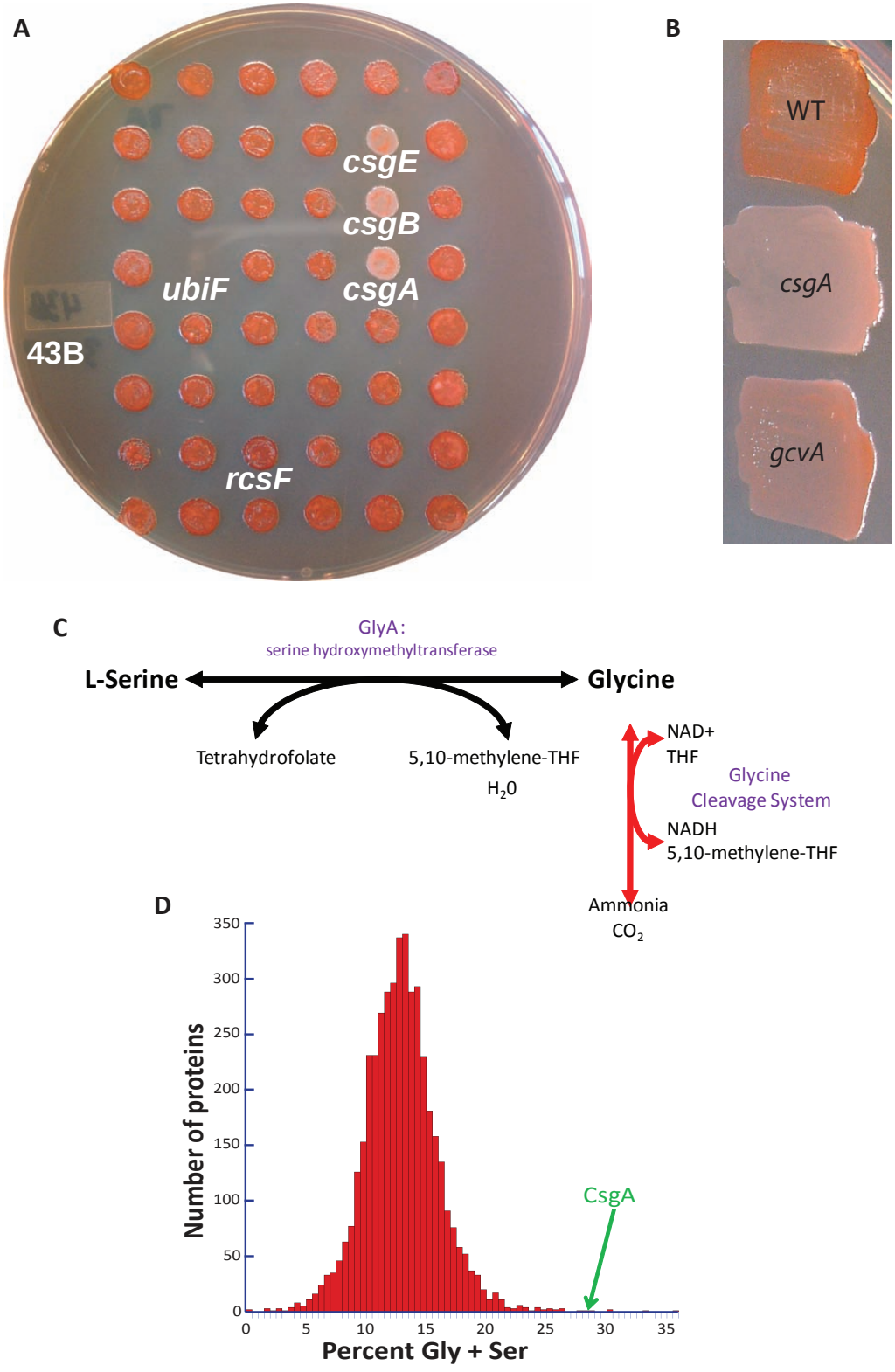


Figure 1.6. Use of Congo Red binding in the Keio collection screen found many mutants including *gcvA* were defective for curli formation leading to an exploration of glycine composition in curli specific genes including CsgA.

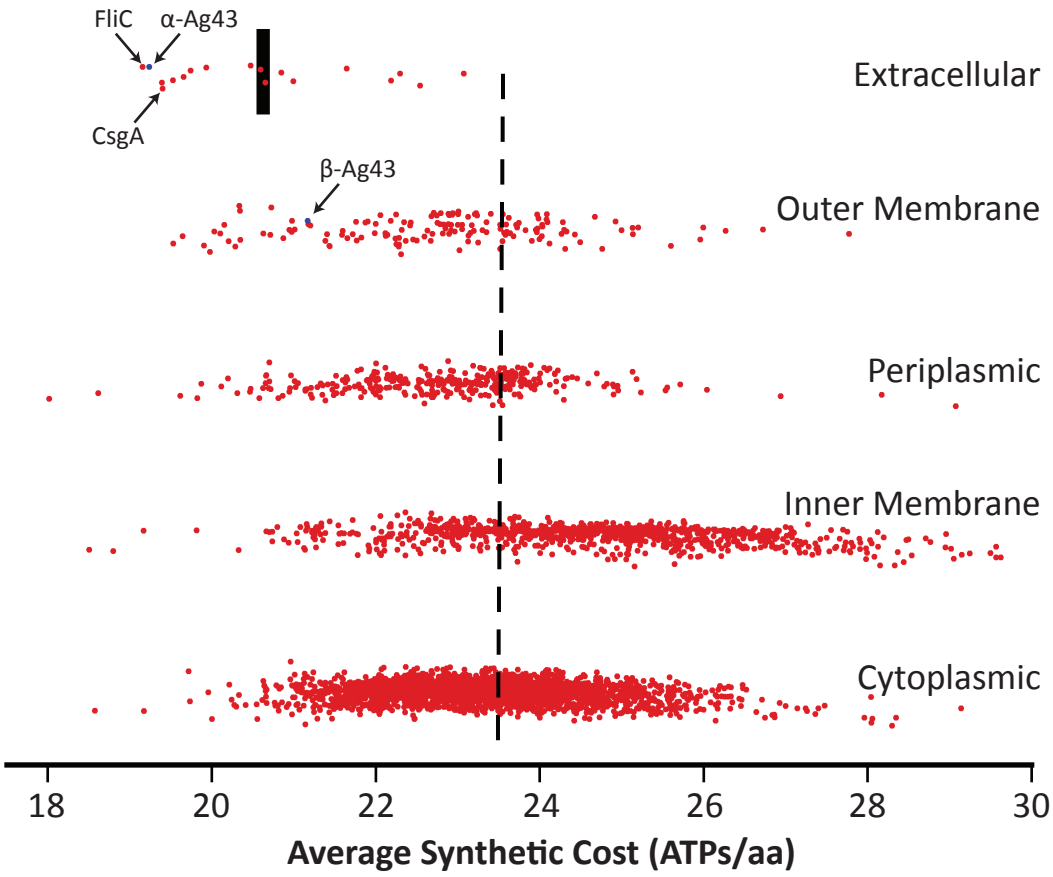


Figure 1.7. Economy of extracellular proteins in *Escherichia coli* strain MG1655.

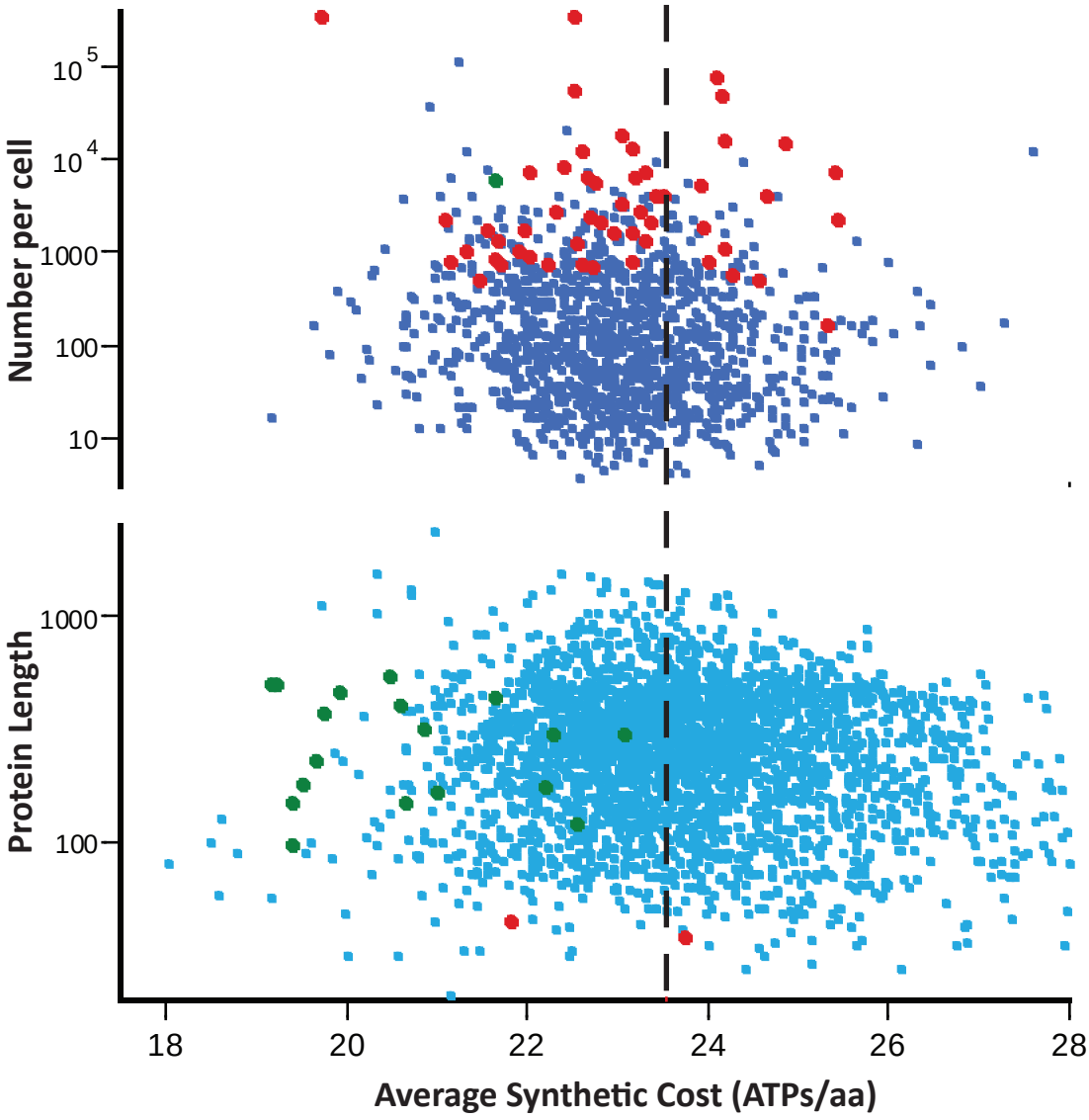


Figure 1.8. The effects of location and abundance on protein amino acid cost.

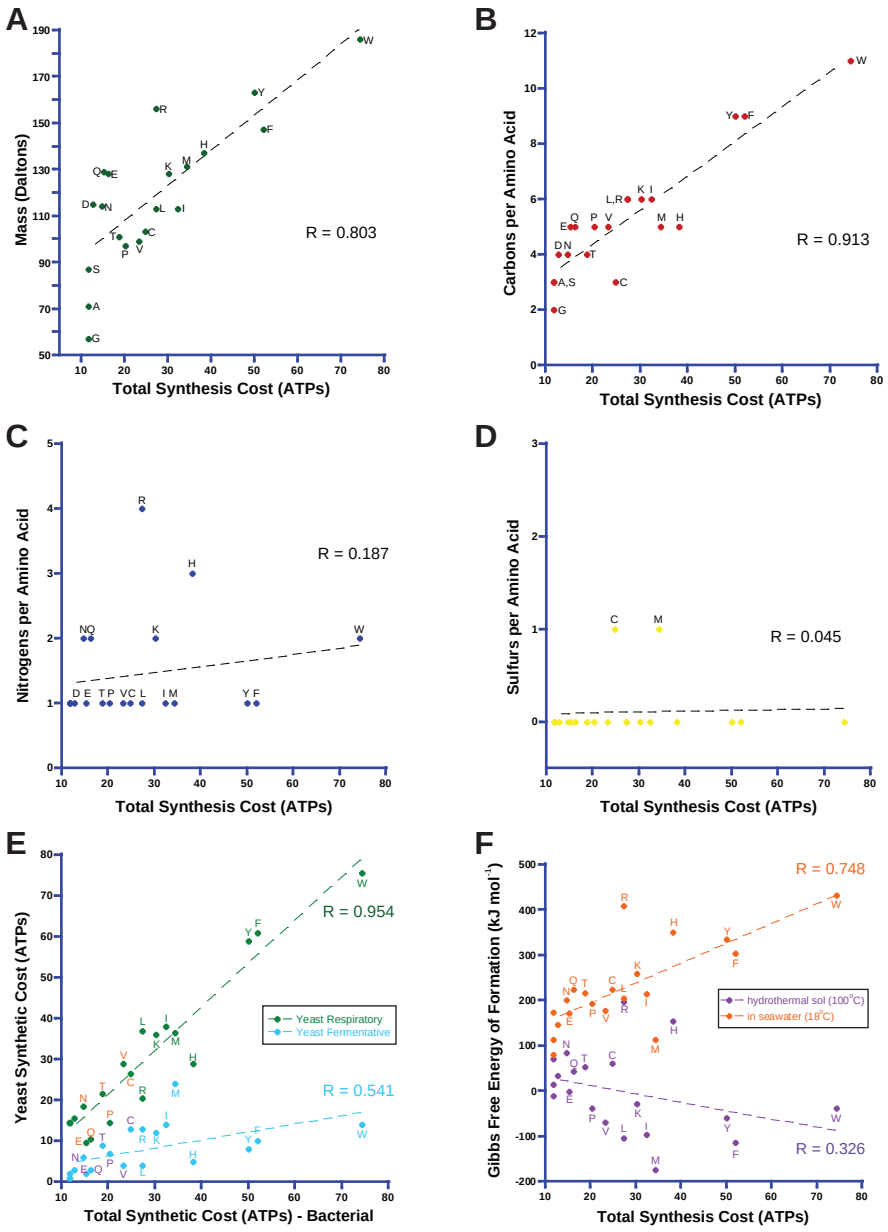


Figure 1.9. Comparison of the synthetic cost of amino acids in bacteria with their atomic mass; number of carbons, nitrogens, or sulfurs; yeast synthetic costs, and Gibbs free energy of formation.

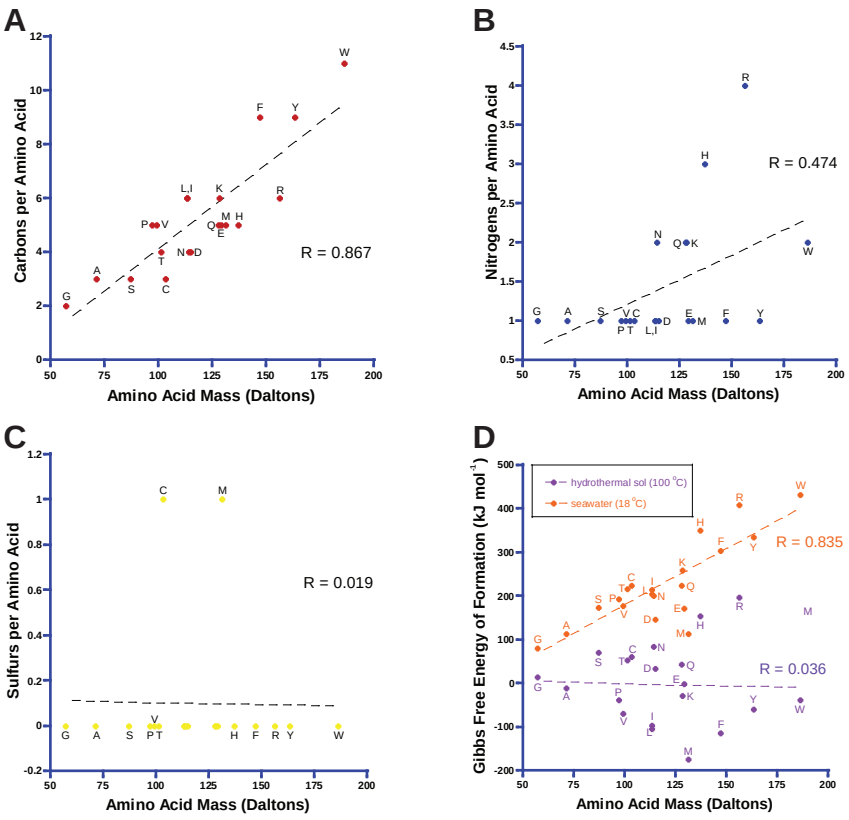


Figure 1.10 Comparison of mass of amino acids to their number of carbons, nitrogens, or sulfurs and Gibbs free energy of formation.

References

1. **Akashi, H., and T. Gojobori.** 2002. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc Natl Acad Sci U S A **99:**3695-700.
2. **Alteri, C. J., J. Xicohtencatl-Cortes, S. Hess, G. Caballero-Olin, J. A. Giron, and R. L. Friedman.** 2007. Mycobacterium tuberculosis produces pili during human infection. Proc Natl Acad Sci U S A **104:**5145-50.
3. **Alves, R., and M. A. Savageau.** 2005. Evidence of selection for low cognate amino acid bias in amino acid biosynthetic enzymes. Mol Microbiol **56:**1017-34.
4. **Amend, J. P., and E. L. Shock.** 1998. Energetics of amino acid synthesis in hydrothermal ecosystems. Science **281:**1659-62.
5. **Arnqvist, A., A. Olsen, and S. Normark.** 1994. Sigma S-dependent growth-phase induction of the csgBA promoter in Escherichia coli can be achieved in vivo by sigma 70 in the absence of the nucleoid-associated protein H-NS. Mol Microbiol **13:**1021-32.
6. **Arnqvist, A., A. Olsen, J. Pfeifer, D. G. Russell, and S. Normark.** 1992. The Crl protein activates cryptic genes for curli formation and fibronectin binding in Escherichia coli HB101. Mol Microbiol **6:**2443-52.
7. **Austin, J. W., G. Sanders, W. W. Kay, and S. K. Collinson.** 1998. Thin aggregative fimbriae enhance Salmonella enteritidis biofilm formation. FEMS Microbiol Lett **162:**295-301.
8. **Baba, T., T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori.** 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol **2:**2006 0008.
9. **Badtke, M. P., N. D. Hammer, and M. R. Chapman.** 2009. Functional amyloids signal their arrival. Sci Signal **2:**pe43.
10. **Barak, J. D., L. Gorski, P. Naraghi-Arani, and A. O. Charkowski.** 2005. Salmonella enterica virulence genes are required for bacterial attachment to plant tissue. Appl Environ Microbiol **71:**5685-91.
11. **Barak, J. D., L. C. Whitehand, and A. O. Charkowski.** 2002. Differences in attachment of Salmonella enterica serovars and Escherichia coli O157:H7 to alfalfa sprouts. Appl Environ Microbiol **68:**4758-63.
12. **Barnhart, M. M., and M. R. Chapman.** 2006. Curli biogenesis and function. Annu Rev Microbiol **60:**131-47.
13. **Barton, M. D., D. Delneri, S. G. Oliver, M. Rattray, and C. M. Bergman.** 2010. Evolutionary systems biology of amino Acid biosynthetic cost in yeast. PLoS ONE **5**.
14. **Baudouin-Cornu, P., Y. Surdin-Kerjan, P. Marliere, and D. Thomas.** 2001. Molecular evolution of protein atomic composition. Science **293:**297-300.
15. **Beloin, C., A. Houry, M. Froment, J. M. Ghigo, and N. Henry.** 2008. A Short-Time Scale Colloidal System Reveals Early Bacterial Adhesion Dynamics. PLoS Biol **6:**e167.

16. **Ben Nasr, A., A. Olsen, U. Sjobring, W. Muller-Esterl, and L. Bjorck.** 1996. Assembly of human contact phase proteins and release of bradykinin at the surface of curli-expressing Escherichia coli. Mol Microbiol **20:**927-35.
17. **Betillison, S., O. Berglund, D. M. Karl, and S. W. Chisholm.** 2003. Elemental composition of marine Prochlorococcus and Synechococcus: Implications for the ecological stoichiometry of the sea. Limnology and Oceanography **48:**1721-1731.
18. **Bian, Z., A. Brauner, Y. Li, and S. Normark.** 2000. Expression of and cytokine activation by Escherichia coli curli fibers in human sepsis. J Infect Dis **181:**602-12.
19. **Bian, Z., and S. Normark.** 1997. Nucleator function of CsgB for the assembly of adhesive surface organelles in Escherichia coli. Embo J **16:**5827-36.
20. **Bian, Z., Z. Q. Yan, G. K. Hansson, P. Thoren, and S. Normark.** 2001. Activation of inducible nitric oxide synthase/nitric oxide by curli fibers leads to a fall in blood pressure during systemic Escherichia coli infection in mice. J Infect Dis **183:**612-9.
21. **Bieler, S., L. Estrada, R. Lagos, M. Baeza, J. Castilla, and C. Soto.** 2005. Amyloid formation modulates the biological activity of a bacterial protein. J Biol Chem **280:**26880-5.
22. **Boer, V. M., J. H. de Winde, J. T. Pronk, and M. D. Piper.** 2003. The genome-wide transcriptional responses of Saccharomyces cerevisiae grown on glucose in aerobic chemostat cultures limited for carbon, nitrogen, phosphorus, or sulfur. J Biol Chem **278:**3265-74.
23. **Bokranz, W., X. Wang, H. Tschape, and U. Romling.** 2005. Expression of cellulose and curli fimbriae by Escherichia coli isolated from the gastrointestinal tract. J Med Microbiol **54:**1171-82.
24. **Bougdour, A., C. Lelong, and J. Geiselmann.** 2004. Crl, a low temperature-induced protein in Escherichia coli that binds directly to the stationary phase sigma subunit of RNA polymerase. J Biol Chem **279:**19540-50.
25. **Box, G. E. P., and N. R. Drape.** 1987. Empirical Model-Building and Response Surfaces 688.
26. **Boyer, R. R., S. S. Sumner, R. C. Williams, M. D. Pierson, D. L. Popham, and K. E. Kniel.** 2007. Influence of curli expression by Escherichia coli 0157:H7 on the cell's overall hydrophobicity, charge, and ability to attach to lettuce. J Food Prot **70:**1339-45.
27. **Bragg, J. G., D. Thomas, and P. Baudouin-Cornu.** 2006. Variation among species in proteomic sulphur content is related to environmental conditions. Proc Biol Sci **273:**1293-300.
28. **Bragg, J. G., and A. Wagner.** 2007. Protein carbon content evolves in response to carbon availability and may influence the fate of duplicated genes. Proc Biol Sci **274:**1063-70.
29. **Brown, P. K., C. M. Dozois, C. A. Nickerson, A. Zuppardo, J. Terlonge, and R. Curtiss, 3rd.** 2001. MlrA, a novel regulator of curli (AgF) and extracellular matrix synthesis by Escherichia coli and Salmonella enterica serovar Typhimurium. Mol Microbiol **41:**349-63.

30. **Castonguay, M. H., S. van der Schaaf, W. Koester, J. Krooneman, W. van der Meer, H. Harmsen, and P. Landini.** 2006. Biofilm formation by Escherichia coli is stimulated by synergistic interactions and co-adhesion mechanisms with adherence-proficient bacteria. Res Microbiol **157:**471-8.
31. **Chapman, M. R., L. S. Robinson, J. S. Pinkner, R. Roth, J. Heuser, M. Hammar, S. Normark, and S. J. Hultgren.** 2002. Role of Escherichia coli curli operons in directing amyloid fiber formation. Science **295:**851-5.
32. **Chirwa, N. T., and M. B. Herrington.** 2003. CsgD, a regulator of curli and cellulose synthesis, also regulates serine hydroxymethyltransferase synthesis in Escherichia coli K-12. Microbiology **149:**525-35.
33. **Chirwa, N. T., and M. B. Herrington.** 2004. Role of MetR and PurR in the activation of glyA by CsgD in Escherichia coli K-12. Can J Microbiol **50:**683-90.
34. **Chiti, F., P. Webster, N. Taddei, A. Clark, M. Stefani, G. Ramponi, and C. M. Dobson.** 1999. Designing conditions for in vitro formation of amyloid protofilaments and fibrils. Proc Natl Acad Sci U S A **96:**3590-4.
35. **Claessen, D., W. de Jong, L. Dijkhuizen, and H. A. Wosten.** 2006. Regulation of Streptomyces development: reach for the sky! Trends Microbiol **14:**313-9.
36. **Claessen, D., R. Rink, W. de Jong, J. Siebring, P. de Vreugd, F. G. Boersma, L. Dijkhuizen, and H. A. Wosten.** 2003. A novel class of secreted hydrophobic proteins is involved in aerial hyphae formation in Streptomyces coelicolor by forming amyloid-like fibrils. Genes Dev **17:**1714-26.
37. **Coleman, M. L., and S. W. Chisholm.** 2007. Code and context: Prochlorococcus as a model for cross-scale biology. Trends Microbiol **15:**398-407.
38. **Collinson, S. K., P. C. Doig, J. L. Doran, S. Clouthier, T. J. Trust, and W. W. Kay.** 1993. Thin, aggregative fimbriae mediate binding of Salmonella enteritidis to fibronectin. J Bacteriol **175:**12-8.
39. **Collinson, S. K., L. Emody, K. H. Muller, T. J. Trust, and W. W. Kay.** 1991. Purification and characterization of thin, aggregative fimbriae from Salmonella enteritidis. J Bacteriol **173:**4773-81.
40. **Collinson, S. K., J. M. Parker, R. S. Hodges, and W. W. Kay.** 1999. Structural predictions of AgfA, the insoluble fimbrial subunit of Salmonella thin aggregative fimbriae. J Mol Biol **290:**741-56.
41. **Cookson, A. L., W. A. Cooley, and M. J. Woodward.** 2002. The role of type 1 and curli fimbriae of Shiga toxin-producing Escherichia coli in adherence to abiotic surfaces. Int J Med Microbiol **292:**195-205.
42. **Coustou-Linares, V., M. L. Maddelein, J. Begueret, and S. J. Saupe.** 2001. In vivo aggregation of the HET-s prion protein of the fungus Podospora anserina. Mol Microbiol **42:**1325-35.
43. **Coustou, V., C. Deleu, S. Saupe, and J. Begueret.** 1997. The protein product of the het-s heterokaryon incompatibility gene of the fungus Podospora anserina behaves as a prion analog. Proc Natl Acad Sci U S A **94:**9773-8.
44. **Coustou, V., C. Deleu, S. J. Saupe, and J. Begueret.** 1999. Mutational analysis of the [Het-s] prion analog of Podospora anserina. A short N-terminal peptide allows prion propagation. Genetics **153:**1629-40.

45. **Craig, C. L., and R. S. Weber.** 1998. Selection costs of amino acid substitutions in ColE1 and ColIa gene clusters harbored by Escherichia coli. Mol Biol Evol **15:**774-6.
46. **Cuhel, R. L. C., C. D. Taylor, and H. W. Jannasch.** 1981. Assimilatory sulfur metabolism in marine microorganisms: Sulfur metabolism, growth, and protein synthesis of Pseudomonas halodurans and Alteromonas luteo-violaceus during sulfate limitation. Archives of Microbiology **130:**1-7.
47. **Eady, R. R.** 1988. The vanadium-containing nitrogenase of Azotobacter. Biofactors **1:**111-6.
48. **Elser, J. J., W. F. Fagan, S. Subramanian, and S. Kumar.** 2006. Signatures of ecological resource availability in the animal and plant proteomes. Mol Biol Evol **23:**1946-51.
49. **Epstein, E. A., and M. R. Chapman.** 2008. Polymerizing the fibre between bacteria and host cells: the biogenesis of functional amyloid fibres. Cell Microbiol **10:**1413-20.
50. **Epstein, E. A., M. A. Reizian, and M. R. Chapman.** 2009. Spatial clustering of the curlin secretion lipoprotein requires curli fiber assembly. J Bacteriol **191:**608-15.
51. **Evin, G., A. Zhu, R. M. Holsinger, C. L. Masters, and Q. X. Li.** 2003. Proteolytic processing of the Alzheimer's disease amyloid precursor protein in brain and platelets. J Neurosci Res **74:**386-92.
52. **Fauchon, M., G. Lagniel, J. C. Aude, L. Lombardia, P. Soularue, C. Petat, G. Marguerie, A. Sentenac, M. Werner, and J. Labarre.** 2002. Sulfur sparing in the yeast proteome in response to sulfur demand. Mol Cell **9:**713-23.
53. **Ferrieres, L., and D. J. Clarke.** 2003. The RcsC sensor kinase is required for normal biofilm formation in Escherichia coli K-12 and controls the expression of a regulon in response to growth on a solid surface. Mol Microbiol **50:**1665-82.
54. **Fowler, D. M., A. V. Koulov, C. Alory-Jost, M. S. Marks, W. E. Balch, and J. W. Kelly.** 2006. Functional amyloid formation within mammalian tissue. PLoS Biol **4:**e6.
55. **Fowler, D. M., A. V. Koulov, W. E. Balch, and J. W. Kelly.** 2007. Functional amyloid--from bacteria to humans. Trends Biochem Sci **32:**217-24.
56. **Gerstel, U., C. Park, and U. Romling.** 2003. Complex regulation of csgD promoter activity by global regulatory proteins. Mol Microbiol **49:**639-54.
57. **Ghrist, A. C., G. Heil, and G. V. Stauffer.** 2001. GcvR interacts with GcvA to inhibit activation of the Escherichia coli glycine cleavage operon. Microbiology **147:**2215-21.
58. **Gibson, D. L., A. P. White, C. M. Rajotte, and W. W. Kay.** 2007. AgfC and AgfE facilitate extracellular thin aggregative fimbriae synthesis in Salmonella enteritidis. Microbiology **153:**1131-40.
59. **Giovannoni, S. J., H. J. Tripp, S. Givan, M. Podar, K. L. Vergin, D. Baptista, L. Bibbs, J. Eads, T. H. Richardson, M. Noordewier, M. S. Rappe, J. M. Short, J. C. Carrington, and E. J. Mathur.** 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science **309:**1242-5.

60. **Gophna, U., M. Barlev, R. Seijffers, T. A. Oelschlager, J. Hacker, and E. Z. Ron.** 2001. Curli fibers mediate internalization of Escherichia coli by eukaryotic cells. Infect Immun **69:**2659-65.
61. **Gophna, U., T. A. Oelschlaeger, J. Hacker, and E. Z. Ron.** 2002. Role of fibronectin in curli-mediated internalization. FEMS Microbiol Lett **212:**55-8.
62. **Hales, B. J., E. E. Case, J. E. Morningstar, M. F. Dzeda, and L. A. Mauterer.** 1986. Isolation of a new vanadium-containing nitrogenase from Azotobacter vinelandii. Biochemistry **25:**7251-5.
63. **Hammar, M., A. Arnqvist, Z. Bian, A. Olsen, and S. Normark.** 1995. Expression of two csg operons is required for production of fibronectin- and congo red-binding curli polymers in Escherichia coli K-12. Mol Microbiol **18:**661-70.
64. **Hammar, M., Z. Bian, and S. Normark.** 1996. Nucleator-dependent intercellular assembly of adhesive curli organelles in Escherichia coli. Proc Natl Acad Sci U S A **93:**6562-6.
65. **Hammarstrom, P., R. L. Wiseman, E. T. Powers, and J. W. Kelly.** 2003. Prevention of transthyretin amyloid disease by changing protein misfolding energetics. Science **299:**713-6.
66. **Hammer, N. D., J. C. Schmidt, and M. R. Chapman.** 2007. The curli nucleator protein, CsgB, contains an amyloidogenic domain that directs CsgA polymerization. Proc Natl Acad Sci U S A **104:**12494-9.
67. **Hammer, N. D., X. Wang, B. A. McGuffie, and M. R. Chapman.** 2008. Amyloids: friend or foe? J Alzheimers Dis **13:**407-19.
68. **Hawkins, C. L., D. I. Pattison, and M. J. Davies.** 2003. Hypochlorite-induced oxidation of amino acids, peptides and proteins. Amino Acids **25:**259-74.
69. **Heil, G., L. T. Stauffer, and G. V. Stauffer.** 2002. Glycine binds the transcriptional accessory protein GcvR to disrupt a GcvA/GcvR interaction and allow GcvA-mediated activation of the Escherichia coli gcvTHP operon. Microbiology **148:**2203-14.
70. **Heizer, E. M., Jr., D. W. Raiford, M. L. Raymer, T. E. Doom, R. V. Miller, and D. E. Krane.** 2006. Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. Mol Biol Evol **23:**1670-80.
71. **Henderson, I. R., F. Navarro-Garcia, M. Desvaux, R. C. Fernandez, and D. Ala'Aldeen.** 2004. Type V protein secretion pathway: the autotransporter story. Microbiol Mol Biol Rev **68:**692-744.
72. **Herwald, H., M. Morgelin, A. Olsen, M. Rhen, B. Dahlback, W. Muller-Esterl, and L. Bjorck.** 1998. Activation of the contact-phase system on bacterial surfaces--a clue to serious complications in infectious diseases. Nat Med **4:**298-302.
73. **Hubmacher, D., B. F. Matzanke, and S. Anemuller.** 2003. Effects of iron limitation on the respiratory chain and the membrane cytochrome pattern of the Euryarchaeon Halobacterium salinarum. Biol Chem **384:**1565-73.
74. **Hughes, K. T., K. L. Gillen, M. J. Semon, and J. E. Karlinsey.** 1993. Sensing structural intermediates in bacterial flagellar assembly by export of a negative regulator. Science **262:**1277-80.

75. **Hung, D. L., T. L. Raivio, C. H. Jones, T. J. Silhavy, and S. J. Hultgren.** 2001. Cpx signaling pathway monitors biogenesis and affects assembly and expression of P pili. Embo J **20:**1508-18.
76. **Hurst, L. D., E. J. Feil, and E. P. Rocha.** 2006. Protein evolution: causes of trends in amino-acid gain and loss. Nature **442:**E11-2; discussion E12.
77. **Ishihama, A.** 2010. Prokaryotic genome regulation: multifactor promoters, multitarget regulators and hierarchic networks. FEMS Microbiol Rev.
78. **Jensen, L. J., M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering.** 2009. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res **37:**D412-6.
79. **Jeter, C., and A. G. Matthysse.** 2005. Characterization of the binding of diarrheagenic strains of E. coli to plant surfaces and the role of curli in the interaction of the bacteria with alfalfa sprouts. Mol Plant Microbe Interact **18:**1235-42.
80. **Johansson, C., T. Nilsson, A. Olsen, and M. J. Wick.** 2001. The influence of curli, a MHC-I-binding bacterial surface structure, on macrophage-T cell interactions. FEMS Immunol Med Microbiol **30:**21-9.
81. **Jones, J.** Unpublished observations.
82. **Jourdan, A. D., and G. V. Stauffer.** 1999. GcvA-mediated activation of gcvT-lacZ expression involves the carboxy-terminal domain of the alpha subunit of RNA polymerase. FEMS Microbiol Lett **181:**307-12.
83. **Karp, P. D., I. M. Keseler, A. Shearer, M. Latendresse, M. Krummenacker, S. M. Paley, I. Paulsen, J. Collado-Vides, S. Gama-Castro, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penaloza-Spinola, C. Bonavides-Martinez, and J. Ingraham.** 2007. Multidimensional annotation of the Escherichia coli K-12 genome. Nucleic Acids Res **35:**7577-90.
84. **Kayed, R., E. Head, J. L. Thompson, T. M. McIntire, S. C. Milton, C. W. Cotman, and C. G. Glabe.** 2003. Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. Science **300:**486-9.
85. **Khurana, R., C. Coleman, C. Ionescu-Zanetti, S. A. Carter, V. Krishna, R. K. Grover, R. Roy, and S. Singh.** 2005. Mechanism of thioflavin T binding to amyloid fibrils. J Struct Biol **151:**229-38.
86. **Khurana, R., V. N. Uversky, L. Nielsen, and A. L. Fink.** 2001. Is Congo red an amyloid-specific dye? J Biol Chem **276:**22715-21.
87. **Kikuchi, T., Y. Mizunoe, A. Takade, S. Naito, and S. Yoshida.** 2005. Curli fibers are required for development of biofilm architecture in Escherichia coli K-12 and enhance bacterial adherence to human uroepithelial cells. Microbiol Immunol **49:**875-84.
88. **Kim, J. K., and M. A. Harrison.** 2009. Surrogate selection for Escherichia coli O157:H7 based on cryotolerance and attachment to romaine lettuce. J Food Prot **72:**1385-91.

89. **Kim, S. H., and Y. H. Kim.** 2004. Escherichia coli O157:H7 adherence to HEp-2 cells is implicated with curli expression and outer membrane integrity. J Vet Sci **5:**119-24.
90. **King, J. L., and T. H. Jukes.** 1969. Non-Darwinian evolution. Science **164:**788-98.
91. **Klunk, W. E., R. F. Jacob, and R. P. Mason.** 1999. Quantifying amyloid by congo red spectral shift assay. Methods Enzymol **309:**285-305.
92. **Klunk, W. E., J. W. Pettegrew, and D. J. Abraham.** 1989. Quantitative evaluation of congo red binding to amyloid-like proteins with a beta-pleated sheet conformation. J Histochem Cytochem **37:**1273-81.
93. **Krebs, M. R., E. H. Bromley, and A. M. Donald.** 2005. The binding of thioflavin-T to amyloid fibrils: localisation and implications. J Struct Biol **149:**30-7.
94. **La Ragione, R. M., W. A. Cooley, and M. J. Woodward.** 2000. The role of fimbriae and flagella in the adherence of avian strains of Escherichia coli O78:K80 to tissue culture cells and tracheal and gut explants. J Med Microbiol **49:**327-38.
95. **La Ragione, R. M., I. M. McLaren, G. Foster, W. A. Cooley, and M. J. Woodward.** 2002. Phenotypic and genotypic characterization of avian Escherichia coli O86:K61 isolates possessing a gamma-like intimin. Appl Environ Microbiol **68:**4932-42.
96. **Lane, T. W., and F. M. Morel.** 2000. A biological function for cadmium in marine diatoms. Proc Natl Acad Sci U S A **97:**4627-31.
97. **Lane, T. W., and F. M. Morel.** 2000. Regulation of carbonic anhydrase expression by zinc, cobalt, and carbon dioxide in the marine diatom Thalassiosira weissflogii. Plant Physiol **123:**345-52.
98. **Lane, T. W., M. A. Saito, G. N. George, I. J. Pickering, R. C. Prince, and F. M. Morel.** 2005. Biochemistry: a cadmium enzyme from a marine diatom. Nature **435:**42.
99. **Lapidot, A., and S. Yaron.** 2009. Transfer of Salmonella enterica serovar Typhimurium from contaminated irrigation water to parsley is dependent on curli and cellulose, the biofilm matrix components. J Food Prot **72:**618-23.
100. **Lawley, T. D., K. Chan, L. J. Thompson, C. C. Kim, G. R. Govoni, and D. M. Monack.** 2006. Genome-wide screen for Salmonella genes required for long-term systemic infection of the mouse. PLoS Pathog **2:**e11.
101. **LeVine, H., 3rd.** 1999. Quantification of beta-sheet amyloid fibril structures with thioflavin T. Methods Enzymol **309:**274-84.
102. **LeVine, H., 3rd.** 1997. Stopped-flow kinetics reveal multiple phases of thioflavin T binding to Alzheimer beta (1-40) amyloid fibrils. Arch Biochem Biophys **342:**306-16.
103. **LeVine, H., 3rd.** 1993. Thioflavine T interaction with synthetic Alzheimer's disease beta-amyloid peptides: detection of amyloid aggregation in solution. Protein Sci **2:**404-10.
104. **Li, N., J. Lv, and D. K. Niu.** 2009. Low contents of carbon and nitrogen in highly abundant proteins: evidence of selection for the economy of atomic composition. J Mol Evol **68:**248-55.

105. **Loferer, H., M. Hammar, and S. Normark.** 1997. Availability of the fibre subunit CsgA and the nucleator protein CsgB during assembly of fibronectin-binding curli is limited by the intracellular concentration of the novel lipoprotein CsgG. Mol Microbiol **26:**11-23.
106. **Lv, J., N. Li, and D. K. Niu.** 2008. Association between the availability of environmental resources and the atomic composition of organismal proteomes: evidence from Prochlorococcus strains living at different depths. Biochem Biophys Res Commun **375:**241-6.
107. **Macnab, R. M.** 2003. How bacteria assemble flagella. Annu Rev Microbiol **57:**77-100.
108. **Maji, S. K., M. H. Perrin, M. R. Sawaya, S. Jessberger, K. Vadodaria, R. A. Rissman, P. S. Singru, K. P. Nilsson, R. Simon, D. Schubert, D. Eisenberg, J. Rivier, P. Sawchenko, W. Vale, and R. Riek.** 2009. Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. Science **325:**328-32.
109. **Maji, S. K., D. Schubert, C. Rivier, S. Lee, J. E. Rivier, and R. Riek.** 2008. Amyloid as a depot for the formulation of long-acting drugs. PLoS Biol **6:**e17.
110. **Masuda, T., N. Saito, M. Tomita, and Y. Ishihama.** 2009. Unbiased quantitation of Escherichia coli membrane proteome using phase-transfer surfactants. Mol Cell Proteomics.
111. **Mazel, D., and P. Marliere.** 1989. Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. Nature **341:**245-8.
112. **Minamino, T., B. Gonzalez-Pedrajo, K. Yamaguchi, S. I. Aizawa, and R. M. Macnab.** 1999. FliK, the protein responsible for flagellar hook length control in Salmonella, is exported during hook assembly. Mol Microbiol **34:**295-304.
113. **Misra, R. V., R. S. Horler, W. Reindl, Goryanin, II, and G. H. Thomas.** 2005. EchoBASE: an integrated post-genomic database for Escherichia coli. Nucleic Acids Res **33:**D329-33.
114. **Morel, F. M.** 2008. The co-evolution of phytoplankton and trace element cycles in the oceans. Geobiology **6:**318-24.
115. **Morel, F. M., and N. M. Price.** 2003. The biogeochemical cycles of trace metals in the oceans. Science **300:**944-7.
116. **Nenninger, A. A., L. S. Robinson, and S. J. Hultgren.** 2009. Localized and efficient curli nucleation requires the chaperone-like amyloid assembly protein CsgF. Proc Natl Acad Sci U S A **106:**900-5.
117. **Ogasawara, H., A. Hasegawa, E. Kanda, T. Miki, K. Yamamoto, and A. Ishihama.** 2007. Genomic SELEX search for target promoters under the control of the PhoQP-RstBA signal relay cascade. J Bacteriol **189:**4791-9.
118. **Ogasawara, H., K. Yamada, A. Kori, K. Yamamoto, and A. Ishihama.** 2010. Regulation of the E. coli csgD Promoter: Interplay between Five Transcription Factors. Microbiology.
119. **Oh, J., J. G. Kim, E. Jeon, C. H. Yoo, J. S. Moon, S. Rhee, and I. Hwang.** 2007. Amyloidogenesis of type III-dependent harpins from plant pathogenic bacteria. J Biol Chem **282:**13601-9.

120. **Olsen, A., A. Arnqvist, M. Hammar, S. Sukupolvi, and S. Normark.** 1993. The RpoS sigma factor relieves H-NS-mediated transcriptional repression of csgA, the subunit gene of fibronectin-binding curli in Escherichia coli. Mol Microbiol **7:**523-36.
121. **Olsen, A., H. Herwald, M. Wikstrom, K. Persson, E. Mattsson, and L. Bjorck.** 2002. Identification of two protein-binding and functional regions of curli, a surface organelle and virulence determinant of Escherichia coli. J Biol Chem **277:**34568-72.
122. **Olsen, A., A. Jonsson, and S. Normark.** 1989. Fibronectin binding mediated by a novel class of surface organelles on Escherichia coli. Nature **338:**652-5.
123. **Olsen, A., M. J. Wick, M. Morgelin, and L. Bjorck.** 1998. Curli, fibrous surface proteins of Escherichia coli, interact with major histocompatibility complex class I molecules. Infect Immun **66:**944-9.
124. **Otoo, H. N., K. G. Lee, W. Qiu, and P. N. Lipke.** 2008. Candida albicans Als adhesins have conserved amyloid-forming sequences. Eukaryot Cell **7:**776-82.
125. **Otzen, D., and P. H. Nielsen.** 2008. We find them here, we find them there: functional bacterial amyloid. Cell Mol Life Sci **65:**910-27.
126. **Palenik, B., J. Grimwood, A. Aerts, P. Rouze, A. Salamov, N. Putnam, C. Dupont, R. Jorgensen, E. Derelle, S. Rombauts, K. Zhou, R. Otillar, S. S. Merchant, S. Podell, T. Gaasterland, C. Napoli, K. Gendler, A. Manuell, V. Tai, O. Vallon, G. Piganeau, S. Jancek, M. Heijde, K. Jabbari, C. Bowler, M. Lohr, S. Robbens, G. Werner, I. Dubchak, G. J. Pazour, Q. Ren, I. Paulsen, C. Delwiche, J. Schmutz, D. Rokhsar, Y. Van de Peer, H. Moreau, and I. V. Grigoriev.** 2007. The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci U S A **104:**7705-10.
127. **Pardee, A. B.** 1966. Purification and properties of a sulfate-binding protein from Salmonella typhimurium. J Biol Chem **241:**5886-92.
128. **Pawar, D. M., M. L. Rossman, and J. Chen.** 2005. Role of curli fimbriae in mediating the cells of enterohaemorrhagic Escherichia coli to attach to abiotic surfaces. J Appl Microbiol **99:**418-25.
129. **Peers, G., and N. M. Price.** 2004. A role for manganese in superoxide dismutases and growth of iron-deficient diatoms. Limnology and Oceanography **49:**1774-1783.
130. **Perlstein, E. O., B. L. de Bivort, S. Kunes, and S. L. Schreiber.** 2007. Evolutionarily conserved optimization of amino acid biosynthesis. J Mol Evol **65:**186-96.
131. **Persson, K., W. Russell, M. Morgelin, and H. Herwald.** 2003. The conversion of fibrinogen to fibrin at the surface of curliated Escherichia coli bacteria leads to the generation of proinflammatory fibrinopeptides. J Biol Chem **278:**31884-90.
132. **Pratt, L. A., and T. J. Silhavy.** 1998. Crl stimulates RpoS activity during stationary phase. Mol Microbiol **29:**1225-36.
133. **Prigent-Combaret, C., G. Prensier, T. T. Le Thi, O. Vidal, P. Lejeune, and C. Dorel.** 2000. Developmental pathway for biofilm formation in curli-producing Escherichia coli strains: role of flagella, curli and colanic acid. Environ Microbiol **2:**450-64.

134. **Raiford, D. W., E. M. Heizer, Jr., R. V. Miller, H. Akashi, M. L. Raymer, and D. E. Krane.** 2008. Do amino acid biosynthetic costs constrain protein evolution in Saccharomyces cerevisiae? J Mol Evol **67:**621-30.
135. **Richmond, R. C.** 1970. Non-Darwinian evolution: a critique. Nature **225:**1025-8.
136. **Robinson, L. S., E. M. Ashman, S. J. Hultgren, and M. R. Chapman.** 2006. Secretion of curli fibre subunits is mediated by the outer membrane-localized CsgG protein. Mol Microbiol **59:**870-81.
137. **Rocha, E. P., and A. Danchin.** 2002. Base composition bias might result from competition for metabolic resources. Trends Genet **18:**291-4.
138. **Rocha, E. P., A. Sekowska, and A. Danchin.** 2000. Sulphur islands in the Escherichia coli genome: markers of the cell's architecture? FEBS Lett **476:**8-11.
139. **Romero, D., C. Aguilar, R. Losick, and R. Kolter.** Amyloid fibers provide structural integrity to Bacillus subtilis biofilms. Proc Natl Acad Sci U S A **107:**2230-4.
140. **Romling, U.** 2005. Characterization of the rdar morphotype, a multicellular behaviour in Enterobacteriaceae. Cell Mol Life Sci **62:**1234-46.
141. **Romling, U., Z. Bian, M. Hammar, W. D. Sierralta, and S. Normark.** 1998. Curli fibers are highly conserved between Salmonella typhimurium and Escherichia coli with respect to operon structure and regulation. J Bacteriol **180:**722-31.
142. **Romling, U., W. D. Sierralta, K. Eriksson, and S. Normark.** 1998. Multicellular and aggregative behaviour of Salmonella typhimurium strains is controlled by mutations in the agfD promoter. Mol Microbiol **28:**249-64.
143. **Roterman, I., M. KrUl, M. Nowak, L. Konieczny, J. Rybarska, B. Stopa, B. Piekarska, and G. Zemanek.** 2001. Why Congo red binding is specific for amyloid proteins - model studies and a computer analysis approach. Med Sci Monit **7:**771-84.
144. **Ryu, J. H., and L. R. Beuchat.** 2005. Biofilm formation by Escherichia coli O157:H7 on stainless steel: effect of exopolysaccharide and Curli production on its resistance to chlorine. Appl Environ Microbiol **71:**247-54.
145. **Ryu, J. H., H. Kim, J. F. Frank, and L. R. Beuchat.** 2004. Attachment and biofilm formation on stainless steel by Escherichia coli O157:H7 as affected by curli production. Lett Appl Microbiol **39:**359-62.
146. **Saldana, Z., J. Xicohtencatl-Cortes, F. Avelino, A. D. Phillips, J. B. Kaper, J. L. Puente, and J. A. Giron.** 2009. Synergistic role of curli and cellulose in cell adherence and biofilm formation of attaching and effacing Escherichia coli and identification of Fis as a negative regulator of curli. Environ Microbiol **11:**992-1006.
147. **Sankaran, B., S. A. Bonnett, K. Shah, S. Gabriel, R. Reddy, P. Schimmel, D. A. Rodionov, V. de Crecy-Lagard, J. D. Helmann, D. Iwata-Reuyl, and M. A. Swairjo.** 2009. Zinc-independent folate biosynthesis: genetic, biochemical, and structural investigations reveal new metal dependence for GTP cyclohydrolase IB. J Bacteriol **191:**6936-49.
148. **Seligmann, H.** 2003. Cost-minimization of amino acid usage. J Mol Evol **56:**151-61.

149. **Shorter, J., and S. Lindquist.** 2008. Hsp104, Hsp70 and Hsp40 interplay regulates formation, growth and elimination of Sup35 prions. Embo J **27:**2712-24.
150. **Si, K., S. Lindquist, and E. R. Kandel.** 2003. A neuronal isoform of the aplysia CPEB has prion-like properties. Cell **115:**879-91.
151. **Sjobring, U., G. Pohl, and A. Olsen.** 1994. Plasminogen, absorbed by Escherichia coli expressing curli or by Salmonella enteritidis expressing thin aggregative fimbriae, can be activated by simultaneously captured tissue-type plasminogen activator (t-PA). Mol Microbiol **14:**443-52.
152. **Solomon, E. B., B. A. Niemira, G. M. Sapers, and B. A. Annous.** 2005. Biofilm formation, cellulose production, and curli biosynthesis by Salmonella originating from produce, animal, and clinical sources. J Food Prot **68:**906-12.
153. **Stadtman, E. R., and B. S. Berlett.** 1991. Fenton chemistry. Amino acid oxidation. J Biol Chem **266:**17201-11.
154. **Stauffer, G. V., L. T. Stauffer, and M. D. Plamann.** 1989. The Salmonella typhimurium glycine cleavage enzyme system. Mol Gen Genet **220:**154-6.
155. **Stauffer, L. T., S. J. Fogarty, and G. V. Stauffer.** 1994. Characterization of the Escherichia coli gcv operon. Gene **142:**17-22.
156. **Stauffer, L. T., M. D. Plamann, and G. V. Stauffer.** 1986. Cloning and characterization of the glycine-cleavage enzyme system of Escherichia coli. Gene **44:**219-26.
157. **Stauffer, L. T., and G. V. Stauffer.** 2005. GcvA interacts with both the alpha and sigma subunits of RNA polymerase to activate the Escherichia coli gcvB gene and the gcvTHP operon. FEMS Microbiol Lett **242:**333-8.
158. **Steensma, D. P.** 2001. "Congo" red: out of Africa? Arch Pathol Lab Med **125:**250-2.
159. **Stopa, B., B. Piekarska, L. Konieczny, J. Rybarska, P. Spolnik, G. Zemanek, I. Roterman, and M. Krol.** 2003. The structure and protein binding of amyloid-specific dye reagents. Acta Biochim Pol **50:**1213-27.
160. **String.** http://string.embl.de/.
161. **Sukupolvi, S., R. G. Lorenz, J. I. Gordon, Z. Bian, J. D. Pfeifer, S. J. Normark, and M. Rhen.** 1997. Expression of thin aggregative fimbriae promotes interaction of Salmonella typhimurium SR-11 with mouse small intestinal epithelial cells. Infect Immun **65:**5320-5.
162. **Szabo, E., A. Skedsmo, A. Sonnevend, K. Al-Dhaheri, L. Emody, A. Usmani, and T. Pal.** 2005. Curli expression of enterotoxigenic Escherichia coli. Folia Microbiol (Praha) **50:**40-6.
163. **Torres, A. G., C. Jeter, W. Langley, and A. G. Matthysse.** 2005. Differential binding of Escherichia coli O157:H7 to alfalfa, human epithelial cells, and plastic is mediated by a variety of surface structures. Appl Environ Microbiol **71:**8008-15.
164. **Tukel, C., J. H. Nishimori, R. P. Wilson, M. G. Winter, A. M. Keestra, J. P. van Putten, and A. J. Baumler.** 2010. Toll-like receptors 1 and 2 cooperatively mediate immune responses to curli, a common amyloid from enterobacterial biofilms. Cell Microbiol.

165. **Tukel, C., M. Raffatellu, A. D. Humphries, R. P. Wilson, H. L. Andrews-Polymenis, T. Gull, J. F. Figueiredo, M. H. Wong, K. S. Michelsen, M. Akcelik, L. G. Adams, and A. J. Baumler.** 2005. CsgA is a pathogen-associated molecular pattern of Salmonella enterica serotype Typhimurium that is recognized by Toll-like receptor 2. Mol Microbiol **58:**289-304.
166. **Tukel, C., R. P. Wilson, J. H. Nishimori, M. Pezeshki, B. A. Chromy, and A. J. Baumler.** 2009. Responses to amyloids of microbial and host origin are mediated through toll-like receptor 2. Cell Host Microbe **6:**45-53.
167. **Uhlich, G. A., P. H. Cooke, and E. B. Solomon.** 2006. Analyses of the red-dry-rough phenotype of an Escherichia coli O157:H7 strain and its role in biofilm formation and resistance to antibacterial agents. Appl Environ Microbiol **72:**2564-72.
168. **Uhlich, G. A., N. W. t. Gunther, D. O. Bayles, and D. A. Mosier.** 2009. The CsgA and Lpp proteins of an Escherichia coli O157:H7 strain affect HEp-2 cell invasion, motility, and biofilm formation. Infect Immun **77:**1543-52.
169. **Uhlich, G. A., J. E. Keen, and R. O. Elder.** 2002. Variations in the csgD promoter of Escherichia coli O157:H7 associated with increased virulence in mice and increased invasion of HEp-2 cells. Infect Immun **70:**395-9.
170. **Uniprot.** http://www.uniprot.org/.
171. **Van Mooy, B. A., H. F. Fredricks, B. E. Pedler, S. T. Dyhrman, D. M. Karl, M. Koblizek, M. W. Lomas, T. J. Mincer, L. R. Moore, T. Moutin, M. S. Rappe, and E. A. Webb.** 2009. Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity. Nature **458:**69-72.
172. **Van Mooy, B. A., G. Rocap, H. F. Fredricks, C. T. Evans, and A. H. Devol.** 2006. Sulfolipids dramatically decrease phosphorus demand by picocyanobacteria in oligotrophic marine environments. Proc Natl Acad Sci U S A **103:**8607-12.
173. **Vianney, A., G. Jubelin, S. Renault, C. Dorel, P. Lejeune, and J. C. Lazzaroni.** 2005. Escherichia coli tol and rcs genes participate in the complex network affecting curli synthesis. Microbiology **151:**2487-97.
174. **Vidal, O., R. Longin, C. Prigent-Combaret, C. Dorel, M. Hooreman, and P. Lejeune.** 1998. Isolation of an Escherichia coli K-12 mutant strain able to form biofilms on inert surfaces: involvement of a new ompR allele that increases curli expression. J Bacteriol **180:**2442-9.
175. **Vieira-Silva, S., and E. P. Rocha.** 2008. An assessment of the impacts of molecular oxygen on the evolution of proteomes. Mol Biol Evol **25:**1931-42.
176. **Wagner, A.** 2005. Energy constraints on the evolution of gene expression. Mol Biol Evol **22:**1365-74.
177. **Wang, X., and M. R. Chapman.** 2008. Curli provide the template for understanding controlled amyloid propagation. Prion **2:**57-60.
178. **Wang, X., N. D. Hammer, and M. R. Chapman.** 2008. The molecular basis of functional bacterial amyloid polymerization and nucleation. J Biol Chem.

179. **Wang, X., M. Rochon, A. Lamprokostopoulou, H. Lunsdorf, M. Nimtz, and U. Romling.** 2006. Impact of biofilm matrix components on interaction of commensal Escherichia coli with the gastrointestinal cell line HT-29. Cell Mol Life Sci **63:**2352-63.
180. **Wang, X., D. R. Smith, J. W. Jones, and M. R. Chapman.** 2007. In Vitro Polymerization of a Functional Escherichia coli Amyloid Protein. J Biol Chem **282:**3713-9.
181. **White, A. P., D. L. Gibson, W. Kim, W. W. Kay, and M. G. Surette.** 2006. Thin aggregative fimbriae and cellulose enhance long-term survival and persistence of Salmonella. J Bacteriol **188:**3219-27.
182. **White, A. P., and M. G. Surette.** 2006. Comparative genetics of the rdar morphotype in Salmonella. J Bacteriol **188:**8395-406.
183. **Wickner, R. B., H. K. Edskes, B. T. Roberts, U. Baxa, M. M. Pierce, E. D. Ross, and A. Brachmann.** 2004. Prions: proteins as genes and infectious entities. Genes Dev **18:**470-85.
184. **Wickner, R. B., H. K. Edskes, and F. Shewmaker.** 2006. How to find a prion: [URE3], [PSI+] and [beta]. Methods **39:**3-8.
185. **Wickner, R. B., K. L. Taylor, H. K. Edskes, M. L. Maddelein, H. Moriyama, and B. T. Roberts.** 2000. Prions of yeast as heritable amyloidoses. J Struct Biol **130:**310-22.
186. **Wilson, R. L., and G. V. Stauffer.** 1994. DNA sequence and characterization of GcvA, a LysR family regulatory protein for the Escherichia coli glycine cleavage enzyme system. J Bacteriol **176:**2862-8.
187. **Wilson, R. L., L. T. Stauffer, and G. V. Stauffer.** 1993. Roles of the GcvA and PurR proteins in negative regulation of the Escherichia coli glycine cleavage enzyme system. J Bacteriol **175:**5129-34.
188. **Wilson, R. L., M. L. Urbanowski, and G. V. Stauffer.** 1995. DNA binding sites of the LysR-type regulator GcvA in the gcv and gcvA control regions of Escherichia coli. J Bacteriol **177:**4940-6.
189. **Wonderling, L. D., M. L. Urbanowski, and G. V. Stauffer.** 2000. GcvA binding site 1 in the gcvTHP promoter of Escherichia coli is required for GcvA-mediated repression but not for GcvA-mediated activation. Microbiology **146 (Pt 11):**2909-18.
190. **Xu, G., and M. R. Chance.** 2005. Radiolytic modification of sulfur-containing amino acid residues in model peptides: fundamental studies for protein footprinting. Anal Chem **77:**2437-49.
191. **Xu, Y., L. Feng, P. D. Jeffrey, Y. Shi, and F. M. Morel.** 2008. Structure and metal exchange in the cadmium carbonic anhydrase of marine diatoms. Nature **452:**56-61.
192. **Zakikhany, K., C. R. Harrington, M. Nimtz, J. C. Hinton, and U. Romling.** 2010. Unphosphorylated CsgD controls biofilm formation in Salmonella enterica serovar Typhimurium. Mol Microbiol.
193. **Zogaj, X., W. Bokranz, M. Nimtz, and U. Romling.** 2003. Production of cellulose and curli fimbriae by members of the family Enterobacteriaceae isolated from the human gastrointestinal tract. Infect Immun **71:**4151-8.

Chapter II

A Complex Genetic Network Governs Curli Biogenesis in *Escherichia coli*

Abstract

Curli are stable proteinaceous fibers assembled within the extracellular matrix of many enteric bacteria including *Escherichia coli* and *Salmonella typhimurium*. Curli expression and assembly are known to require two fiber subunit genes, three secretion factors, and several transcriptional regulators. Using the Keio strain collection of single gene deletions, we screened for mutations that affect curli production in *E. coli* K-12 by assessing colony phenotypes on Congo Red indicator plates. We discovered more than three hundred genes modulate curli production. Genes affecting many fundamental cellular processes including metabolism, environmental sensing, gene regulation, outer membrane biosynthesis, molecular transport, and protein turnover had altered curli production. For example, deletion of the sodium antiporter NhaA led to loss of curli and an increase in motility. Null mutants in several purine, pyrimidine, and inner core lipopolysaccharide (LPS) biosynthetic genes also made less curli. The alternative sigma RpoS and RpoE appear to have opposing roles in curli production. Mutations such as *nuoG, clpX,* and *sdhC* that increased RpoS levels produced more curli, while strains with lower RpoS including *dnaK* and *dksA* made less curli. Mutants predicted to induce RpoE expression, such as *rseA*, inner core LPS and enteric common antigen (ECA) mutants, were correlated with a reduction in curli gene expression. Using previously published studies, we found only little overlap between genes idenfied as important for curli biogenesis with genes affecting swimming and swarming motility. The large number of genes identified in this study that affect curli biogenesis, suggest that curli production is part of a highly regulated and complex developmental pathway in *Escherichia coli*.

Introduction

Curli are thin aggregative fimbriae produced by many *Enterobacteriaceae* including *Escherichia coli* and *Salmonella typhimurium*. (18, 23) (See Fig. 2.11). As the major protein component of the extracellular matrix, curli are vital for mature biofilms (57). Curliated bacteria better adhere to and colonize a variety of surfaces (5, 50, 106, 122, 126) and are more resistant to environmental damage from desiccation or antimicrobial agents (106, 122, 135). Curli fibers have also been implicated in host cell adhesion, invasion, and immune system activation (8, 38, 129).

Curli assembly is a complex process that requires six proteins encoded by the divergent operons *csgBAC* and *csgDEFG* (18, 41). At the cell surface, the major curli fiber subunit CsgA is nucleated into an extracellular amyloid fiber by the minor fiber subunit CsgB (9, 18, 42). Secretion of both curli fiber subunits requires the lipoprotein CsgG, which forms a putative pore in the outer membrane (18, 30, 69, 98). CsgE and CsgF are accessory proteins that assist in subunit secretion and the proper localization of CsgB, respectively (18, 79). CsgC is dispensable for curli formation; however, it may influence the morphology of curli fibers (36, 41).

Curli production is affected by a number of additional gene products, most of which affect curli through the regulation of the *csgDEFG* operon (See Table 2.1) (6). Previous screens to identify additional accessory factors and regulators have identified only a handful of genes such as *nagA* (7, 40, 41). Therefore, we performed a comprehensive screen for genes involved in curli production using the Keio collection (4). The Keio collection is a near complete set of single gene knockouts in *E. coli* K-12. Using Congo Red (CR) indicator plates, we found more than 300 strains had reproducible changes in curli production. These strains lack genes involved in a variety of cellular processes including cell envelope biogenesis, metabolism, transport, and regulation. Therefore, we propose curli fiber formation requires significant changes in the physiology and lifestyle of *Escherichia coli*.

Results and Discussion

Screening to identify genes involved in curli production.

To better understand curli production and regulation, we screened the Keio collection for additional curli associated genes using Congo Red (CR) indicator plates (Fig. 2.1A). Curliated bacterial colonies turn red when grown on CR plates, while strains that do not produce curli remain white (22). The Keio collection consists of 3985 nonessential gene mutants comprising more than 90% of the 4,390 open reading frames of *E. coli* K-12 (4). Following growth at 26°C for 48 hours, strains with changes in CR phenotypes were isolated and scored according to their color, which varied from the lighter pink of *csgA* to the darker red of *rcsF* (Fig. 2.1A). Some strains including *thyA*, *glnA*, *ubiE* and *ubiF* (Fig. 2.1A) did not grow well under the conditions tested.

Strains with an altered phenotype were regrown on CR indicator plates for verification and rescored numerically according to color from lightest to darkest (Fig. 2.1B and Table 2.5). We isolated 318 Keio mutants with reproducible CR phenotypes; however, 64 strains had two or more phenotypes upon isolation of single colonies. The original gene deletion in the strains with multiple CR phenotypes was verified by PCR, and most had the correct Kan^r insert for each phenotype (Table 2.6). Many of the strains with multiple CR phenotype strains had colonies with WT curli production in addition to a mutant CR phenotype. Therefore, we excluded these strains from further study as they likely have second site suppressors. However, we included the c*mk, fabF, mdoG, pgm, trpD, trpE, yiaK,* and *ymgE* strains because the CR phenotypes were always less than WT cells. The resulting 265 mutants were 13.7% white, 17.9% light pink, 30.8% pink, 19.0% light red, and 18.6% dark red when grown on CR indicator plates (Fig. 2.1D).

To better quantify curli production in the identified mutants, we measured the CsgG and CsgA levels using western blotting (Fig. 2.1C and Table 2.5). The levels of both proteins were generally consistent with CR phenotypes (Compare Fig. 2.1B,C). However, in some light pink mutants, the putative curli secretion protein CsgG (69, 98) was present; whereas, no CsgA was detectable (Table 2.5, Fig. 2.1C and 2.2C).

We divided the genes required for WT levels of CR binding into clusters of orthologous groups (COGs) (4, 67) and used Echobase (75) to assign a cellular location to each gene (Tables 2.3 and 2.5, Fig. 2.5). Collectively, the genes identified in our screen are predicted to perform a variety of functional roles in the cell and are primarily cytoplasmic. Of the genes outside the *csg* operon that are known to affect curli, most are proposed to affect curli production through transcriptional changes (6); exceptions include *ycfR, rfaG, lpxM* (*msbB*), and *ddhC* (Table 2.3). Most of the CR phenotypes of the Keio collection were consistent with the literature: exceptions are primarily found in cyclic-di-GMP and regulatory proteins some of which are known to have strain dependent effects on curli production (Table 2.3) (6). In the following sections, we explore genes from several different functional groups and discuss likely roles of their gene functions in curli production.

Cell Envelope

Genes involved in basic cellular processes like metabolism, regulation, motility and envelope biogenesis were identified in our screen (Tables 2.3). Because curli assembly occurs on the cell surface, we focused on the cell envelope and outer membrane biogenesis COG family group (M). The cell envelope includes the inner membrane, the periplasm, the outer membrane and extracellular structures including fimbriae and extracellular polysaccharides (EPS) (132).

Multiple lipopolysaccharide (LPS) biosynthesis gene mutants, many of which are involved in assembling the inner core region, were defective for curli production (Fig. 2.2A,B). Curli associated genes in the inner core region of LPS encode hexose and heptose transferases, enzymes that produce the inner core sugar building block ADP-L-glycero-D-manno-heptose, the LPS kinase *rfaP*, and the transcriptional antiterminator for the *rfaQGPSBIJYZK* operon *rfaH* (89, 91, 109). Western blot analysis showed that many LPS mutants with CR defects had corresponding decreases in CsgA and CsgG levels (Fig. 2.2A,C). Mutants in the inner core sugar transferases *rfaC, rfaF, rfaG,* and the heptose biosynthesis genes *lpcA* and *rfaE* were the most defective.

Several LPS mutant strains had a notably dry and friable texture including *galU* and *rfaG* and to a lesser degree *rfaF* (84, 109). GalU produces UDP-D-Glucose and is required for growth on galactose and trehalose and the synthesis of trehalose, the LPS outer core, colonic acid and periplasmic glucans (11, 32, 56, 133). Since RfaG adds UTD-D-Glucose to the inner core of LPS, the *galU* and *rfaG* mutants should be functionally similar. Both strains displayed a light pink phenotype, were visibly drier and crumbled when scraped from plates (84), autoagglutinated when resuspended in PBS (109), had similarly low levels of CsgG by western blotting, and displayed identical LPS profiles on silver stained PAGE gels (32) (Table 2.5, Fig. 2.3A,B). We also tested cell hydrophobicity of the inner core sugar transferases using the BATH method (103) and found that strains with more developed inner LPS cores had progressively more hydrophilic membranes (compare *rfaC,F,G* in Fig. 2.2B and 2.3C). As expected, both *galU* and *rfaG* had similar cell hydrophobicity profiles. Surprisingly, a *csgA* strain was more hydrophobic than the wild type curliated strain. Accordingly, the lack of curli in the *ycfR* strain may explain its high hydrophobicity in a previous study (136).

Both *rfaD* and *gmhB* were pinker and had higher Csg protein levels than *lpcA, rfaE,* or *rfaC,* which were lighter pink and contained very little detectable CsgA and CsgG (Fig 2.2A). However, all these gene products are involved in the process of synthesizing or transferring the ADP-L-glycero-D-manno-heptose to Lipid A-KDOII (91). Deletion of the epimerase *rfaD* results in accumulation of a mixture of mostly heptose free LPS, some LPS with the D-glycero-D-manno-heptose isomer, and a trace amount of LPS with the correct L,D heptose isomer (Fig. 2.3A) (21). Strains lacking *gmhB* accumulate a mixture of heptose free and L-glycero-D-manno-heptose core LPS (Fig. 2.3A) (60); another haoloacid-dehalogenase-like hydrolase paralog may partially complement GmhB function (60, 63). The mixture of LPS in the *rfaD* and *gmhB* strains with and without core regions may explain their intermediate curli phenotype (See Fig. 2.3A). Unlike the *rfaG* strain and other inner core mutants, both *rfaD* and *gmhB* strains have a small amount of Glucose I and further modified LPS. Since the *rfaI* strain is unaffected

for curli production, it appears even a small amount of Glucose I modified LPS is capable of supporting curli secretion and polymerization.

Previous studies have implicated LPS biosynthesis genes including *rfaG, ddhC (rfbH)* (1), and *lpxM* (*msbB*) (58) in curli production. In *Salmonella enterica* both *rfaG* and *ddhC* strains produce less curli but more cellulose than wild type (1). Similarly, the *rfaG* strain in K-12 had reduced curli production; however, there was no increase in cellulose production, as many K-12 strains including MC4100 do not produce cellulose (70, 139) (Fig. 2.6). We also looked at other lipid A modifying enzymes to see if they affected curli production. The *lpxL, lpxM* (*msbB), lpxP, pagP,* and *arnT* mutant strains all produced curli at WT levels suggesting the modification state of the lipid A core does not necessarily affect curli production in K-12 strains of *E. coli* (Fig. 2.10).

Other components of the cell envelope include the enteric common antigen (ECA) and the periplasmic glucans. Mutants in ECA biosynthesis with altered curli production include *rfe, rffA, rffC, rffT,* and *wzxE* (Fig. 2.7A). RffA, RffC, and RffT are involved in synthesis or addition of the TDP-Fuc4NAc to lipid II (Fig. 2.7B) (92). Strains lacking these proteins accumulate lipid II and have increased *degP* levels via RpoE and CpxR induction (25). Mutants in *rfe* suppress the *degP* activation in the *rffA, rffH,* and *rffT* strains suggesting undecaprenyl-P and lipid II accumulation have opposing roles in envelope stress responses and curli production (25). Mutants in the lipid III flipase gene *wzxE* accumulate lipid III, which is toxic to the cell and can be overcome by earlier blocks in the ECA synthesis pathway (92, 96). Lipid III also accumulates in the inner membrane of *wzyE* mutants, the ECA chain elongation gene (92), which may be essential (4). These results suggest ECA is not required for curli production as only mutants accumulating lipids II and III are curli deficient. Conversely, the *rfe* strain, which accumulates undecaprenyl-P, produced more curli than wild type cells (Fig. 2.7). Mutants in the osmoregulated periplasmic glucan (OPG) synthesis genes also had changes in curli production. Mutants in OPG structural genes *mdoG* and *mdoH* resulted in less curli whereas the OPG succinyl modification mutants in *mdoC* produced more curli (Table 2.5). The *mdoG* mutant also had white muccoid colonies in addition to pink mutants

suggesting induction of the Rcs pathway (Table 2.6). Others have reported an *mdoH* mutant induced colonic acid biosynthesis (29).

The cell envelope is also composed of several extracellular polysaccharides (EPS) including colonic acid, poly-β-1,6-GlcNAc (PGA), cellulose, and an unidentified curli associated polysaccharide (128, 134); however, only the colonic acid gene *galU* was found to affect curli production. Other mutants in EPS biosynthetic pathways except *treR* (56) all had wild type levels of curli indicating that the *galU* strain's curli phenotype is likely due to defects in LPS or OPG biosynthesis. Knockouts of genes involved in peptidoglycan synthesis or lipid A formation also had normal curli production. However, a strain lacking RodZ, which interacts with peptidoglycan and is essential for the rod shape in *E. coli (112),* was defective for curli production. Many of these genes are essential and omitted from the Keio collection (4). Collectively, these results suggest the minimum cell envelope needed for efficient curli production is partially Glucose I substituted LPS and basic OPG structure.

Nucleotide and amino acid biosynthesis

Mutants in metabolic genes (COG families C, E-I) constitute nearly forty percent of the total strains with altered CR phenotypes (Table 2.1, Fig. 2.5). Deletions in most curli associated metabolic genes are defective for curli production, particularly among nucleotide and amino acid biosynthesis genes. Nutrient availability may play a significant role in the ability of cells to produce curli fibers. We found several pathways where nearly every gene in the pathway had defective CR phenotype, including the *de novo* purine and pyrimidine biosynthesis pathways. Accordingly, the *cytR* strain produced less curli likely due to derepression of nucleoside degradation and utilization genes (56).

Relative to *pyr* strains, mutations in purine biosynthesis genes were more defective for curli production suggesting a limitation of purine under our growth conditions (Table 2.4). Both pathways had relatively few mutants with wild type levels of curli production: *purT, pyrL, pyrI, carA,* and *guaA* (Table 2.4). All but the latter play

either a regulatory role or have isozymes that may duplicate their function (56, 78, 104). Intriguingly, some mutants in pyrimidine synthetic genes including *pyrE, pyrB,* and *carB* had increased CR binding when streaked near other strains indicating a diffusible small molecule may be able to rescue the defect in curli production (See Fig. 2.12).

Mutants in most of the genes for amino acid biosynthesis were absent from our screen. However, all five tryptophan biosynthesis genes were defective for curli production: *trpA, trpB, trpC, trpD,* and *trpE.* Our screen also resulted in multiple chorismate biosynthesis gene mutants with considerable curli defects: *aroA, aroB, aroC, aroD, aroE, aroK, aroH,* and *aroP* (Table 2.5). Chorismate is used in the biosynthetic pathways of aromatic amino acids, ubiquinone, menoquinone, THF, and enterobactin (4); however, with the exception of *aspC* and *ubiE,* curli defectives strains were only found in ubiquinone and tryptophan pathways. Mutants in *aroF, aroG,* and *aroL* had wild type levels of curli; however, their isozymes *aroH* and *aroK* were moderately defective (Table 2.5) (56).

The *gcvA, aspA, aspC*, *asnC,* and *glnA* strains which have defects in serine, glycine, asparagine, and glutamine biosynthesis, respectively, were also defective for curli production. A possible explanation lies in the high content of these amino acids in some curli proteins (Table 3.8). For example, CsgA consists of 28.5% glycine or serine (mean 7.13%), the sixth highest such protein in *E. coli* (Fig. 1.6D). Previously, over expression of CsgD has been shown to increase *glyA* levels (19). GlyA is a serine hydroxymethyltransferase (SHMT) which interconverts Ser to Gly (Fig. 1.6C) (56). The activation of *glyA* by CsgD would increase SHMT activity and may allow sufficient pools of serine and glycine for efficient translation of curli proteins rich in these amino acids such as CsgA. CsgA and the minor curli subunit CsgB also have high levels of Asn and Gln, two amino acids often associated with functional amyloids including the yeast prions Sup35 and Rnq1 (110, 116) (See Fig. 3.9). CsgA and CsgB have 17.9% and 18.5% asparagine or glutamine (mean 8.3%), the 18th and 12th highest proteins in *E. coli*, respectively.

Inorganic ion transport and metabolism

The *nhaA* mutant did not produce detectable amounts of curli (Fig. 2.4A,B). One of three sodium ion antiporters, NhaA uses the proton electrochemical gradient to expel sodium ions (56). Curli fibers are maximally produced in low salt conditions at room temperature (100). The *nhaA* mutant was light pink on CR YESCA plates (Fig. 2.4A), suggesting the lack of NhaA leads to the accumulation of intracellular sodium ions which inhibit curli production. Western blotting for CsgA and CsgG confirmed that the *nhaA* strain had no detectable levels of either curli protein (Fig. 2.4B). After 48 hours, the *nhaA* strains had longer cells, lacked curli fibers, and had considerably more flagellated cells than BW25113 and the *csgA* strain (Fig. 2.4D-F). Western blotting for the flagella component FliC confirmed this observation (Fig. 2.4G). The *nhaA* mutant was also more motile in YESCA motility plates than WT (Fig 2.4C). Surprisingly, *csgA* and *csgG* showed decreased motility relative to WT. It is unclear if the increased hydrophobicity of the *csgA* strain or the induction of a stress response of either *csg* mutant (Margery Evan unpublished data) is related to their reduced motility (Fig. 2.3C).

Several mutants for genes in iron acquisition displayed changes in curli production including *dps, fepB, fepD,* and *fepG.* Low iron conditions have been shown to increase curli production at 37°C in *Salmonella typhimurium* (100). We saw both decreased and increased curli production for iron acquisition mutants, and *fepE* displayed both dark and light pink phenotypes. The affects of iron on curli production are likely complex and need further study.

We found several mutants related to sulfate assimilation including *cysB, cysC, cysE, cysH, cysI,* and *cysG.* CysG is an uroporphyrin III C-methyltransferase necessary for the seroheme cofactor used in CysI (130). Most of the proteins had increased curli production except the *cysH* strain that had both pink and wild type colonies (Table 2.6). CysE modifies serine to O-acetyl-serine and requires CysM or a complex with CysK to fully convert to L-Cys (56). Additionally, the *metC* strain produced less curli; MetC is one of three L-cysteine desulfhydrases and degrades L-Cys. Collectively, these results suggest that lower levels of cysteine may stimulate curli production.

Carbohydrate metabolism and energy production

Several genes involved in global carbohydrate flux and sugar import were found to be defective for curli production: *cyaA, crp, fruR* (*cra*)*, ptsH, ptsI, aceE, fbp, gnd,* and *tpiA.* The metabolic flux changes of the Keio *cyaA, crp,* and *fruR* strains have been examined under different growth conditions; the *crp, cyaA* and *fruR* strains had less phosphoenolpyruvate (PEP) from oxaloacetate in glucose limiting conditions (77, 86). FruR increases PEP production from pyruvate through a combination of *pykF* repression and *ppsA* activation (107). PEP is used by the phosphotransferase system (PTS) to transport and phosphorylate many different sugars. Additionally, the cAMP-CRP complex activates *ptsHI-crr* (138), and *cyaA* and *crp* mutants have low levels of glucose uptake due to a PTS defect (78). Combined with the curli phenotype for the PTS genes *ptsH (*Enzyme I) and *ptsI* (HPr), our results suggest a defect in the PTS system results in lower curli production. However, many of these gene products have global effects on gene transcription: the cAMP/CRP complex regulates multiple genes including direct activation of *csgDEFG* (138).

Enzymes for central metabolism, energy production, and their coenzymes also play an important role in curli production. TCA cycle mutants with reduced curli production included genes encoding enzymes for the complete conversion of α-ketoglutrate to fumarate: *sdhA,B* and *sucA,B,C,D* (Tables 2.5 and 2.6). However, *sucC,D* is typically present during anaerobic growth where curli production is reduced (35, 78, 99). Why curli defective mutants are restricted to this portion of the TCA cycle is unknown and contrasts with the negative regulation of *sdhBACD* by RpoS and Crl (93, 97), both positive regulators of curli production (6). Several mutants for quinol oxidoreductases gave variable curli phenotypes. SdhA and SdhB, the soluble cytoplasmic components of succinate quinone reductase (SQR), are sufficient for succinate oxidation (78) and produced less curli than WT. Intriguingly, *sdhC* and *sdhD* mutants lacking the SQR membrane subunits produced more curli than WT. The *nuoA,B,C,E,F,G,J,K,M,N* strains which are defective for NADH dehydrogenase:quinone reductase I (NDH-I) also expressed more curli than WT and appeared dark red on CR plates (Table 2.5).

Conversely, an *appC* strain lacking the stationary phase (3) induced cytochrome bd-II quinol oxidase (CBD-II) (3) produced very little curli. Unlike SQR, both NHD-I and CBD-II contribute to the proton gradient needed for membrane transport and ATP production; accordingly, mutants in the ATP synthase genes *atpA,B,C,E,F,G,H* grew more slowly and had greatly reduced amounts of curli (Table 2.5).

The opposing CR phenotypes of quinol oxidase and quinone reductase deficient strains suggest quinone has a negative effect on curli production. Previously, a *nuoG* strain had reduced *rpoS* transcription and a CBD-I mutant *cydA* had higher levels (111). Quinone can affect RpoS levels through the Arc two-component system. Quinones cause disulfide bonding in the inner membrane protein ArcB (33, 71). ArcB cross linking limits trans-phosphorylation of ArcA and RssB which increases *rpoS* transcription and RpoS stability, respectively (71, 74). Thus the increased curli production in these quinone reductases could be explained by higher levels of RpoS.

Mutants strains *cysG, lipA,B, lpd, pdxH,* and *ubiE,F,G,H* which lack cofactors and coenzymes required for some of the proteins described above also produced less curli. LipA and LipB are responsible for biosynthesis and transfer of lipoate to Lpd, respectively (56). Lpd is utilized by pyruvate dehydrogenase (*aceEF*), α-ketoglutrate dehydrogenase (*sucAB*), and the glycine cleavage (*gcvTHP*) multienzyme complexes (56) of which the *sucA* and *sucB* strains were the most defective (Tables 2.5 and 2.6). The ubiquinone biosynthesis mutants produced very little curli and grew slower than normal.

Previously, our lab investigated the affect of *nagA* mutants on curli production. In the Keio collection, we found both *nagA* and *nagC* mutants to be defective for curli production. For the *nagA* strain, the decrease in curli production was similar to other K-12 strains and much less than seen in the C600 strain (7). Intriguingly, we found *yhbJ* and *pcnB* strains were dark red on CR plates, and produced significantly more curli than BW25113. YhbJ has recently been shown to destabilize the RNA *glmZ* which increases the *glmS* transcript stability (55, 94). Typically this *glmS* transcript is polyadenylated by PcnB and rapidly degraded (124). These results suggest that the curli defect in a *nagA*

strain may be due to more than an accumulation of GlcNac-6P. Instead defects in the levels of UDP-GlcNac needed for lipid A, ECA, and peptidoglycan biosynthesis may lead to a compromised envelope and subsequently to lower curli expression. The curli defect of the Keio *glmM* mutant would seem to confirm this; however, *glmM* is apparently an essential gene (73), and therefore its presence in the Keio collection suggests that the *glmM* strain used in our study has acquired suppressor mutations. We also found decreased curli in *nanK* and *nanE*. Both are involved in sialic acid biosynthesis which ultimately is converted to GlcNAc-6P (56).

Regulatory networks control curli gene expression

At 754 bases the non-coding region between *csgD* and *csgB* is the fifth largest intergenic region between divergent operons in *E. coli* K-12 (Table 2.8, Fig 2.8A) and the thirteenth largest overall (Fig 2.8B) (105). The intergenic region between *csgD* and *csgB* has strong inherent curvature (85) and multiple regulatory proteins binding sites (6, 47, 56, 82). Most of the proteins previously shown to affect curli production are proposed to do so through transcriptional changes at one or both curli operons and include several proteins that modify cyclic-di-GMP metabolism (120) (Table 2.2). We found strains with defects in transcriptional regulation and signal transduction that were defective for curli production. The large number of regulatory mutants affecting curli production is consistent with the multiple regulatory binding sites afforded by the large *csgD_csgB* intergenic region.

CsgD has been shown to regulate a number of genes outside the curli operon (14, 15, 19, 20, 37, 39, 65) including some found in this screen: *dps, ompF,* and *pyrB* (Table 2.9). Over-expression of CsgD results in down regulation of the iron acquisition genes *fhuE, fecR,* and *dps* (14). Dps is a nucleoid protein expressed in stationary phase that binds and stores iron (56, 137). The down regulation of iron acquisition by CsgD conflicts with the stimulatory effects of low iron on curli production in *Salmonella typhimurium* (100), and the increased curli produced by many of the *fep* strains (Tables 2.5 and 2.6).

Several mutants identified in our screen have been shown to be involved in directly or indirectly regulating RpoS levels or function including *crp, clpP, clpX, dksA, dnaK, galU, hns, hfq, nuoG, pgm,* and *mdoA (mdoGH)* (10, 44, 68, 111). Recently, CsgD itself has been shown to increase levels of RpoS via activation of *iraP* (39). IraP stabilizes RpoS by preventing RssB mediated ClpXP degradation (13). Like an *rpoS* mutant, an *nlpD* mutant completely lacks curli, probably due to loss of the major *rpoS* promoter (64). Since DksA affects ppGpp induction of RpoS (16), we decided to test whether altering ppGpp production would affect curli production. A *relA spoT* double mutant in MG1655 was light pink on CR plates and produced almost no curli proteins (Fig 2.9). In addition to *rpoS,* we found other sigma factors defective for curli production including *rpoN* and *rpoZ.*

The low curli production in the ECA, LPS, and *tol pal* strains indicates induction of the sigma E stress response system has a negative effect on curli production. As described above, ECA mutants that induce RpoE make less curli than a WT strain (Fig 2.7). Defects in LPS biosynthesis also result in RpoE induction and less curli (119). Inner core LPS mutants such as *rfaC, rfaD,* and *galU* have drastically altered outer membrane protein profiles (84, 109), increased RpoE dependent transcription (51, 59, 76), and little curli production. Additionally, mutations in *tolA* and *pal* have more RpoE and are defective for curli (127). The Cpx two-component system, which negatively regulates curli specific genes (27, 53), is also induced in LPS and ECA mutants (25, 59). However, the LPS mutant strain *rfaC* had only increased sigma E dependent transcription (59). Finally, the *rseA* strain, which lacks the anti-sigma factor of RpoE, produces less curli and is light pink on CR plates (Fig. 2.7C). Thus RpoE appears to have a role in modulating curli production even in cases where RpoS levels are quite high such as in *rfaD* and *galU* strains (10, 52). The sigma E stress response may function as a check to prevent curli fiber formation during outer membrane stress. In a similar fashion, the sigma E inducing strains *galU* and *rfaP* produce less flagella and type I pili (32, 61, 84). However, the mechanism for the effect of RpoE on curli production is unknown; RpoE is not predicted or shown to directly affect *csg* gene transcription (95).

Several genes involved in quorum sensing and virulence were found to affect curli production including *qseB, qseC, aaeR, lsrF, sdiA,* and *flgM*. QseB and QseC had different effects on curli production (Tables 2.5 and 2.6) and were opposite those published elsewhere in a uropathogenic strain of *E. coli* (62). However, our *qseB* strain had both light pink and dark red colonies suggesting suppressor mutations (Table 2.6). The Keio *flgM* strain overproduced flagella and made less curli consistent with the antagonistic relationship between these two extracellular appendages (87). However, *sdiA,* which should also overproduce flagella(28), had higher levels of curli production (Table 2.5).

To further explore the intersection of flagella and curli we compared our results with other screens involving the Keio collection. Inoue et al (2007) screened the Keio collection for defects in swarming motility using Eiken Agar (43, 46) and subsequently checked the swarming mutants for reduced swimming motility. Using GeneVenn (88), we compared swarming and swimming with curli associated genes (See Fig 2.13). Very few genes were found to overlap, especially between swimming motility and curli production. More than half the genes that affect swimming motility and curli encode for either ATP synthase or LPS biosynthesis genes (Table 2.10). Another screen looked for biofilms defective mutants in the Keio collection using crystal violet and 96-well plates; biofilms mutants were subsequently tested for the flagella, type I pili, and curli production (80). Comparison of flagella, curli, and type I genes in these strains also identified little overlap between flagella and curli associated genes. However, there were much fewer curli genes identified in their screen, perhaps due to the use of LB media in the initial biofilm screen (80). LB media has relatively high salt concentration which inhibits curli production (100). Additionally, the CR phenotypes presented here are different for several strains (Fig. 2.10, Table 2.7); for example, several inner core LPS mutants are listed as WT for CR binding (80). Consequently, we tested our strains using similar conditions; only a few are likely due to media, growth, or staining differences (Fig. 2.10, Table 2.7).

Curli requires much more than the dedicated Csg proteins encoded in the divergent operons *csgDEFG* and *csgBAC* (18, 41). In this work, we screened the Keio collection for additional factors that affected curli production. More than 300 genes were found to be associated with curli production in *E. coli* K-12, adding greatly to the number of known factors. Several factors were shown to increase curli production including those that lead to higher RpoS levels. However, mutants that are known to induce the sigma E stress response produced less curli. Comparison of genes associated with curli and motility indicates little overlap in the metabolism and regulation of these two processes. The quantity, diversity, and specificity of function of curli associated genes suggest curli formation is a complex and highly regulated developmental pathway in *Escherichia coli.*

Materials and Methods

Bacterial strains and growth.

This study used the Keio collection, a systematic collection of nonessential single gene knockouts in BW25113 made with the Datsenko and Wanner method (26). The collection was shipped grown on Lysis Broth (LB) agar plates. Freezer stocks made with LB with 20% glycerol were maintained in 96-well microplates at -80°C. Bacteria were grown for 48 h at 26°C on YESCA plates (1 g yeast extract, 10 g Casamino Acids, and 20 g agar per liter) for cell hydrophobicity, western blotting, and electron microscopy. Curli production was monitored by growing organisms under similar conditions on 20mL Congo red (CR) indicator plates (YESCA plates with 50 g/mL CR and 10 g/mL Coomassie Brilliant Blue (CBB)). Colonization factor antigen (CFA) agar (1.5g yeast extract, 10g Casamino Acids, and 20g agar per liter containing 0.4mM $MgS0_4$ and 0.04mM $MnCl_2$ buffered to pH 7.4) with 100g/mL CR or 50g/mL CR and 10g/mL CBB was tested against CR indicator plates (Fig S8). Antibiotics were added when appropriate to the following concentrations: 25 μg/mL kanamycin; 25 μg/mL chloramphenicol; or 100 μg/mL ampicillin.

Screening of curli associated genes.

The Keio collection was shipped as individual colonies grown on LB plates. To screen the collection we used sterile toothpicks and plate bolt replicators to copy the collection onto CR indicator plates amended with 25 μg/mL kanamycin. Following growth for 48 h at 26°C the strains were scored for color. If a colony was pink or darker red than the surrounding strains it was restreaked for single colonies. The CR phenotype of each strain was verified by comparison to BW25113. To emphasize CR phenotypes, levels were uniformly adjusted by setting the gray point to a clear spot on the red agar in Adobe Photoshop

SDS-PAGE and Western blotting.

Bacteria were scraped off YESCA plates and resuspended in PBS (pH 7.4) before normalization by optical density at 600 nm. To allow CsgA solubility, samples were briefly treated with formic acid as described (23). Whole cell samples were electrophoresed on 13% sodium dodecyl sulfate (SDS)-polyacrylamide and blotted onto polyvinylidene difluoride using standard techniques. CsgA and CsgG polyclonal antibodies were raised in rabbits with the purified proteins (Proteintech, Chicago, IL) and were used at 1:10,000 and 1:100,000 dilutions, respectively. The FliC antibodies were kindly provided by Harry Mobley and were used at a 1:5000 dilution. The secondary antibodies were anti-rabbit antibodies conjugated to horseradish peroxidase (Sigma, St. Louis, MO) and were used at a dilution of 1:10,000 for both CsgA and CsgG and a 1:5000 dilution for FliC. Western blots were developed using the Pierce super signal detection system.

PCR verification of strains with multiple CR phenotypes.

Strains with multiple phenotypes were chosen for PCR verification. Primers were designed 2-300 bp upstream of each gene and are listed in Table 2.6. Each strain with an altered CR phenotype was struck from the Keio collection to make a clean freezer stock. At least two independent colonies for each phenotype were resuspended in sterile water and subjected to colony PCR. The individual colony mixtures were added to master mixes of GoTaq Flexi (Promega). Mixture A contained the appropriate upstream primer and primer K1 (4) and was used to verify the location of the Kanamycin insert in the genome. Mixture B contained the primers KT and K2 (4) and was used to verify the presence of the insert. If each phenotype gave a positive PCR product for both mixtures then it was added to Table 2.6. Mutations in energy production, coenzyme metabolism, cell envelope biogenesis, and DNA repair genes were more likely to have more than one CR phenotypes (Table 2.6).

Cell hydrophobicity.

Cell hydrophobicity was measured as described (103) using Xylene (X5-1; Fisher Scientific Co., Pittsburgh, PA), which gave more consistent results than a mixture of linear hexanes (136). The assay was performed with 2mL of cells resuspended to OD600 in PBS pH 7.4 amended with 2M ammonium sulfate (101, 102) which helps distinguish between the relatively hydrophilic K-12 strains . Cells were tested in 16x125mm glass culture tubes overlaid with increasing amounts of Xylene. Following incubation for 10 minutes at room temperature, the mixtures were vortexed for 1 minute using a Barnstead Thermolyne 16700. Following phase separation for 15 minutes at room temperature, a sample from the lower aqueous layer was removed and measured at OD600. Each strain was independently tested at least twice.

Silver stained LPS PAGE gels.

LPS was extracted as described (72) from 10 OD600 of cells grown on YESCA plates for 48 h at 26°C. Samples were normalized to 10 ng of KDO per lane using the Thiobarbituric Acid Assay and were separated and silver stained in a 14% Tricine SDS PAGE gel (72).

Motility assay.

Cells were grown overnight in YESCA with appropriate antibiotics. Saturated cultures were diluted 1/100 in fresh YESCA and grown to mid log phase (OD600 of 0.3-0.6). Strains were normalized to 0.2 OD600 in YESCA and 2uL was inoculated into 0.2% Agar YESCA motility plates. Plates were grown for 8h at 26°C. The strains tested for motility were also tested for growth rates in YESCA at 26°C using a Klett meter. No appreciably growth rate differences were measured.

Electron microscopy.

Bacteria were grown on YESCA plates for 48 h at 26°C. Samples were resuspended in phosphate buffered saline (PBS) and stained with 2% uranyl acetate as previously described (18). Grids were viewed using a Phillips CM10 microscope.

Online databases.

Intergenic region distance was gathered from Ecogene's Intergene Regions (105); intergenes were excluded.

Figure Legends

Fig. 2.1. Overview of screen for mutants affecting curli production. (A-B) Congo Red (CR) plates grown at 26C for two days. (A) Screen of Keio collection inoculated with bolt replicator: JW5581 (*ubiE*), JW1022 (*csgE*), JW1024 (*csgB*), JW1025 (*csgA*), JW0192 (*rcsF*), and JW0230 (*crl*). The collection was screened three times by both individual streaking and replica plating (B) Associated phenotype scoring: 1 = JW1025 (*csgA*), 2 = JW4130 (*hfq*), 3 = JW0663 (*nagA*), 4= JW3148 (*greA*), 5= BW25113 (WT), and 6 = JW3883 (*cpxR*). (C) Whole cell Westerns of strains in Fig. 2.1B. (D) Distribution of CR phenotypes based on scoring from white to dark red.

Fig. 2.2. Lipopolysaccharide (LPS) mutants. (A) CR YESCA plates grown at 26C for two days: BW25113 (WT), JW1025 (*csgA*), JW3603 (*rfaB*), JW3596 (*rfaC*), JW3594 (*rfaD*), JW3024 (*rfaE*), JW3595 (*rfaF*), JW3606 (*rfaG*), JW3818 (*rfaH*), JW3602 (*rfaI*), JW3601 (*rfaJ*), JW3597 (*rfaL*), JW3605 (*rfaP*), JW3607 (*rfaQ*), JW3604 (*rfaS*), JW3600 (*rfaY*), JW3599 (*rfaZ*), JW0212 (*lpcA*), and JW0196 (*gmhB*). (B) LPS structure. Blue lines represent CR defective LPS mutants. Red lines represent LPS genes not in the Keio collection. (C) Whole cell Westerns of LPS mutants. All samples were scraped from YESCA plates grown at 26C for two days and treated with formic acid.

Fig. 2.3. LPS structure and function. (A) Silver stain of a 14% PAGE Tricine gel of LPS mutant and Whole Cell Western blots. (B) BW25113 (WT), JW1025 (*csgA*), JW1224 (*galU*), JW3596 (*rfaC*), JW3595 (*rfaF*), and JW3606 (*rfaG*) grown on CR YESCA for 2 days. A sterile loop was scraped through each strain to show friable phenotypes of some LPS mutant strains. (C) Cell hydrophobicity of LPS mutant strains in phosphate buffered saline (PBS) amended with 2 M ammonium sulfate. Similar results were obtained in indepent replicates (data not shown).

Fig. 2.4. The sodium antiporter gene *nhaA* is required for curli production. (A) Strains were grown on Congo Red (CR) plates at 26C for two days. (B) Western blots probed

with anti-CsgG and anti-CsgA antibodies. (C) YESAC motility plates were inoculated with 2 µL from a 0.5 OD shaking culture of YESCA cells and then grown for 8 hours at 26C. (D-F) EM images from cells grown for 26C for two days on YESCA plates, resuspended in PBS, and stained with 2% uranyl acetate. Scale bar equals 1µM. (D) WT at 15,500X (E) *csgA* at 15,500X (F) *nhaA* at 15,500X. (G) Western blot probed with anti-FliC antibodies.

Fig. 2.5. Distribution of the molecular functions of genes affecting curli production. Genes affecting curli function were separated according to the molecular function (See Table 2.5) using COG families. Genes in bold produce more curli than wild type bacteria.

Fig. 2.6. Congo Red and cellulose phenotypes of different *E. coli* strains. Strains were grown 2 days at 26C on YESCA plates that were amended with (A) Congo Red or (B) Calcuflour. A black light was used to show Calcufluor fluorescence. The UTI89 isolate was used as a positive control for cellulose production.

Fig. 2.7. Congo red phenotypes of Enteric Common Antigen (ECA) and *rseA* gene deletions. (A) Curli deficient strains: *rffA, rffC, rffT, wxzE*. Strains with increased curli production: *rfe.* (B) Diagram of Enteric Common Antigen biosynthetic pathway. (C) A *rseA* deletion strain was defective for Congo Red binding.

Fig. 2.8. Comparison of the length of intergenic regions in *Escherichia coli* K-12. (A) Histogram of the length of divergent intergenic regions. (B) Histogram of the length of all intergenic regions including divergent, convergent, and codirectional. Intergenic regions were downloaded from Ecogene (105) and can be found at http://www.ecogene.org/InterGeneTable.php.

Fig. 2.9. Low ppGpp strains and *dksA* mutants produce less curli. Congo red binding and Western blots of strains deficient for producing ppGpp. Background is BW25113 (4) unless otherwise noted as MG1655. Strain *relA*' is shorthand for BW25113 *relA:kan* in

which the Kanamyacin cassette has been excised using pCP20 which encodes Flp recombinase (26). The *ΔspoT207::cat* was moved into the BW25113 *relA*' strain by P1 transduction; however, the *relA'* strain is more defective than the *relA::kan* strain. Duplicate strains of *relA'* and *relA' ΔspoT207::cat* were independent isolates. Strains *relA', relA' ΔspoT207::cat*, MG1655 *ΔrelA251::kan,* and MG1655 *ΔrelA251::kan ΔspoT207::cat* were from Janine Maddock. Strains were grown on YESCA plates for two days at 26C.

Fig. 2.10. Differences in Congo Red phenotypes of Keio collection strains on YESCA and CFA plates and CFA plates without Coommassie Brilliant Blue counterstain. Niba et al 2007 (80) performed a screen of the Keio collection for biofilm defective mutants. The biofilm mutants were subsequently tested for their motility, Type I pili formation, and curli production. For curli production they grew strains on CFA agar with twice the usual amount of Congo Red and no Coommassie Brilliant Blue (CBB) counterstain. Some of their CR phenotypes which were scored as +, -, or ± were different than ours. We tested these strains on YESCA and CFA plates with CR and CBB and on CFA plates without CBB and twice the CR as in Niba et al 2007 (80). Sets (A) (B) and (C) were scored based on their CR phenotypes from 1-6 at 24, 48, and 72 hours at 26C (See Table 2.7). Shown here is 48 hours growth. Some of the phenotypic differences are due to plating differences. CBB aids in detecting subtle differences between strains.

Fig. 2.11. Distribution of Curli Specific Genes in other organism according to String. Found primarily in the genomes of Enterobacteriaceae, *csg* genes are also present in different arrangements and combinations in other Gammaproteobacteria families including Pseudomonadaceae, Shewanellaceae, and Vibrionaceae as well as in a few distantly related bacteria in the Bradyrhizobiaceae, Burkholderiaceae, and Flavobacteriaceae families (49, 118). Darker colors indicated higher similarity to the Csg proteins in *E. coli* K-12 than the lighter pinks.

Fig. 2.12. Rescue of Congo Red binding in some *pyr* mutants when struck near other strains. When struck near other strains several pyrimidine biosynthesis mutants from the Keio collection displayed increased Congo Red binding. This was seen for *pyrB, pyrD, pyrE, pyrF,* and *carB.* In many cases the 'donating' strain did not matter; *rpoS,* BW25113, and various *pur* mutants were able to rescue *pyrE.* Exceptions include the effects of the lack of donation from *purD* to *pyrB* and *carB,* and the fact that other *pyr* mutants were unable to donate the missing compound; however, *pyrE* was able to donate to *carB.*

Fig. 2.13. Venn diagrams demonstrating few genes associated with curli affect swarming or swimming. Comparison of genes found to affect curli with swimming and swarming associated genes. (A) Excludes and (B) includes the less defective swarming mutants found in Table S2 of Inoue et al (46). Sections of Venn diagrams: **a** (orange) curli associated genes that do not affect swarming or swimming motility; **b** (dark green) curli associated genes that affect swarming motility; **c** (dark blue) curli associated genes that affect swimming motility; **d** (light blue) swimming defective genes that do not affect curli; **e** (light green) swarming defective genes that do not affect curli. Venn diagram is drawn to scale with volumn fairly representative of number of genes in each section.

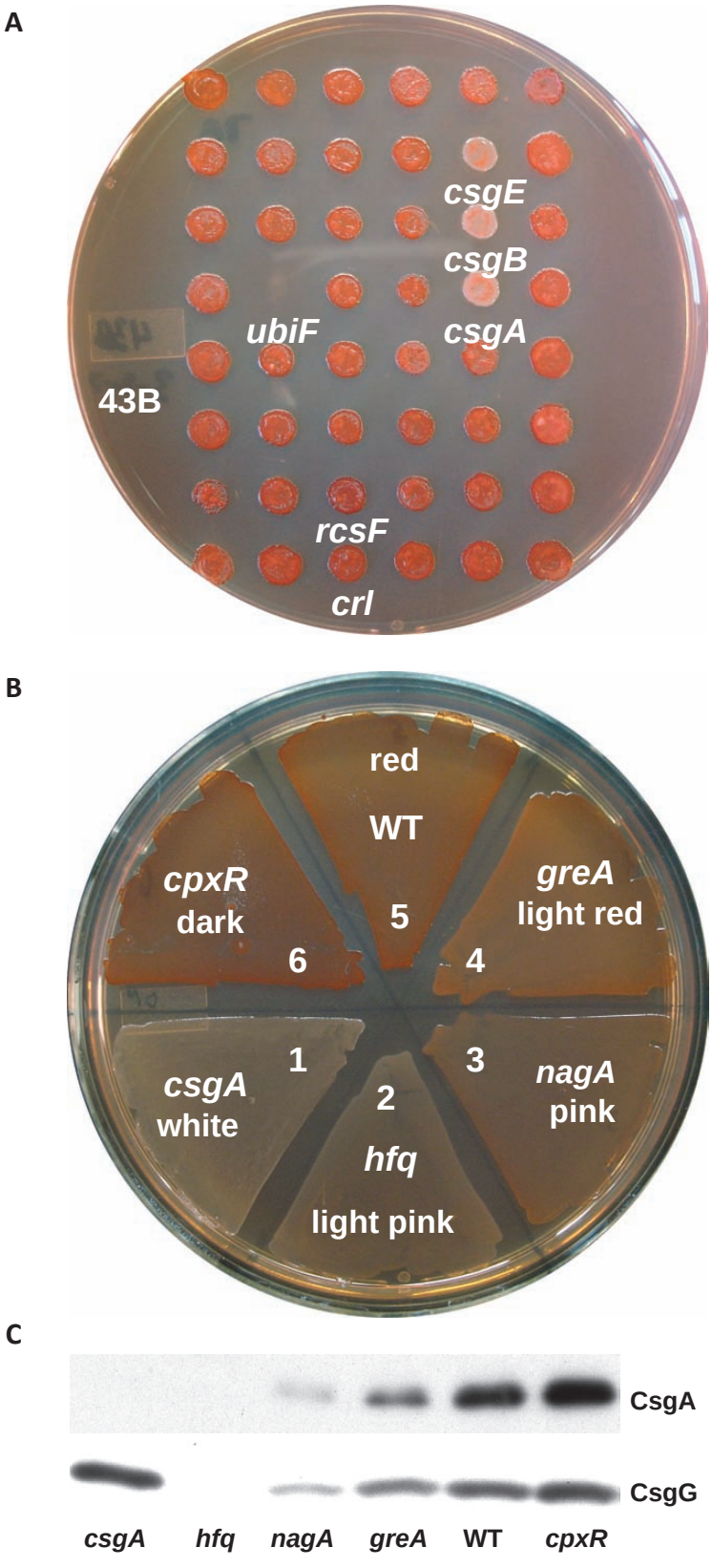


Figure 2.1. Overview of screen for mutants affecting curli production.

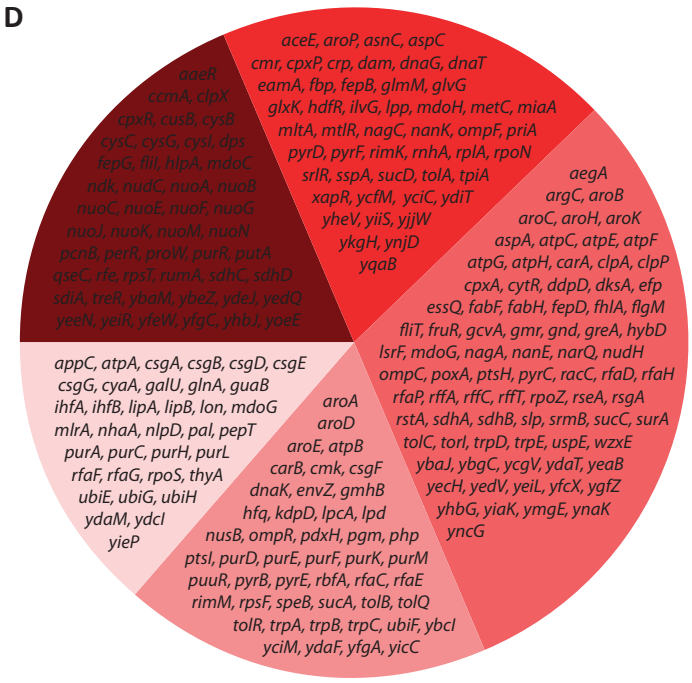


Figure 2.1 (continued). Overview of screen for mutants affecting curli production.

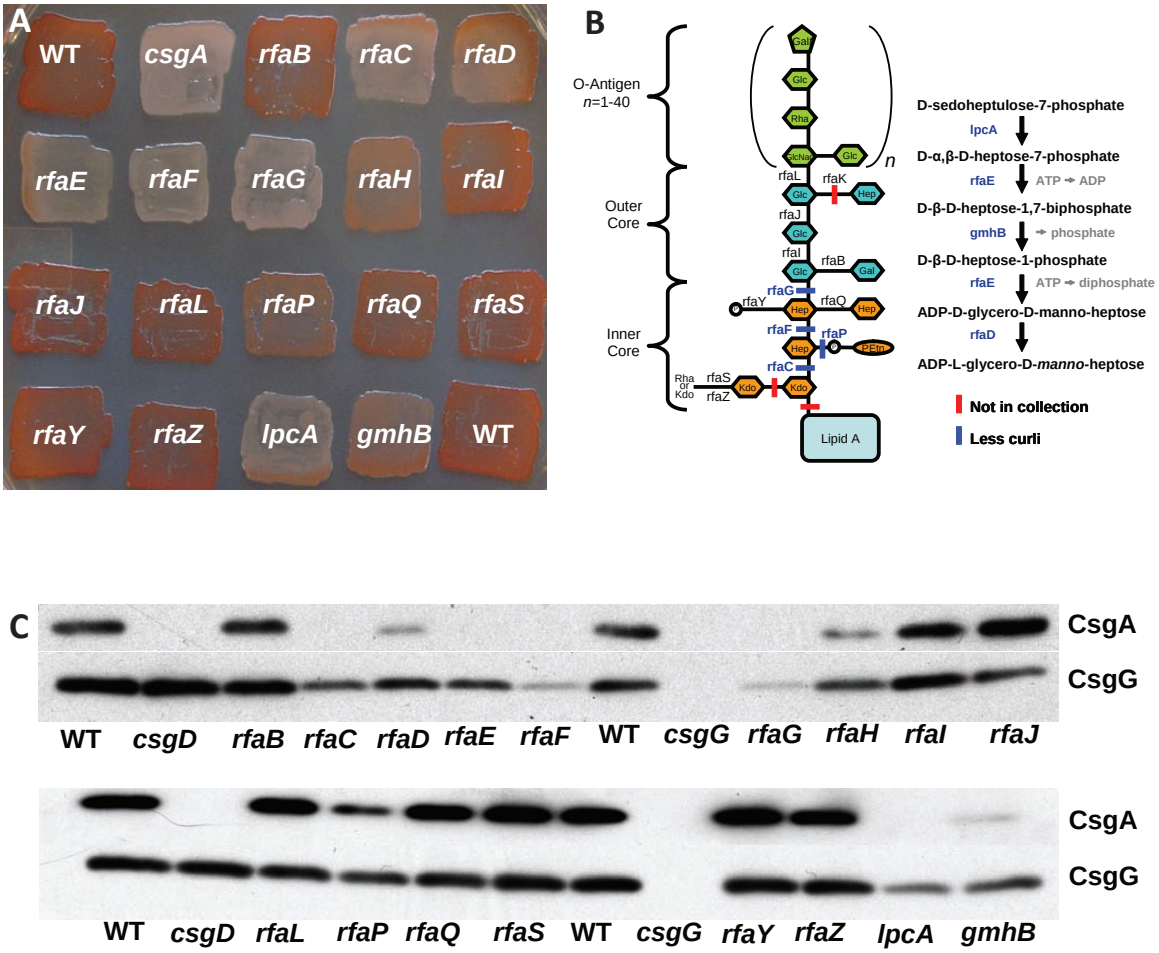


Figure 2.2. Lipopolysaccharide (LPS) mutants.

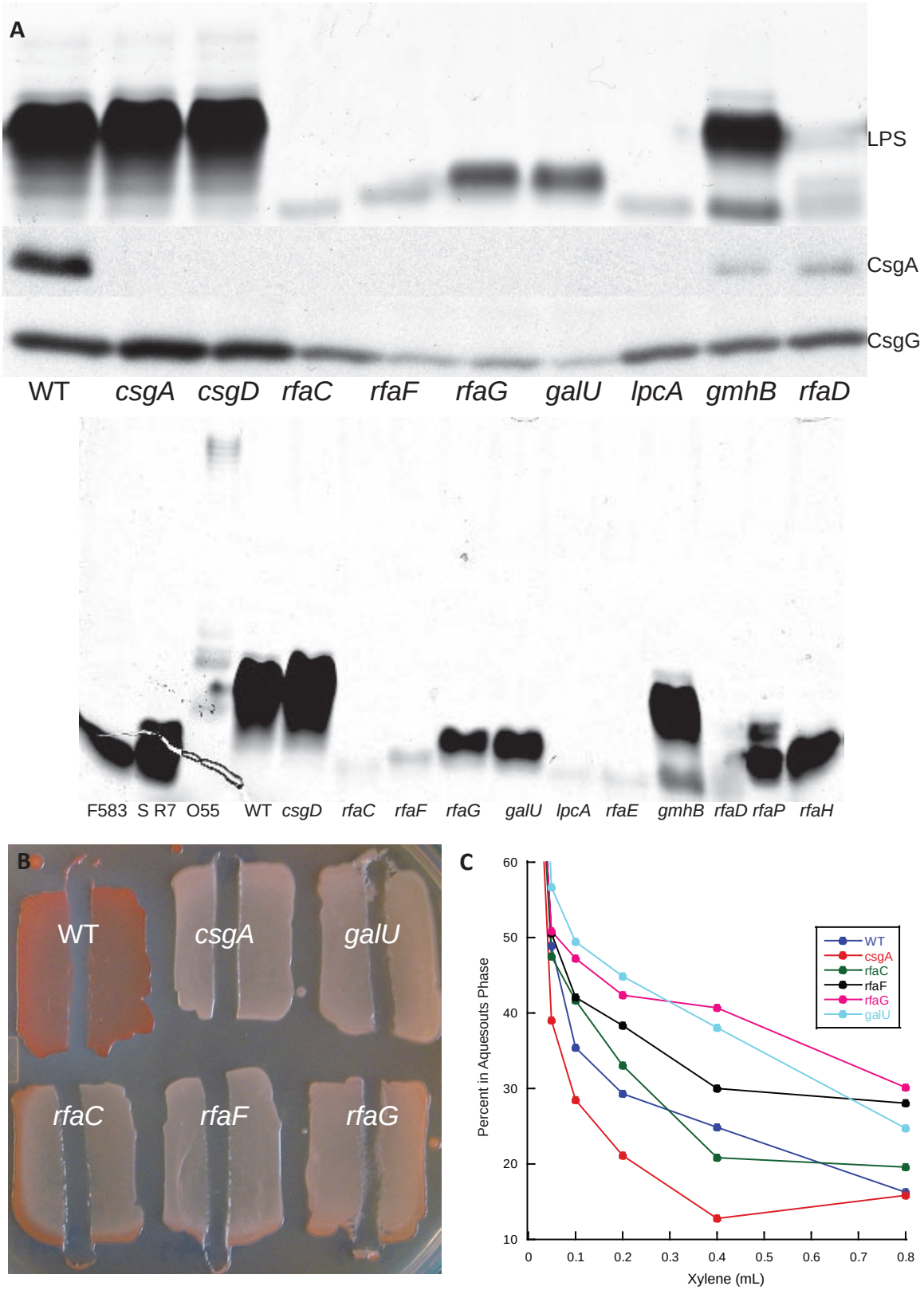


Figure 2.3. Phenotypes of deletions in inner core LPS biosynthetic genes.

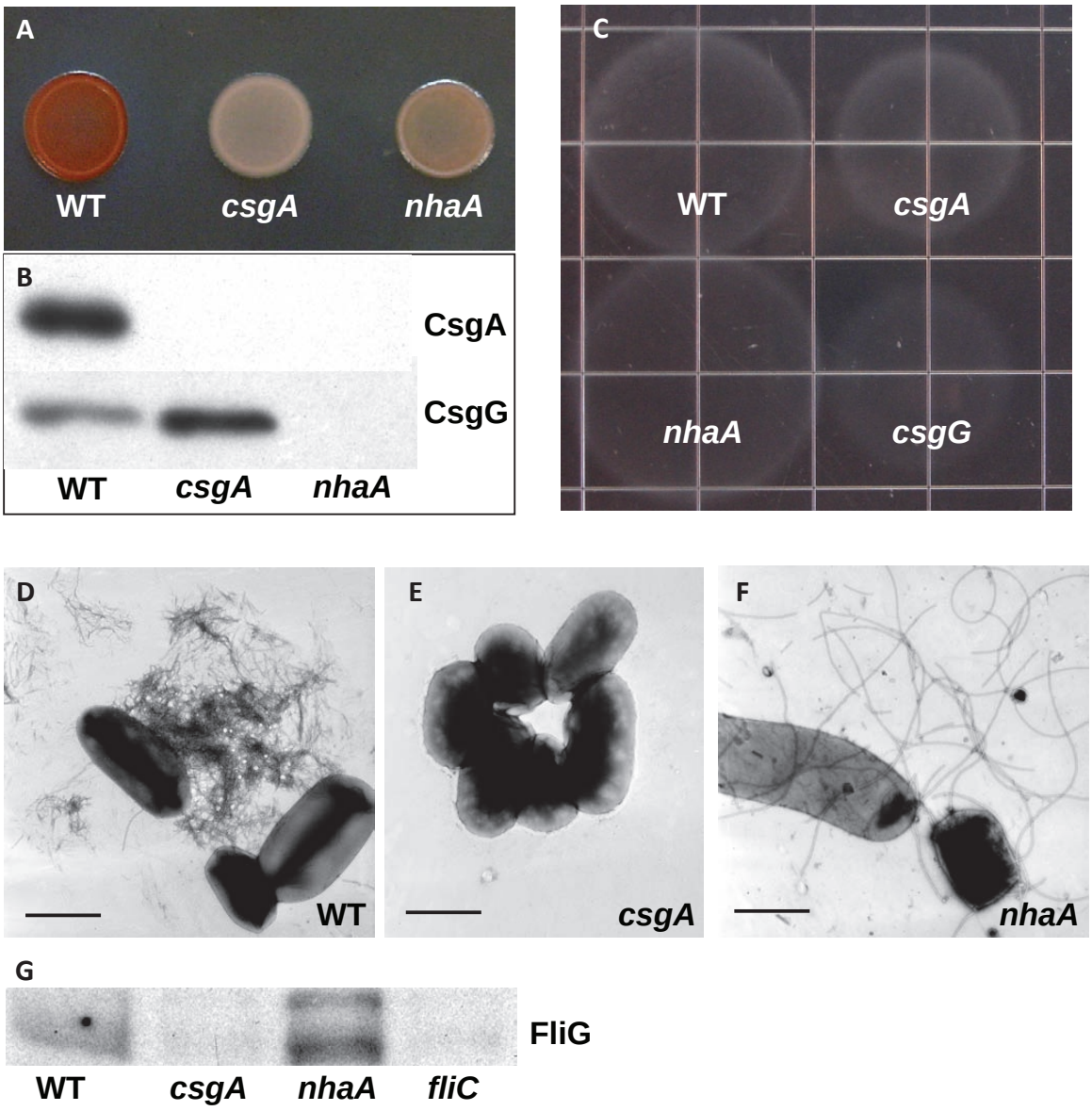


Figure 2.4. The sodium antiporter gene *nhaA* is required for curli production.

Figure 2.5. Distribution of molecular functions of genes affecting curli production.

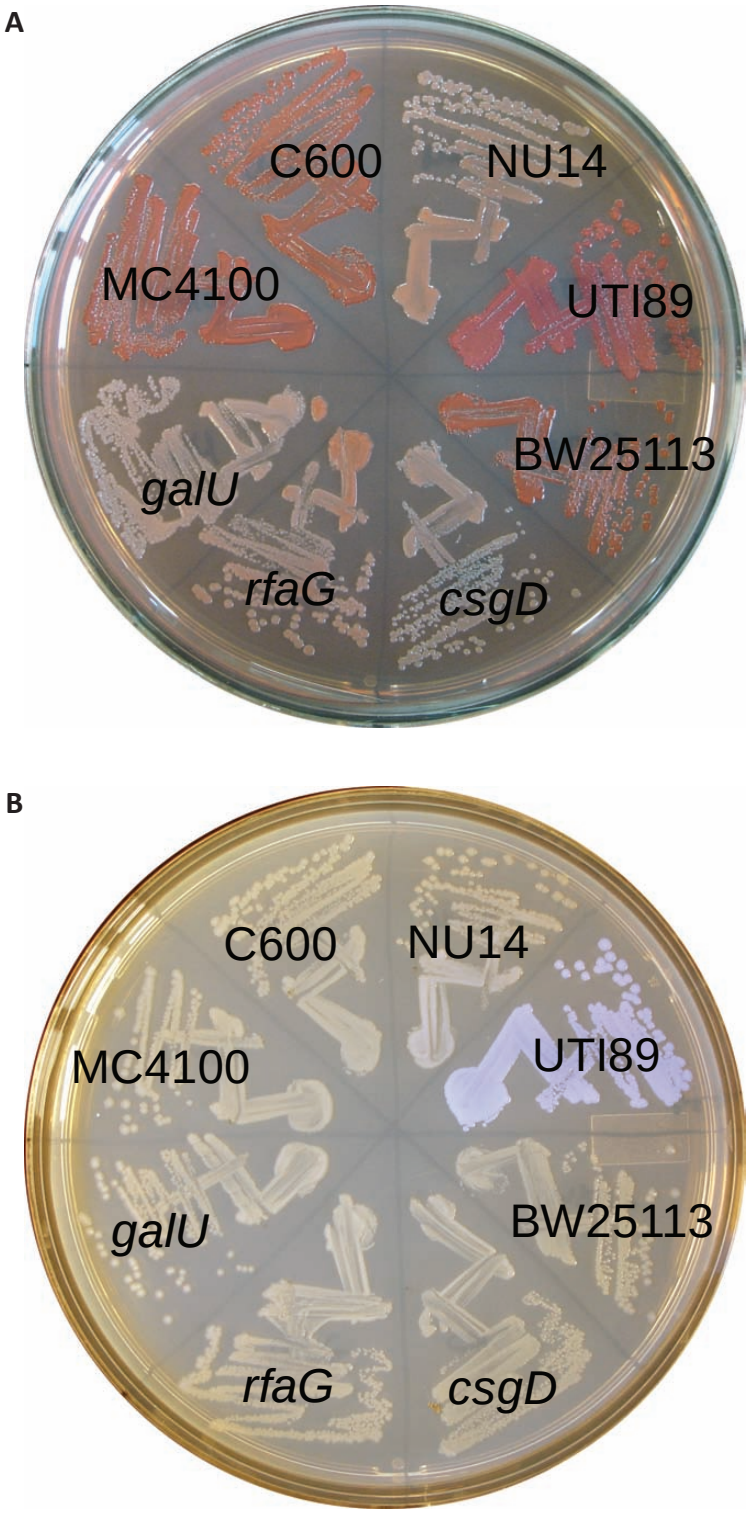


Figure 2.6. Congo Red and cellulose phenotypes of different *E. coli* strains.

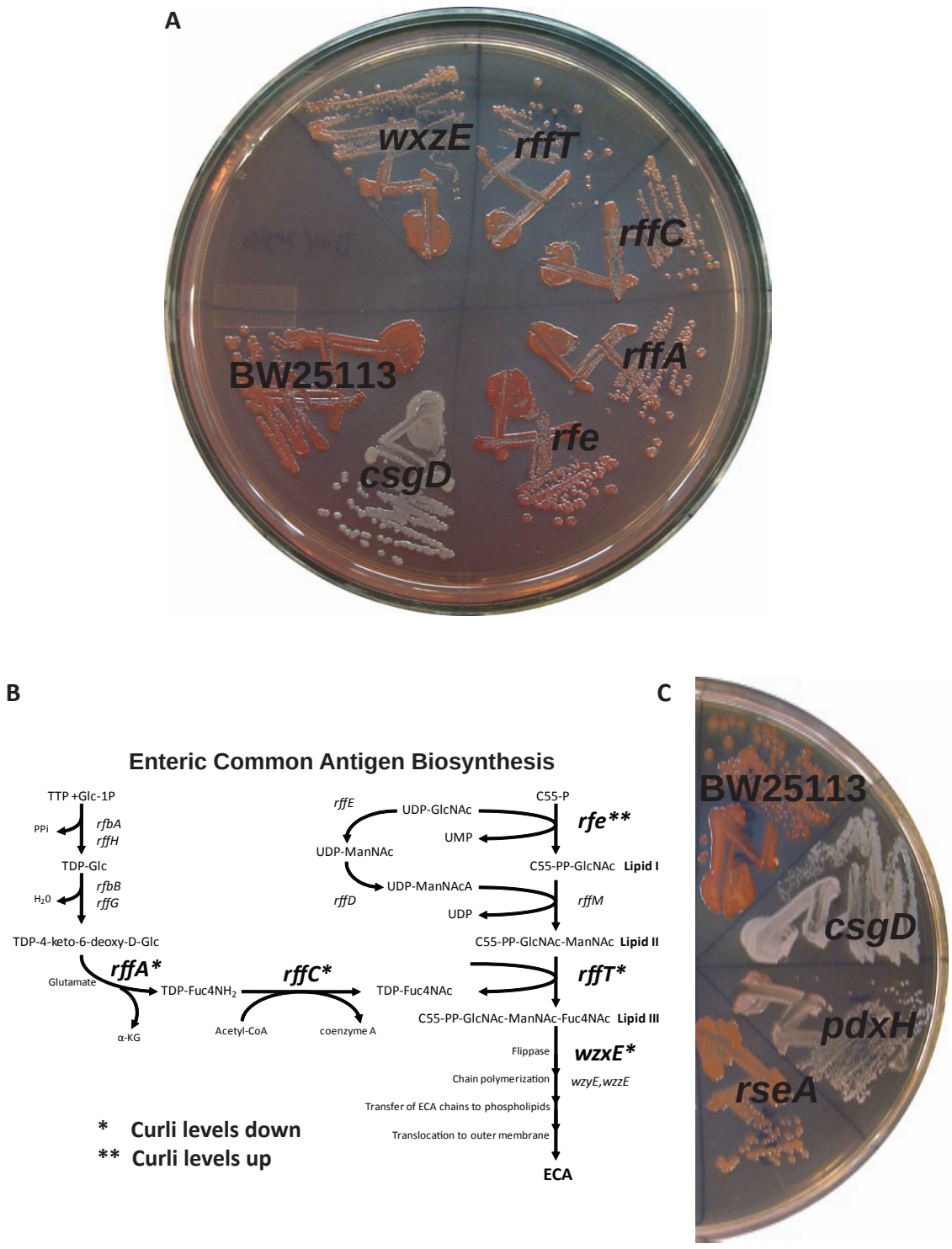


Figure 2.7. Congo red phenotypes of Enteric Common Antigen (ECA) and *rseA* gene deletions.

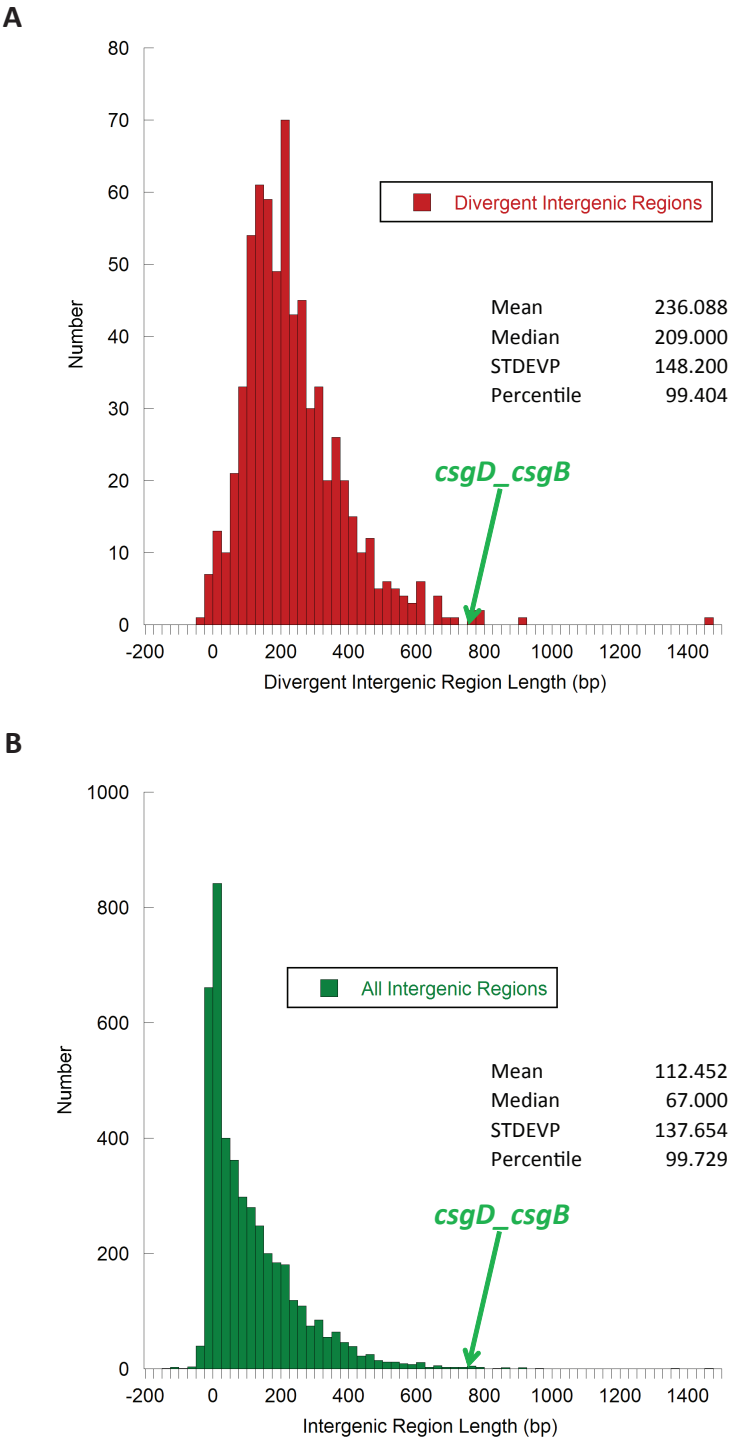


Figure 2.8. Comparison of the length of intergenic regions in *Escherichia coli* K-12.

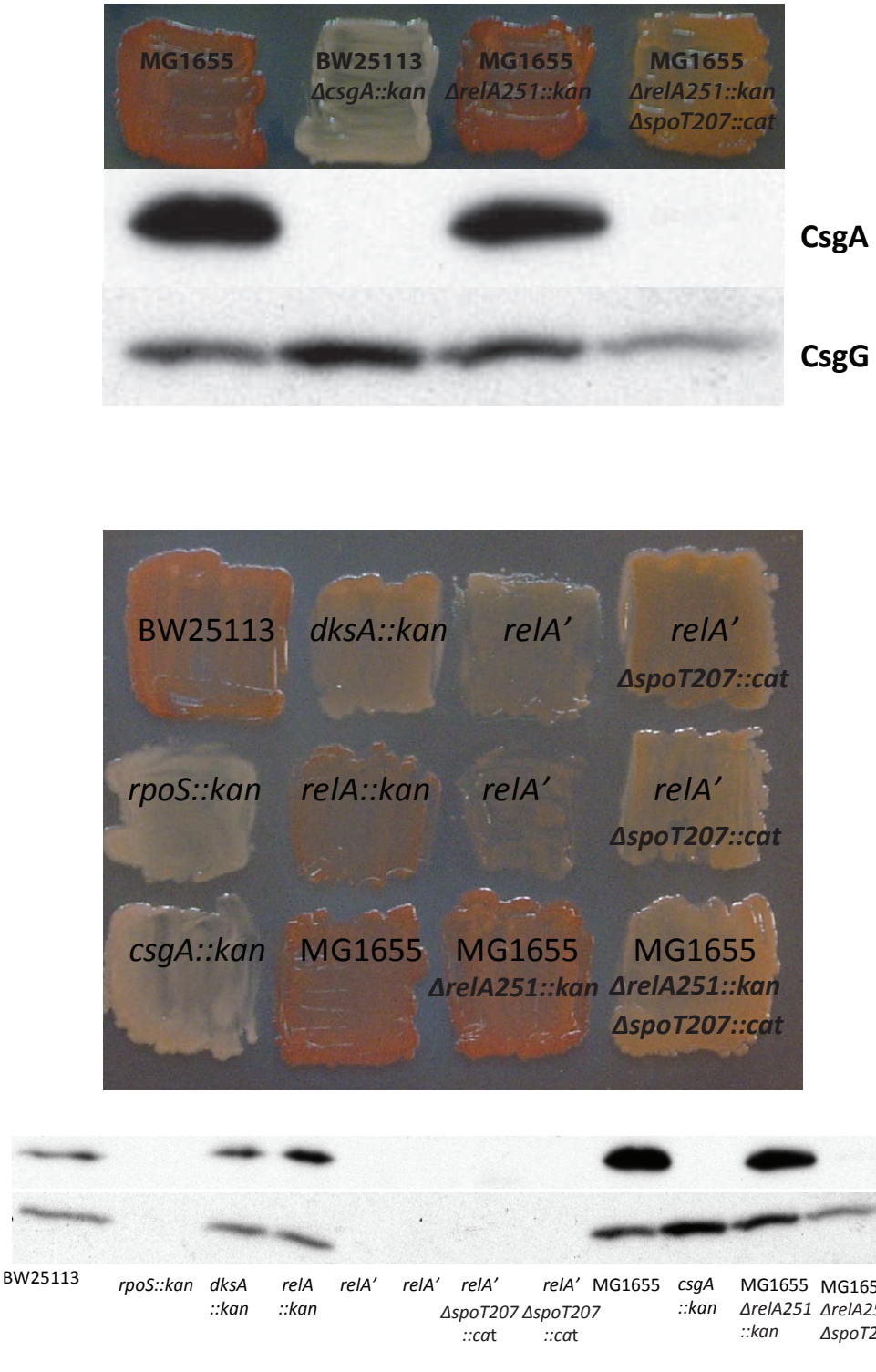


Figure 2.9. Low ppGpp and *dksA* mutants produce less curli.

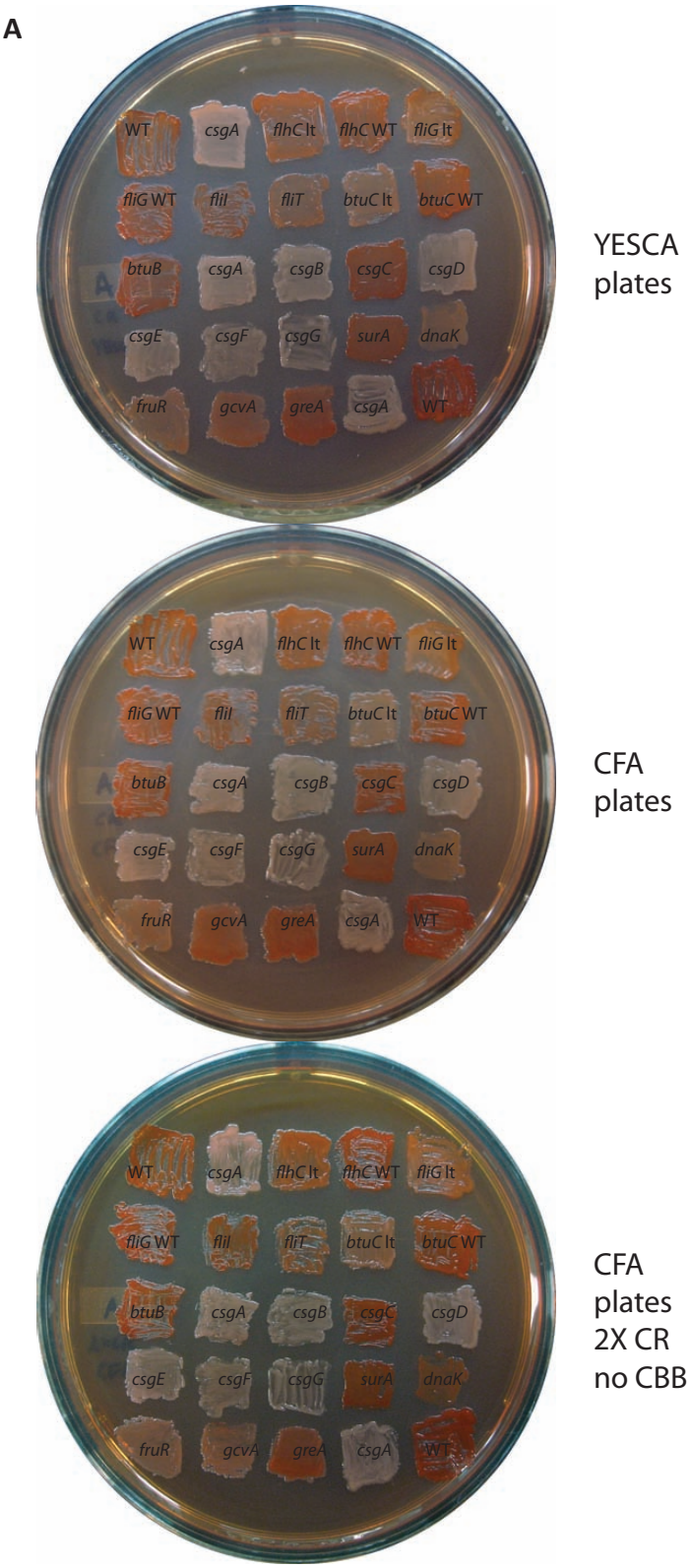


Figure 2.10. Differences in Congo Red phenotypes of Keio collection strains on YESCA and CFA plates and CFA plates without Coommassie Brilliant Blue counterstain.

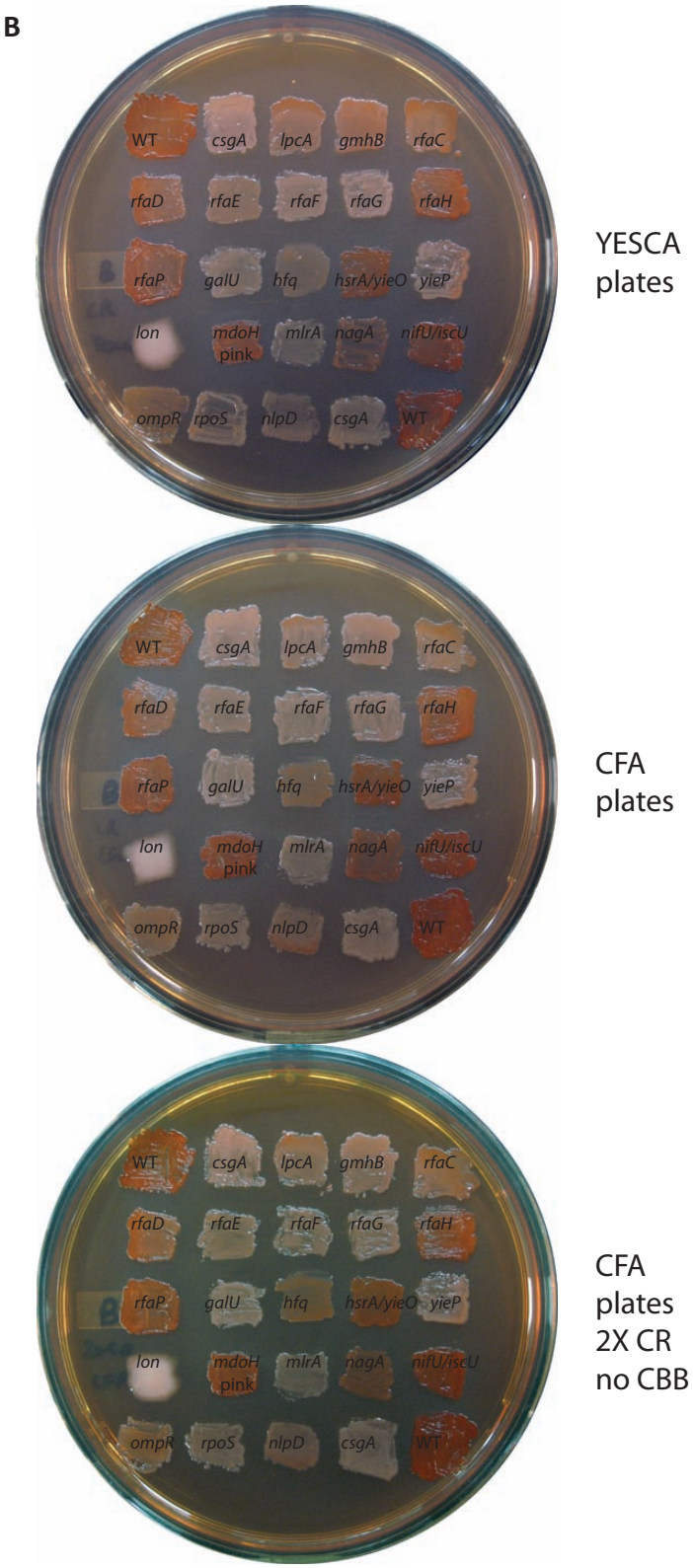


Figure 2.10 (continued). Differences in Congo Red phenotypes of Keio collection strains on YESCA and CFA plates and CFA plates without Coommassie Brilliant Blue counterstain.

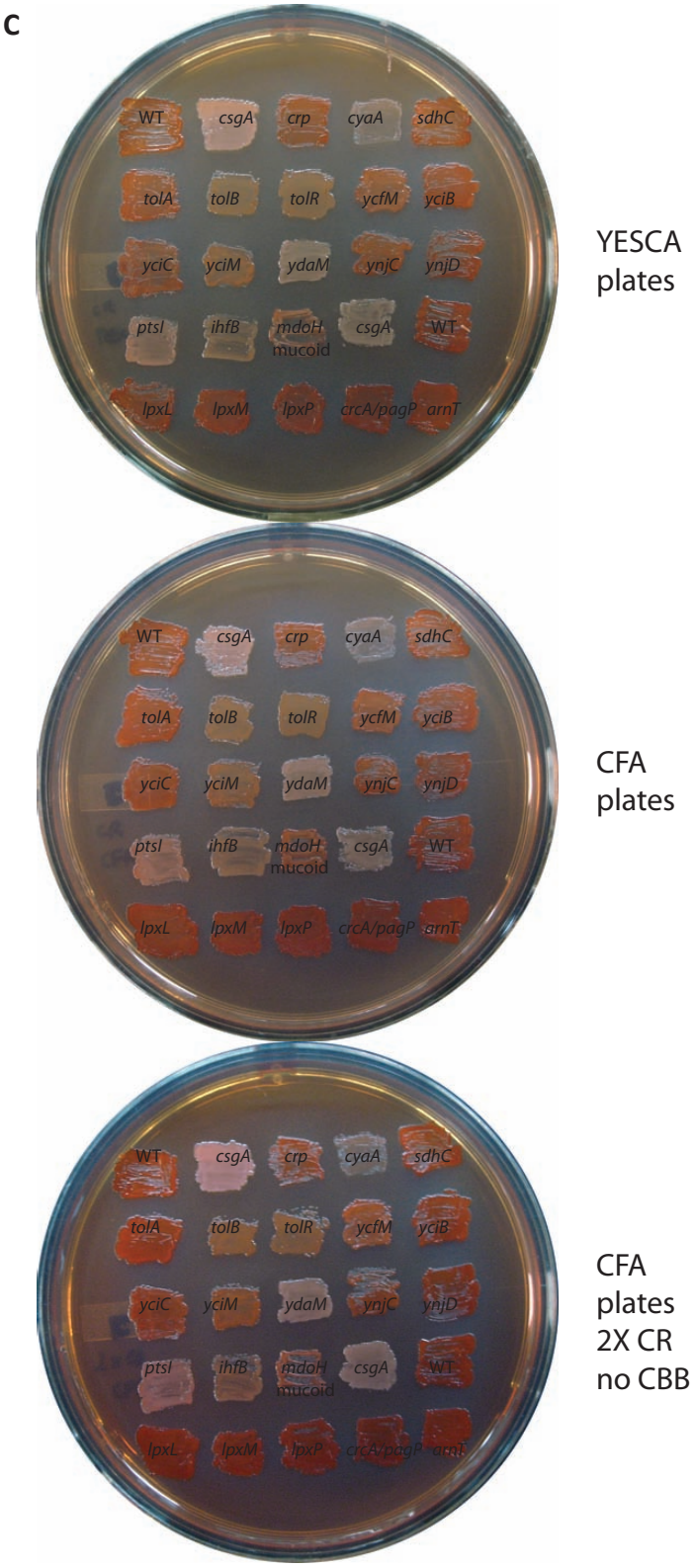


Figure 2.10 (continued). Differences in Congo Red phenotypes of Keio collection strains on YESCA and CFA plates and CFA plates without Coommassie Brilliant Blue counterstain.

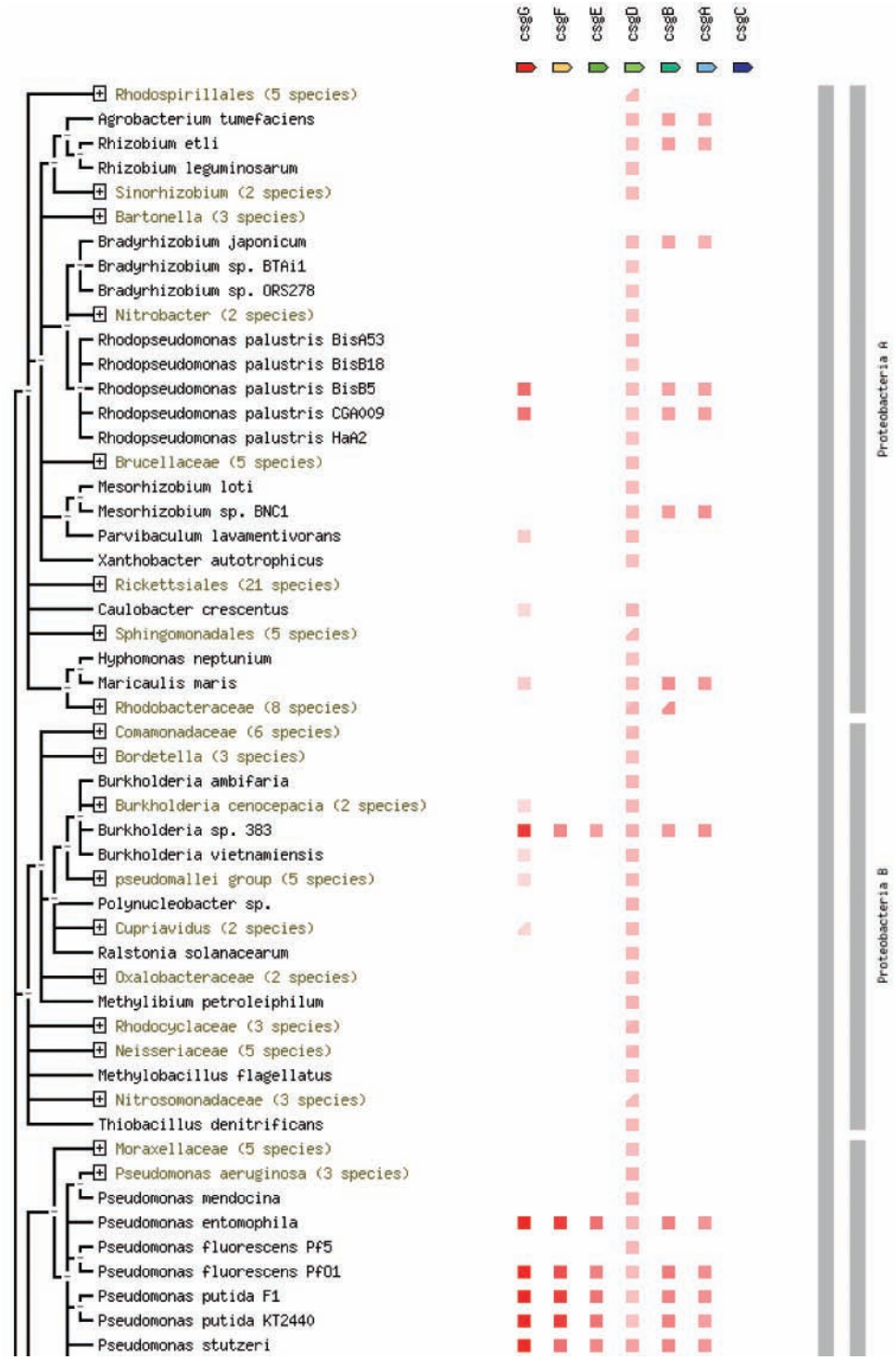


Figure 2.11. Distribution of Curli Specific Genes in other organism according to String.

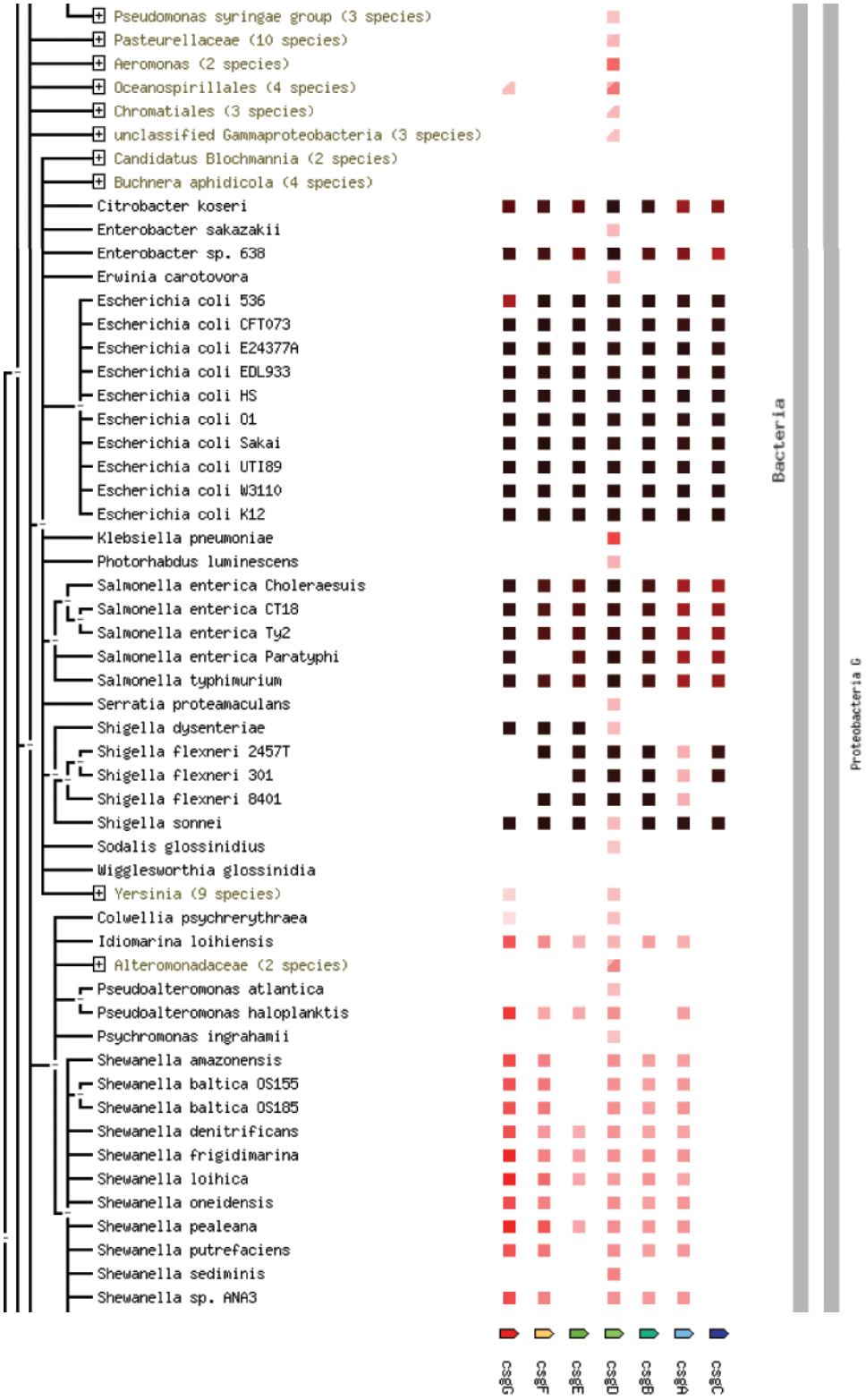


Figure 2.11. Distribution of Curli Specific Genes in other organism according to String.

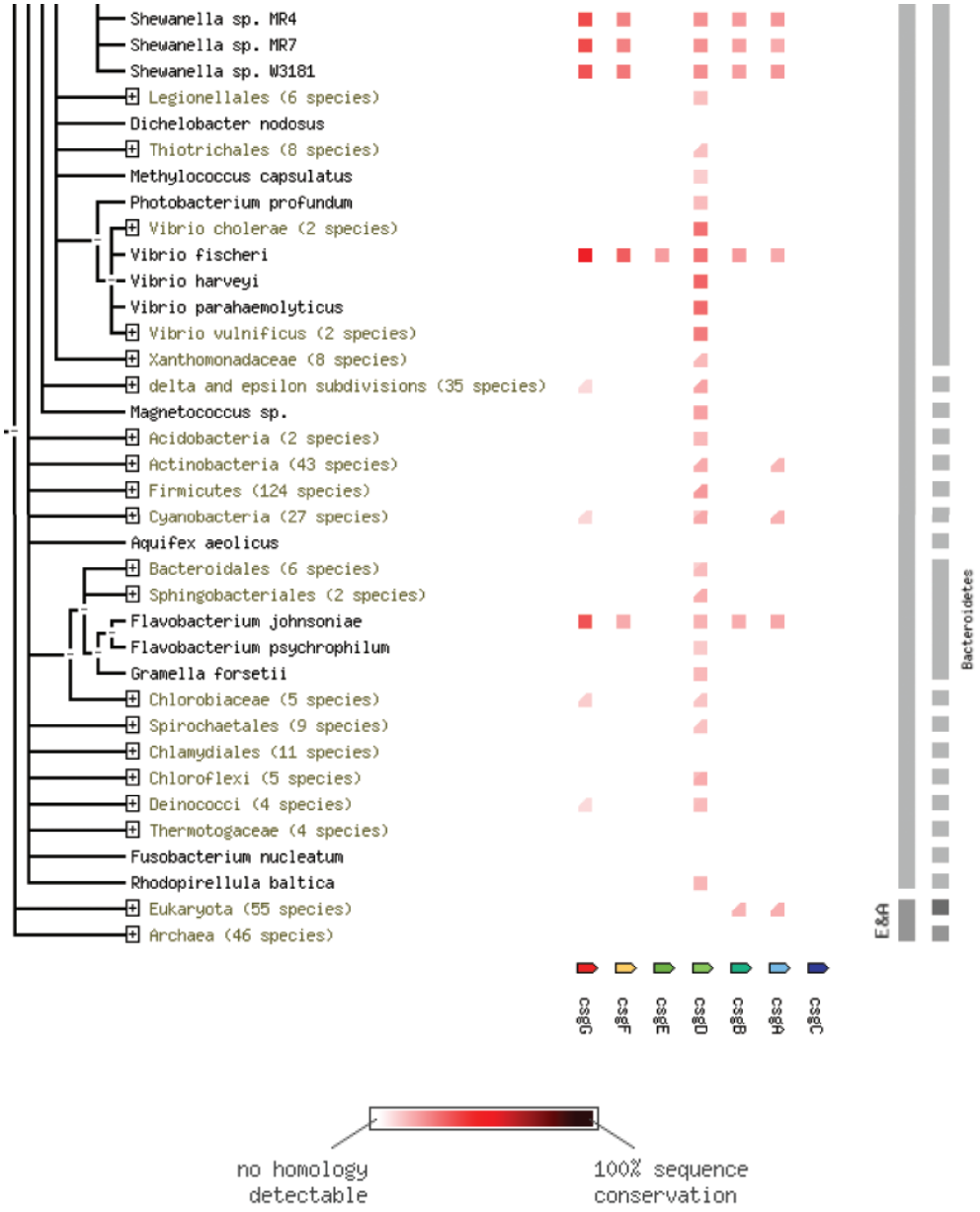


Figure 2.11. Distribution of Curli Specific Genes in other organism according to String.

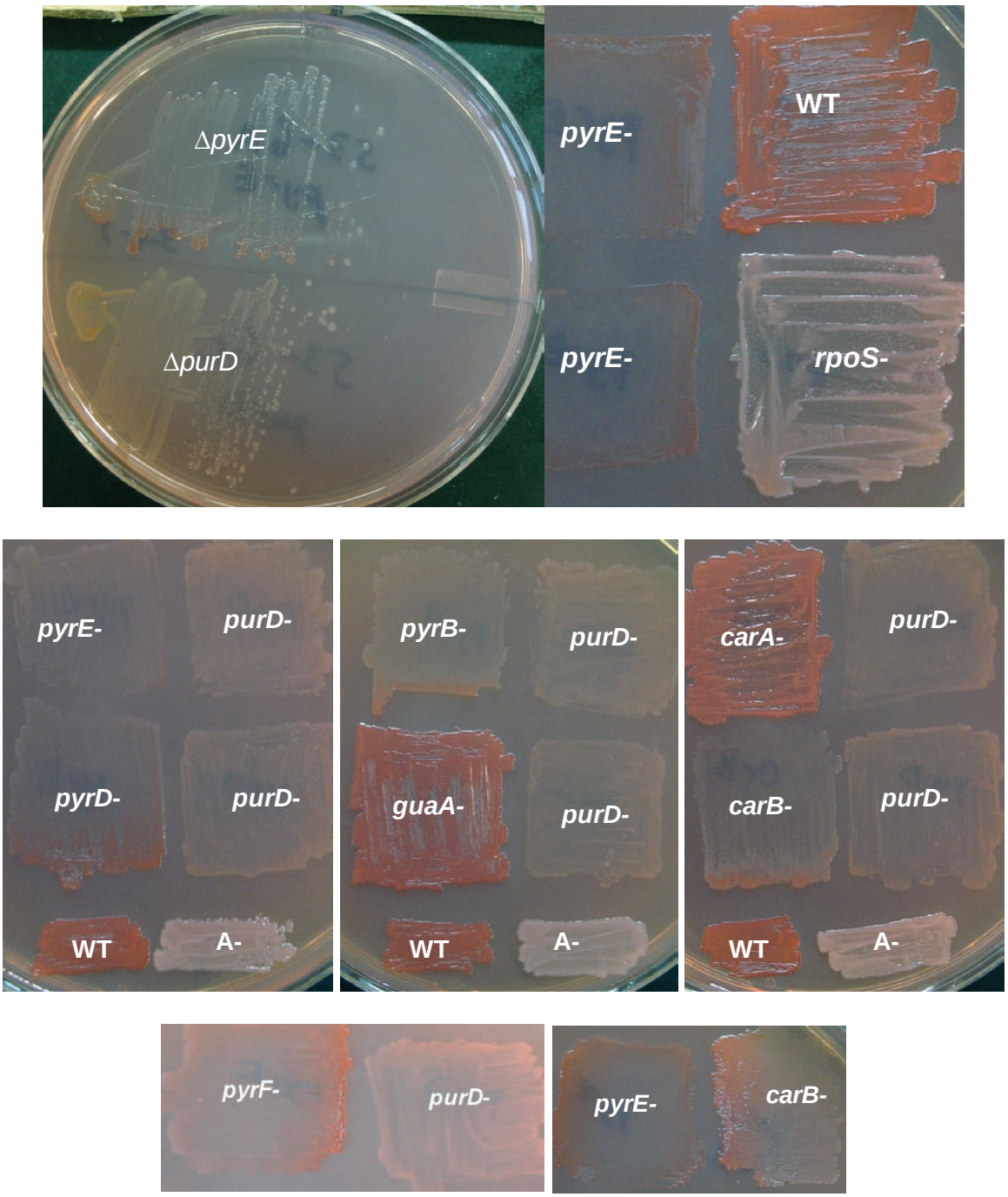


Figure 2.12. Rescue of Congo Red binding in some *pyr* mutants when struck near other strains.

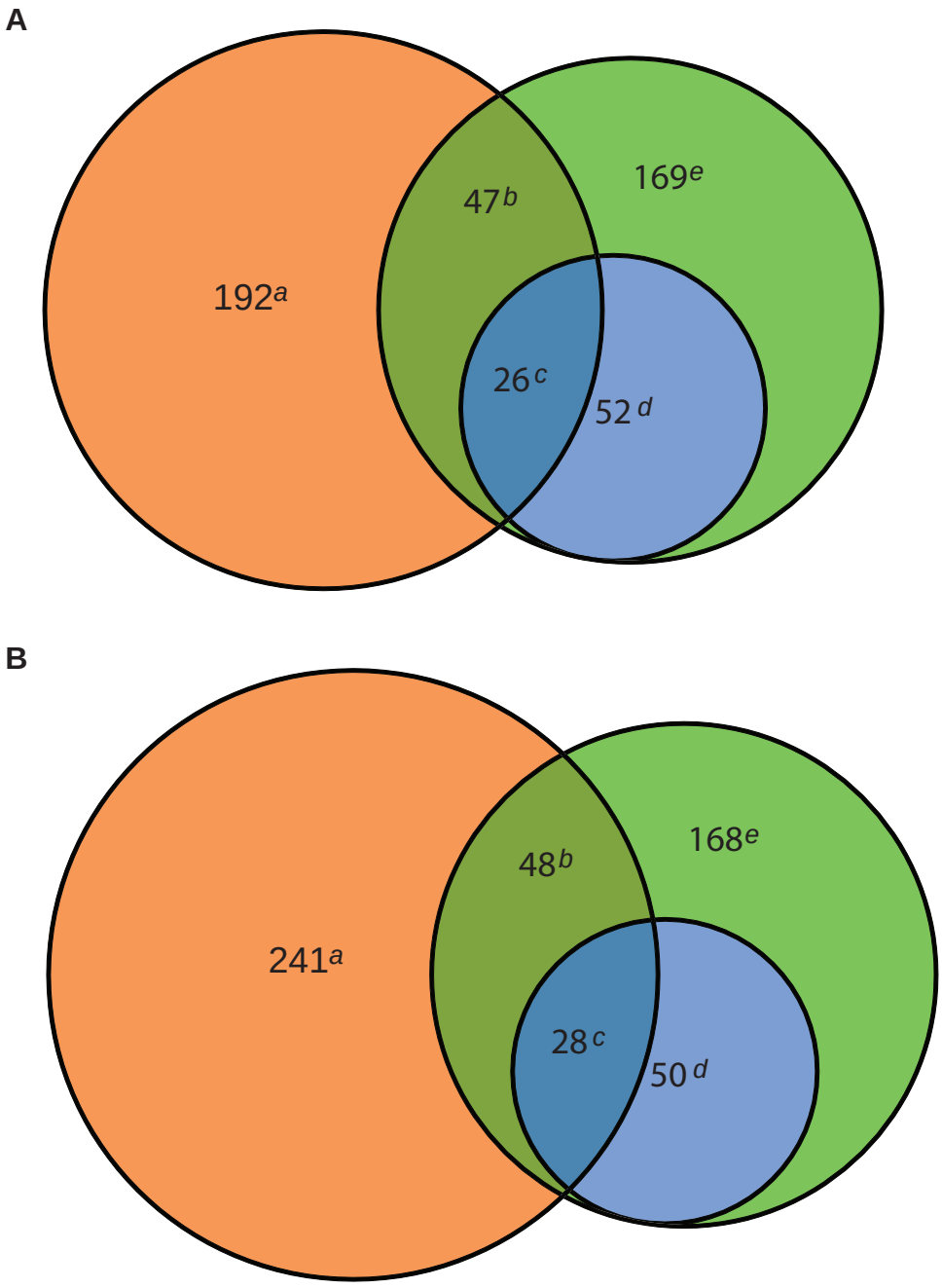


Figure 2.13. Venn diagrams demonstrating few genes associated with curli affect swarming or swimming.

Table 2.1. Genes required for or known to affect curli production

Gene	Effect on Curli production	Source	Color in screen
***cpxA**,R*	-	(27, 53)	WT, dark red
crl	+	(12, 90)	WT
csgA,B	Fiber subunits	(18, 41)	white
csgC	Fiber morphology [b]	(36)	WT
csgD	*csgBAC* transcription	(41)	white
csgE,F[a]*,G*	Secretion, chaperone	(18)	white
ddhC (rfbH)	+[d]	(1)	N/A [e]
dksA	+[o]	(40)	pink
envZ	+	(34, 53)	white
hns	variable[j]	(2, 34, 83)	WT
ihfA,B	+	(34)	light pink
fis	-[f]	(108)	WT
flhC,D	-	(87)	WT[l]
fliZ	-	(87)	WT
lon	+	(125)	mucoid white
lpp	+[f]	(123)	light red
lpxM (msbB)	+[f]	(58)	WT
mlrA	+	(17)	white
nagA [o]*,C*	+	(7, 40)	pink
ompR	+[o]	(34, 40, 53)	white
pal	+[c]	(125)	light pink
rcsA,B,C	-	(31, 53, 125)	WT[m]
rfaG	+[d]	(1)	light pink
rpoS	+	(2, 12, 83)	white
rssB	+[o]	(40)	pink & WT [n]
rstA,B	-	(81)	light red, WT
qseB,C	none, +	(62)	variable, dark red
sdiA	-	(66)	dark red
tolA,B,Q,R	+	(125)	light pink
ycfR	+	(136)	white & WT [g]
Cyclic-di-GMP (GGDEF or EAL)			
adrA (*yaiC*)	+[i]	(54)	WT[i]
ycgF	- (via Rcs)	(115, 121)	WT
***yciR* (*gmr*,STM1703)**	-	(113, 115, 131)	pink[h]
ydaM	+	(115, 131)	white
yeaP	+	(115)	WT
yedQ	Cellulose only[k]	(24, 115, 131)	dark red
***yegE* (STM2123)**	+	(54, 87, 115)	WT
yhdA	+	(115)	WT
***yhjH* (STM3611)**	-	(54, 87, 113, 115)	WT
***yjcC* (STM4264)**	-	(113)	WT
STM3388	+	(54)	N/A[e]
STM1827	-	(113)	N/A[e]
STM4551	+	(114)	N/A[e]

Table 2.1 (continued). Genes required for or known to affect curli production

Notes: Genes in bold differed from results found in the literature when applicable.
[a] Some curli production occurs but is not cell associated (79). [b] Larger, more stable fibers (36) . [c] Inferred from (125). [d] In *Salmonella enterica* serovar Typhimurium (DT104 Rv or LT2) (1). Abequose is not used for *E. coli* O-antigens (117). [e] No homolog in BW25113 (4, 56). [f] In *Eschrelichia coli* O157:H57 EHEC strain 4304. [g] Mostly white colonies with few red suppressors. [h] Pink on CR plates but WT levels of CsgG and CsgA. [i] AdrA is also regulated by CsgD (139); positive feedback on CsgD was seen in *S. Typhimurium* MAE52 but not UMR1(54). [j] (-) in *E. coli* K-12 (+) in *S. typhimurium*. [k] No effect when tested in *E. coli* W3110, a BW25113 relative (4, 48, 115). [l] *flhC* had mostly pinkish/light red colonies with a few WT colonies. [m] *rcsF* is dark red. [n] Mostly pink colonies with few red suppressors. [o] Found in screens for additional genes affecting curli production (40). Table excludes the RNAs *omrA* and *omrB* which postranscriptionally inhibit *csgD* translation (45).

Table 2.2. Cellular localization of identified gene products.

Location	Number (% tot.)	Gene(s)
Cytoplasmic	193 (75.6)	*aaeR, aceE, aegA, argC, aroA, aroB, aroC, aroD, aroE, aroK, asnC, aspA, aspC, atpA, atpC, atpG, atpH, carA, carB, ccmA, clpA, clpP, clpX, cmk, cpxR, crp, csgD, cyaA, cysB, cysC, cysG, cysI, cytR, dam, ddpD, dksA, dnaG, dnaK, dnaT, dps, efp, fabF, fabH, fbp, fhlA, flgM, fliI, fliT, fruR, galM, galU, gcvA, glmM, glnA, glxK, gmhB, gmr, gnd, greA, guaB, hdfR, hfq, hns, hybD, ihfA, ihfB, ilvG, lipA, lipB, lon, lpcA, lpd, lsrF, metC, miaA, mlrA, mtlR, nagA, nanK, ndk, nudC, nudH, nuoB, nuoC, nuoE, nuoF, nuoG, nusB, ompR, pdxH, pepT, perR, pgm, php, poxA, priA, ptsH, ptsI, purA, purC, purD, purE, purF, purH, purK, purL, purM, purR, putA, puuR, pyrB, pyrC, pyrD, pyrE, pyrF, racC, rbfA, rfaC, rfaD, rfaE, rfaF, rfaG, rfaH, rfaP, rffC, rimK, rimM, rnhA, rplA, rpoN, rpoS, rpoZ, rpsF, rpsT, rsgA, rstA, rumA, sdhA, sdhB, sdiA, speB, srlR, srmB, sspA, sucA, sucC, sucD, thyA, torI, tpiA, treR, trpA, trpB, trpC, trpD, trpE, ubiE, ubiF, ubiG, ubiH, uspE, xapR, ybaJ, ybaM, ybeZ, ybgC, ycgV, ydaF, ydaM, ydaT, ydcI, ydeJ, ydiT, yeaB, yecH, yeeN, yeiL, yfcX, ygfZ, ygiF, yhbG, yhbJ, yheV, yiaK, yicC, yieP, yiiS, yjjW, ynaK, yncG, ynjD, yoeE, yqaB*
Periplasmic	11 (4.3)	*cpxP, csgE, csgF, fepB, hlpA, mdoG, surA, tolB, yciM, yfeW, yfgC*
Integral Membrane Proteins	32 (12.5)	*appC, atpB, atpE, cmr, cpxA, eamA, envZ, essQ, fepD, fepG, kdpD, mdoC, mdoH, narQ, nhaA, nuoA, nuoJ, nuoK, nuoM, nuoN, proW, qseC, rseA, sdhC, sdhD, tolQ, wzxE, ybcI, yciC, yedQ, ykgH, ymgE*
Membrane Anchored	7 (2.7)	*atpF, cusB, glvG, tolA, tolR, yedV, yfgA*
Outer membrane Lipoproteins	7 (2.7)	*csgG, lpp, mltA, nlpD, pal, slp, ycfM*
Outer Membrane ß-barrel proteins	3 (1.2)	*ompC, ompF, tolC*
Extracellular	2 (0.8)	*csgA, csgB*

Table 2.3. Curli associated genes and their primary cellular role.

Function	COG Group	Number (% tot.)	Genes
Information storage			
Translation, ribosome structure, & biogenesis	J	13 (5.0)	*efp, miaA,* ***pcnB****, poxA, rbfA****, rimK****, rimM, rplA, rpsF,* ***rpsT****, rsgA, rumA, srmB*
Transcription	K	30 (11.5)	***aaeR****,* ***asnC****, csgD,* ***cysB****, cytR, flgM, fruR, gcvA, greA, hdfR, hns, mlrA, mtlR, nusB,* ***perR****,* ***purR****, puuR, rfaH, rpoN, rpoS, rpoZ, rstA, sdiA, srlR, torI,* ***treR****, xapR, ydcI, yieP, ynaK*
DNA replication, recombination, and repair	L	8 (3.1)	*dam, dnaG, dnaT, ihfA, ihfB, priA, rnhA, yeaB*
Cellular processes			
Cell envelope biogenesis, outer membrane	M	34 (12.2)	*csgA, csgB, csgE, csgF, csgG,* ***cusB****, galU, gmhB, hlpA, lpcA, lpp,* ***mdoC****, mdoG, mdoH, mltA, nlpD, ompC, ompF, pal, rfaC, rfaD, rfaE, rfaF, rfaG, rfaP,* ***rfe****, rffA, rffC, rffT, slp, wzxE, ycgV,* ***yfeW****, yhbG*
Cell motility & secretion	N	6 (2.3)	***fliI****, tolA, tolB, tolC, tolQ, tolR*
Posttranslational modification, protein turnover, chaperones	O	11 (4.2)	***ccmA****, clpA, clpP,* ***clpX****, dnaK, fliT, lon, sspA, surA, yjjW, yncG*
Inorganic ion transport & metabolism	P	10 (3.8)	*cmr,* ***cysC****,* ***cysI****, ddpD,* ***dps****, fepB, fepD,* ***fepG****, nhaA,* ***yoeE***
Signal transduction mechanisms	T	18 (6.9)	*cpxA, cpxP****, cpxR****, crp, dksA, envZ, gmr, kdpD, narQ, ompR,* ***qseC****, rseA, uspE****, ybeZ****, ydaM,* ***yedQ****, yedV, yeiL*
Metabolism			
Energy production & conversion	C	33 (12.6)	*aceE, aegA, appC, atpA, atpB, atpC, atpE, atpF, atpG, atpH, hybD, lpd,* ***nuoA****,* ***nuoB****,* ***nuoC****,* ***nuoE****,* ***nuoF****,* ***nuoG****,* ***nuoJ****,* ***nuoK****,* ***nuoM****,* ***nuoN****, putA, racC, sdhA, sdhB,* ***sdhC****,* ***sdhD****, sucA, sucC, sucD, ydiT, yiaK*
Amino Acid transport & metabolism	E	24 (9.2)	*argC, aroA, aroB, aroC, aroD, aroE,aroH, aroK, aroP, aspA, aspC, carA, carB, eamA, glnA, metC, pepT,* ***proW****, speB, trpA, trpB, trpC, trpD, trpE*
Nucleotide transport & metabolism	F	20 (7.7)	*cmk, cyaA, guaB, ndk,* ***nudC****, purA, purC, purD, purE, purF, purH, purK, purL, purM, pyrB, pyrC, pyrD, pyrE, pyrF, thyA*
Carbohydrate transport & metabolism	G	15 (5.7)	*fbp, glmM, glvG, glxK, gnd, lsrF, nagA, nagC, nanE, nanK, pgm, ptsH, ptsI, tpiA, yciM*
Coenzyme metabolism	H	8 (3.1)	***cysG****, lipA, lipB, pdxH, ubiE, ubiF, ubiG, ubiH*
Lipid metabolism	I	3 (1.1)	*fabF, fabH, yfcX*
Poorly characterized			
General function prediction only	R	15 (5.7)	*essQ, hfq, ilvG, nudH, php, ybcI, ybgC, ycfM, ydeJ,* ***yeiR,*** *yfgC, ygfZ****, yhbJ****, ynjD, yqaB*
Unknown/Other	S,U,V	13 (5.0)	*ybaJ, ybaM, yciC, ydaF, ydaT, yecH, yeeN, yfgA, yheV, yicC, yiiS, ykgH, ymgE* *******add cysE and update*******

Table 2.4. Phenotypes of mutants in genes for purine and pyrimidine biosynthesis.

Gene	CR phenotype	CgG Levels	CsgA Levels
BW25113	red	++++	++++
purA	white	+/-	-
purB	essential[a]	ND	ND
purC	white	+/-	-
purD	white	-	-
purE	light pink	-	-
purF	white	-	-
purH	white	+/-	-
purK	light pink	+/-	-
purL	white	-	-
purM	light pink	-	-
purN	light pink and red[b]	ND	ND
purR[d]	dark red	++++	+++++
purT	red	++++	++++
purU[c]	red	ND	ND
guaA	red	++++	++++
guaB	white	-	-
pyrB	light pink	+/-	-
pyrC	pink to light red	+++	++
pyrD	light red	+++	++
pyrE	light pink	+	+
pyrF	pink	+++	+
pyrG,H	essential[a]	ND	ND
pyrI	red	ND	ND
pyrL[d]	red	ND	ND
carA	pink	+++	+
carB	light pink	+	+/-

[a] Not in Keio collection, essential (4) [b] See Table S2 [c] In formylTHF biosynthesis I pathway not *de novo* purine biosynthesis [d] Regulatory role

Table 2.5. Congo Red and Western blot data for Keio strains.

Name	Main COG Class	Echobase Echolocation	Original Color	Color @ 26C	Color @ 37C	Wester CsgG Level	Western CsgA Level
aaeR	K	cytoplasmic	dark red	4	3	++++	+++++
aceE	C	cytoplasmic	light red	4	3	+++	+++
aegA	C	cytoplasmic	pink to orange	3	3	+	++
appC	C	Integral Membrane Protein	white to lt pink	1	4	+/-	+/-
argC	E	cytoplasmic	light red	3	4	+	+/-
aroA	E	cytoplasmic	white to lt pink	2-3,2	2,3	++,++	+/-,+/-
aroB	E	cytoplasmic	white to lt pink	2-3,4	3,2	+/-,+	-,+/-
aroC	E	cytoplasmic	pink to orange	4,3	2,3	+,++	+/-,++
aroD	E	cytoplasmic	white to lt pink	1,2	2-3,3	+,+/-	+/-,-
aroE	E	cytoplasmic	white to lt pink	2-3,2	2,3	++,++	+/-,+
aroH				3	4	+++	++
aroK	E	cytoplasmic	pink to orange	4,3	4,4	+++,+++	++++,++
aroP				4	4	++	+++
asnC	K	cytoplasmic	dark red	4,4	4,4	+++++,++	+++++,+++
aspA	E	cytoplasmic	pink to orange	3(SG)	4(SG)	++	+
aspC	E	cytoplasmic	pink to orange	4	2-3	+++	+++
atpA	C	cytoplasmic	slow growth/other	1,1(SG)	2,1(SG)	+/-,-	-,-
atpB	C	Integral Membrane Protein	slow growth/other	2,3(SG)	1,1(SG)	++,+	-,+++
atpC	C	cytoplasmic	slow growth/other	3,2-3(SG),3(SG)	2,1(SG),2(SG)	++++,+/-	++++,-
atpE	C	Integral Membrane Protein	slow growth/other	3,3(SG)	2,2(SG)	++,+++	++,+++
atpF	C	Membrane Anchored	slow growth/other	3,3(SG)	2,2(SG)	++,+++	+,++
atpG	C	cytoplasmic	slow growth/other	3,3(SG)	2,2(SG)	+++,+++	++++,++++
atpH	C	cytoplasmic	slow growth/other	3,3(SG)	1,1(SG)	+,+	+/-,+++
carA	E	cytoplasmic	light red	3	4	+++	+
carB	E	cytoplasmic	white to lt pink	2	5(SG)	+	+/-

Table 2.5 (continued). Congo Red and Western blot data for Keio strains.

Name	Main COG Class	Echobase Echolocation	Original Color	Color @ 26C	Color @ 37C	Wester CsgG Level	Western CsgA Level
ccmA	O	cytoplasmic	dark red	5	4	++	+++++
clpA	O	cytoplasmic	red	3	4	++++	+++
clpP	O	cytoplasmic	red	3	4	+++	+++
clpX	O	cytoplasmic	dark red	6	4	+++++	+++++
cmk	F	cytoplasmic	pink to orange	1 dry	4	-	-
cmk	'	'	pink to orange	2	4	+/-	+/-
cmr	P	Integral Membrane Protein	dark red	4	4	++	++
cpxA	T	Integral Membrane Protein	pink	3	4	++++	++++
cpxP	T	periplasmic*	light red	4	4	+++	++++
cpxR	T	cytoplasmic	dark red	6	4	+++++	++++
crp	T	cytoplasmic	light red	5-6,5-6	4,3-4	+++++,++	+++++,++++
csgA	M	Extracellular	white to lt pink	1	4-5	+++++	-
csgB	M	Extracellular	white to lt pink	1	4	+++++	-
csgD	K	cytoplasmic	white to lt pink	1	4	++++	-
csgE	M	periplasmic	white to lt pink	1	4	+	-
csgF	M	periplasmic	white to lt pink	2-3	4	++	+
csgG	M	Outer membrane Lipoprotein	white to lt pink	1	4	-	-
cusB	M	Membrane anchored	dark red	5-6	4	+++++	++++
cyaA	F	cytoplasmic	white to lt pink	1	3	+/-	-
cysB	K	cytoplasmic	dark red	6	3-4	+++++	++++
cysC	P	cytoplasmic	dark red	6,5	4,4	++++	+++++
cysE			dark red	5-6	3	++++	+++++
cysG	H	cytoplasmic	dark red	6	4	+++++	+++++
cysI	P	cytoplasmic	dark red	6,5	4,3	++++	+++++
cytR	K	cytoplasmic	white to lt pink	3, 4	4, 4	+++, +	+/-, ++

Table 2.5 (continued). Congo Red and Western blot data for Keio strains.

Name	Main COG Class	Echobase Echolocation	Original Color	Color @ 26C	Color @ 37C	Wester CsgG Level	Western CsgA Level
dam	L	cytoplasmic	pink to orange	4,4	4,4	++++,+++	++++,+++
ddpD	P	cytoplasmic	pink to orange	3	4	++	++
dksA	T	cytoplasmic	light red	3	3	++++	++
dnaG	L	cytoplasmic	pink to orange	4	4	++++	+++
dnaK	O	cytoplasmic	white to lt pink	2	6	+/-	-
dnaT	L	cytoplasmic	pink to orange	4	5	+/-	+/-
dps	P	cytoplasmic	dark red	6	4	++++	+++++
eamA	E*	Integral Membrane Protein	pink to orange	4	4	+++	++++
efp	J	cytoplasmic	pink to orange	3	4	++	+
envZ	T	Integral Membrane Protein	white to lt pink	2	4-5	+/-	+/-
essQ	R*	Integral Membrane Protein	pink to orange	3	4	++++	+
fabF	I	cytoplasmic	pink to orange	2	4	+/-	-
fabF	'	'	pink to orange	3	4	++	+
fabH	I	cytoplasmic	pink to orange	3	4	++	+++
fadJ	I	cytoplasmic	pink to orange	3	3	++++	++++
fbp	G	cytoplasmic	pink to orange	4,2	4,4	+++,++	+,++
fepB	P	periplasmic	light red	4	3	++	+++++
fepD	P	Integral Membrane Protein	white to lt pink	3	3	++	++++
fepG	P	Integral Membrane Protein	red	6	3	+++++	++++
fhlA	K	cytoplasmic	light red	3	4	+	+
flgM	K*	cytoplasmic	light red	3-4	4	++++	++
fliI	N	cytoplasmic	dark red	6,6,4	4,4,4	+++++,+++++,++++	++++,+,++++
fliT	O*	cytoplasmic	pink to orange	3,3	4,4	+++	+++
fruR	K	cytoplasmic	white to lt pink	3	4	+++	+/-
galU	M	cytoplasmic	white to lt pink	1	4	+/-	-

Table 2.5 (continued). Congo Red and Western blot data for Keio strains.

Name	Main COG Class	Echobase Echolocation	Original Color	Color @ 26C	Color @ 37C	Wester CsgG Level	Western CsgA Level
gcvA	K	cytoplasmic	white to lt pink	3	4	++	+++
glmM	G	cytoplasmic	light red	4,3-4	4,4	++,+++	++++,+++
glnA	E	cytoplasmic	slow growth/other	1(VSG),1(VSG),1(VSG)	1(VSG),1(VSG),1(VSG)	NA,NA,-	NA,NA,-
glvG	G	Membrane anchored	red	4,4	4,4	+++,+++	++++,++++
glxK	G	cytoplasmic	light red	4	4	+++	+++
gmhB	M*	cytoplasmic	white to lt pink	1,2-3,2-3	4,ND,ND	++,+,+	+,-,+
gmr	T	cytoplasmic	red	3	4	++++	++++
gnd	G	cytoplasmic	pink to orange	3	4	+	++
greA	K*	cytoplasmic	pink to orange	4,3-4	4,4	+++++,+++	+++++,+++
guaB	F	cytoplasmic	slow growth/other	1(SG)	5(SG)	-	-
hdfR	K	cytoplasmic	light red	3-4	4	+++	++
hfq	R	cytoplasmic	white to lt pink	2,1-2	5-6,5	+/-	-,-
hybD	C	cytoplasmic	light red	3	4	+	+
ihfA	L	cytoplasmic	white to lt pink	1	4	-	-
ihfB	L	cytoplasmic	white to lt pink	1	4	-	-
ilvG	R*	cytoplasmic	dark red	4	4	++	+++++
kdpD	T	Integral Membrane Protein	pink to orange	2	4-5	++	++
lipA	H	cytoplasmic	slow growth/other	1,2,2(VSG)	1,2,1(VSG)	+/-,++,+/-	NA,+/-,+/-
lipB	H	cytoplasmic	slow growth/other	2(SG),2(VSG)	1(SG),1(VSG)	ND,+/-	ND,+++
lon	O	cytoplasmic	growth/other	1(mucoid)	4	-	-
lpcA	M*	cytoplasmic	white to lt pink	2,1,2,2	5,5,ND,ND	++,++,+,+/-	-,-,-,-
lpd	C	cytoplasmic	slow growth/other	3, 1	2, 1	+/-, -	-, -
lpp	M	Outer membrane Lipoprotein	light red	4, 4	5, 4	++++, ++++	+, ++++
lptB	M*	cytoplasmic	light red	3,4,4	3,4,4	NA,++++,++	++++,++++,++
lsrF	G	cytoplasmic	dark red	3	4-5	++	++++

Table 2.5 (continued). Congo Red and Western blot data for Keio strains.

Name	Main COG Class	Echobase Echolocation	Original Color	Color @ 26C	Color @ 37C	Wester CsgG Level	Western CsgA Level
mdoC	M*	Integral Membrane Protein	red	6,6	4,4	+++++, +++++	++++,++++
mdoG	M*	periplasmic	light red	3	4	++	+++
mdoG	'	'	light red	1 mucoid	4	-	-
mdoH	M	Integral Membrane Protein	light red	3	4	++	+++
metC	E	cytoplasmic	pink to orange	3-4	4	++	+++
miaA	J	cytoplasmic	light red	4	4	++++	+++
mlrA	K	cytoplasmic	white to lt pink	1	3	-	-
mltA	M	Outer membrane Lipoprotein	red	4	4	+++	+++++
mtlR	K	cytoplasmic	light red	4	4	++++	++
nagA	G	cytoplasmic	pink to orange	3	4	++	++
nagC	G	cytoplasmic	pink				
nanE				3-4	ND	+++	++
nanK	G	cytoplasmic	light red	4,3-4	4,ND	++,+++	++++,++
narQ	T	Integral Membrane Protein	light red	3	4	++	+
ndk	F	cytoplasmic	dark red	5	4	++++	++
nhaA	P	Integral Membrane Protein	white to lt pink	1,1	5,5	-,-	-,-
nlpD	M	Outer membrane Lipoprotein	white to lt pink	1	4	+/-	-
nudC	F*	cytoplasmic	red	5-6	4	++++	+++++
nudH	R	cytoplasmic	pink to orange	3	3	+	+
nudL	R*	cytoplasmic	light red	3	4	++	++++
nuoA	C*	Integral Membrane Protein	dark red	6	3	+++++	++++
nuoB	C	cytoplasmic	red	6	3	+++++	++++
nuoC	C	cytoplasmic	dark red	6	3	++++	+++++
nuoE	C	cytoplasmic	red	6	3	+++++	++++
nuoF	C	cytoplasmic	red	5-6	2	++++	+++++

Table 2.5 (continued). Congo Red and Western blot data for Keio strains.

Name	Main COG Class	Echobase Echolocation	Original Color	Color @ 26C	Color @ 37C	Wester CsgG Level	Western CsgA Level
nuoG	C*	cytoplasmic	red	6	3	+++++	+++++
nuoJ	C	Integral Membrane Protein	dark red	6	3	++++	++++
nuoK	C	Integral Membrane Protein	dark red	6	3	++++	++++
nuoM	C	Integral Membrane Protein	red	6	3	+++++	++++
nuoN	C	Integral Membrane Protein	dark red	6	3	++++	++++
nusB	K	cytoplasmic	white to lt pink	2, 3	5-6, 6	+/-, -	-, -
ompC	M	Outer Membrane ß-barrel protein	light red	3,3	3,3	++++,++	++,+++
ompF	M	Outer Membrane ß-barrel protein	light red	4	3	++++	++++
ompR	T	cytoplasmic	white to lt pink	2	5	-	-
pal	M	Outer membrane Lipoprotein	white to lt pink	1	5	+	-
pcnB				6	ND	+++	+++++
pdxH	H	cytoplasmic	pink to orange	2	3	+	+
pepT	O*	cytoplasmic	white to lt pink	1	4	+/-	+/-
perR	K	cytoplasmic	dark red	5-6	4	+++++	++++
pgm	G	cytoplasmic	white to lt pink	1 mucoid	5	-	-
pgm	'	'	white to lt pink	2 mucoid	5	+/-	+/-
php	R	cytoplasmic	white to lt pink	1,2	4,4	-	-
poxA	J	cytoplasmic	pink to orange	3	4	+	++
priA	L	cytoplasmic	pink to orange	4-5	3	+++	+++
proW	E	Integral Membrane Protein	dark red	6,6	4,4	++++,+++++	++++,+++++
ptsH	G	cytoplasmic	light red	3,3-4,2-3	3,3,4	++,++,++	++,++,+++
ptsI	G	cytoplasmic	white to lt pink	2,3,2	2,3,1	+,+/-,+/-	+,+/-,+
purA	F	cytoplasmic	pink to orange	1(SG),3	4(SG),3	+/-,-	-,ND
purC	F	cytoplasmic	pink to orange	1(SG)	5	+/-	-
purD	F	cytoplasmic	white to lt pink	1(SG),2	5-6,6	-,-	-,ND

Table 2.5 (continued). Congo Red and Western blot data for Keio strains.

Name	Main COG Class	Echobase Echolocation	Original Color	Color @ 26C	Color @ 37C	Wester CsgG Level	Western CsgA Level
purE	F	cytoplasmic	white to lt pink	2(SG)	2(SG)	-	-
purF	F	cytoplasmic	white to lt pink	1,3	5(SG),3	-,-	-,ND
purH	F	cytoplasmic	pink to orange	1(SG)	5	+/-	-
purK	F	cytoplasmic	white to lt pink	2	4	+/-	-
purL	F	cytoplasmic	white to lt pink	1(SG),2	4(SG),1	-,-	-,ND
purM	F	cytoplasmic	white to lt pink	2,3	5-6,3	-.-	-,ND
purR	K	cytoplasmic	dark red	5, 6	5, 4	++++, ++++	+++++, ++++
putA	C	cytoplasmic	dark red	6,5	4,4	+++,+++++	+++++
puuR	K	cytoplasmic	light red	2	4	+	+
pyrB	F	cytoplasmic	white to lt pink	2	4	+/-	-
pyrC	F	cytoplasmic	light red	3(SG)	4(SG)	+++	++
pyrD	F	cytoplasmic	light red	4	2	+++	++
pyrE	F	cytoplasmic	white to lt pink	4-5(SG),5,2(SG)	3,2,3(SG)	+++,+/-,+	+,ND,+
pyrF	F	cytoplasmic	lt red	4-5(SG),5	3,2	+++	+,ND
qseC	T	Integral Membrane Protein	dark red	6	3-4	+++++	++
racC	C	cytoplasmic	pink to orange	3	4	-	++++
rbfA	J	cytoplasmic	pink to orange	2	4	+	+/-
rcsB	again		dark red				
rcsF			dark red	6		+++++	+++++
rfaC	M	cytoplasmic	white to lt pink	2,2,2,2	5,5,ND,ND	++,++,+,+	NA,-,-,-
rfaD	M	cytoplasmic	white to lt pink	2-3,2-3,2-3,2-3	4-5,4,ND,ND	++,++,++	-,+/-,+/-
rfaE	M	cytoplasmic	white to lt pink	2,2,2-3,2	5,5,ND,ND	++,+,+,+	NA,-,-,-
rfaF	M	cytoplasmic	white to lt pink	1,1,1-2,2	4-5,4,ND,ND	+/-,+/-,+/-,+/-	-,-,-,-
rfaG	M	cytoplasmic	white to lt pink	1,2,2	4-5,ND,ND	+/-,+/-,+/-	-.-,-
rfaH	K	cytoplasmic	light red	3,3,3-4	4,ND,ND	++,++,++	+.+,++

Table 2.5 (continued). Congo Red and Western blot data for Keio strains.

Name	Main COG Class	Echobase Echolocation	Original Color	Color @ 26C	Color @ 37C	Wester CsgG Level	Western CsgA Level
rfaP	M	cytoplasmic	pink to orange	4,3-4,3,3	4,4,ND,ND	+++,+++,+++,+++	+,++,+/-,++
rfe	M			6	ND	+++	+++++
rffA	M			3-4	ND	+++	+++
rffC	M	cytoplasmic	light red	3,3,4,3	4,4,4,ND	++,++++,+++,+++	+++++,++,+++,++
rffT	M			3	ND	+++	+++++
rimK	J*	cytoplasmic	dark red	5,4	4,4	++++,++	+++++,++
rimM	J*	cytoplasmic	white to lt pink	2	5	+/-	-
rnhA	L	cytoplasmic	pink to orange	4	4	++++	+++
rodZ	M*	Membrane anchored	white to lt pink	2-3,2, 4	4,4-5, 4	++,++	NA,+++
rplA	J	cytoplasmic	pink to orange	4	4	++	+++
rpoN	K	cytoplasmic	light red	3-4	3	++	++
rpoS	K	cytoplasmic	white to lt pink	1	3	-	-
rpoZ	K	cytoplasmic	pink to orange	3	4	++++	++++
rpsF	J	cytoplasmic	white to lt pink	2	4	-	-
rpsT	J	cytoplasmic	dark red	6	4	++++	+++++
rseA	T	Integral Membrane Protein	light red	3	6	+++++	+++++
rsgA	J*	cytoplasmic	pink to orange	3	4	++	+
rstA	K	cytoplasmic	light red	3	4	++	+
rstB							
rumA	J	cytoplasmic	dark red	5	4	+++++	+++
sdhA	C	cytoplasmic	pink to orange	3	3	+++	++
sdhB	C	cytoplasmic	white to lt pink	3	3	+	++
sdhC	C	Integral Membrane Protein	red/pink	6,5	4,3	++++,+++++	+++++,+++++
sdhD	C	Integral Membrane Protein	pink to orange	6,6	4,3	+++,++++	+++++,+++
sdiA	K	cytoplasmic	dark red	6	4	+++++	+++++

Table 2.5 (continued). Congo Red and Western blot data for Keio strains.

Name	Main COG Class	Echobase Echolocation	Original Color	Color @ 26C	Color @ 37C	Wester CsgG Level	Western CsgA Level
skp	M	periplasmic	dark red	5	4	++++	+++
slp	M*	Outer membrane Lipoprotein	light red	3-4	4	+++	++++
speB	E	cytoplasmic	light red	2-3	4	++	++
srlR	K*	cytoplasmic	white to lt pink	4	4	++	++
srmB	J	cytoplasmic	light red	3	4	++++	++++
sspA	O	cytoplasmic	pink to orange	4	5	++	++
sucA	C	cytoplasmic	light red	2(SG)	3(SG)	+/-	++
sucC	C	cytoplasmic	light red	3	3	+	+
sucD	C	cytoplasmic	dark red	4	1-2	++	+++
surA	O	periplasmic	light red	3-4	5	+++	++++
thyA	F	cytoplasmic	slow growth/other	1(VVSG)	1(VVSG)	slow growth	slow growth
tolA	N*	Membrane anchored	pink to orange	5,4	4,ND	+++++,+++++	++++,+++
tolB	N	periplasmic	white to lt pink	2,2	5,2	+,+/-	NA,-
tolC	N	Outer Membrane ß-barrel protein	red	3,3	4,4	++++,+++	++++,++
tolQ	N	Integral Membrane Protein	white to lt pink	2,2	5,5-6	+,+/-	-,-
tolR	N	Membrane anchored	white to lt pink	2-3	5	+	-
tomB	T*	cytoplasmic	light red	3	3	+	++
torI	K*	cytoplasmic*	pink to orange	3	4	++	+/-
tpiA	G	cytoplasmic	red	4	3	+++++	++++
treR	K	cytoplasmic	dark red	5,6	4,4	++++,+++++	++++,+++++
trpA	E	cytoplasmic	white to lt pink	2,1	3,3	+,+/-	+/-.+/-
trpB	E	cytoplasmic	white to lt pink	2	3	+	+/-
trpC	E	cytoplasmic	white to lt pink	2,3,2	3,3,4	NA,+/-,+/-	++++,-,-
trpD	E	cytoplasmic	white to lt pink	2,2	3,4	+,+	++,+/-
trpD	'	'	white to lt pink	3,2-3	4,3	+,+	++,+

Table 2.5 (continued). Congo Red and Western blot data for Keio strains.

Name	Main COG Class	Echobase Echolocation	Original Color	Color @ 26C	Color @ 37C	Wester CsgG Level	Western CsgA Level
trpE	E	cytoplasmic	white to lt pink	2,2	4	++,+/-	++,+/-
trpE	'	'	white to lt pink	3,3	4	+++,++	++++,++
ubiE	H	cytoplasmic	slow growth/other	1,2,1(SG)	2,1,2(SG)	-,+,-	-,-,-
ubiF	H	cytoplasmic	slow growth/other	2,2(SG)	2,2(SG)	+,+/-	+/-,-
ubiG	H	cytoplasmic	slow growth/other	1(SG),1(SG)	1(SG),1(SG)	ND,-	ND,-
ubiH	H	cytoplasmic	slow growth/other	1,1(SG)	1,1(SG)	-	-
uspE	T	cytoplasmic	dark red	3	4	+++	+
wzxE	M*	Integral Membrane Protein	pink to orange	3,3	4,ND	+++,++	++,++
xapR	K	cytoplasmic	light red	5, 6	4, 4	+++, ++	++++, ++++
ybaM	S*	cytoplasmic	light red	6	4	+++++	+++++
ybcI	R	Integral Membrane Protein	light red	2	3-4	++	++
ybeZ	T	cytoplasmic	dark red	6	4	+++++	+++++
ybgC	R	cytoplasmic	light red	3	4-5	+++	+++
ycfM	R	Outer membrane Lipoprotein	pink to orange	4	4	++++	+++++
ycgV	M	cytoplasmic	light red	3	4	++++	++++
yciC	S*	Integral Membrane Protein	light red	3, 4	4, 4	++++, +++	++, ++
yciM	G	periplasmic	light red	2,3,2	5,4,ND	+, ++,++	++++, ++,+/-
ydaF	S*	cytoplasmic	light red	2-3,2-3	4,4	++,++	NA,++++
ydaM	T	cytoplasmic	white to lt pink	1	4	+	-
ydaT	S*	cytoplasmic	pink to orange	3,2-3	4-5,4-5	++++, ++	+++,+++
ydcI	K	cytoplasmic	white to lt pink	1	4	+	-
ydeJ	R	cytoplasmic	dark red	6	4	+++++	+++++
ydiT	C	cytoplasmic	light red	4	3	+++	++
yecH	S*	cytoplasmic	pink to orange	3	4	++++	++++
yedQ	T	Integral Membrane Protein	red	6	4	+++++	++++

Table 2.5 (continued). Congo Red and Western blot data for Keio strains.

Name	Main COG Class	Echobase Echolocation	Original Color	Color @ 26C	Color @ 37C	Wester CsgG Level	Western CsgA Level
yedV	T	Membrane anchored	light red	3	4	++	+
yeeN	S	cytoplasmic	dark red	6	4-5	+++++	++
yeiL	T*	cytoplasmic	pink to orange	3	4	+	+/-
yeiR				6	ND	+++++	++++
yfeW	M	periplasmic	dark red	6,5-6	3,4	++++,++++	+++++,+++++
yfgC	R	periplasmic	dark red	5-6	4	+++++	+++++
ygfZ	R	cytoplasmic	slow growth/other	3	3	+++	+++
yhbJ	R	cytoplasmic	dark red	6	4	+++	+++++
yheV	S*	cytoplasmic	dark red	4	4	+++++	+++
yiaK	C	cytoplasmic	light red	2,2	4,4	+/-	-
yiaK	'	'	light red	3-4,3-4	4,4	++++	+++
yicC	S	cytoplasmic	light red	2	4	++	++++
yieP	K*	cytoplasmic	white to lt pink	1	4	+/-	-
yiiS	S	cytoplasmic	red	4	3	++++	+++
yjjW	O	cytoplasmic	light red	4	4	+++	+++
ykgH	S	Integral Membrane Protein	light red	4	4	++++	++
ymgE	S	Integral Membrane Protein	pink to orange	3	4	++	++
ymgE	'	'	pink to orange	3-4	4	+	+
ynaK	K	cytoplasmic	light red	3	4	++++	++++
yncG	O	cytoplasmic	pink to orange	3	4	+++	++
ynjD	R	cytoplasmic	light red	4	4	+++	+++++
yoeE	P	cytoplasmic	dark red	5-6,5	4,3	++++,+++++	+++++,+++++
yqaB	R	cytoplasmic	light red	4-5	3	++	+++
Notes	* Best pick or changed	* Changed from Echobase or best idea					

Table 2.5 (continued). Congo Red and Western blot data for Keio strains.

Keio strains that were tested by Western blot or retested for CR and shown to be WT

Name	Original Color	Color @ 26C	Color @ 37C	Wester CsgG Level	Western CsgA Level
alaS	pink to orange	5	4	++++	++++
appB	red	5-6,5	3-4,4	++++,++++	++++,++++
arcA		5	ND	++++	++++
aroF		5	4	++++	++++
aroG		5-6	4	++++	+++++
aroL		5	4	+++++	+++++
argI		5	ND	++++	++++
argF	pink	6,4-5(dry),5	4,4,ND,ND	++++,+++++,+++++	++++,+++++,++++
arnT		5	ND	++++	++++
atpD	pink	4-5	5	++++	++++
bioH	red	4-5	4	++++	++++
btuD	dark red	5,5-6	4,4	++++,++++	++++,++++
cbpA	red	5,4,4	4,4,4	++++,++++,++++	++++,++++,++++
crl	red	4,5	4,4	++++	++++
csgC	red	5, 5	4, 3	++++	++++
cspC	pink to orange	5	4	++++	++++
cusR	light red	5	4	++++	++++
cusS	red	4, 5	4, 4	++++	++++
emtA	red	4-5	4	++++	++++
fadH	light red	5,4	4,4	++++	++++
fecC	dark red	4	4	++++	++++
fepC	dark red	5	2	++++	++++
fes	light red	5	2	++++	++++
fre	red	5	2	++++	++++
fucO	light red	5	4	++++	++++
gltA	dark red	5	4	++++	++++
guaA	red	4,4,5-6	4,4,4	++++,++++,++++	NA,++++,++++
hisG	light red	4	4	++++	++++
lpxL		5	ND	++++	bad
lpxM		5	ND	++++	++++
lpxP		5	ND	bad	+++
ilvB	red	5	4	++++	++++
mltB	red	4-5	4	++++	++++
nanA		5	ND	++++	++++
paaF	dark red	5	5	++++	++++
paaI	pink to orange	4-5	4	++++	++++
pagP		5	ND	bad	
pdxJ	pink	5	3-4	++++	++++
phoP	red	5	4	++++	++++
phoQ	red	5	4	++++	++++
purT	dark red	5,5	4,4	++++,++++	++++,++++
rfaB	red	5,5,5	4,ND,ND	++++,++++,++++	++++,++++,++++

Table 2.5 (continued). Congo Red and Western blot data for Keio strains.

Keio strains that were tested by Western blot or retested for CR and shown to be WT

Name	Original Color	Color @ 26C	Color @ 37C	Wester CsgG Level	Western CsgA Level
rfaI	red	5,4-5,4-5,4-5	4,4,ND,ND	++++,++++,++++,++++	+++,++++,+++++,++++
rfaJ	red	4-5,5,5,5	4,4,ND,ND	++++,++++,++++,++++	++++,++++,+++++,+++++
rfaL	red	4-5,4-5,5,5	4,4,ND,ND	++++,++++,++++,++++	++++,++++,NA,++++
rfaQ	red	4-5,5,5	4,ND,ND	++++,++++,+++++	++++,++++,+++++
rfaS	red	5,5,5,5	4,4,ND,ND	++++,++++,++++,++++	++++,+++++,+++++,++++
rfaY	red	5,5,5,5	4,4,ND,ND	++++,++++,++++,++++	++++,+++++,++++,++++
rfaZ	red	5,5,5	4,ND,ND	++++,++++,++++	++++,++++,++++
rph	red	4	4	++++	++++
rpmE	red	4	4	++++	++++
rstB	dark red	5-6	4	++++	++++
sbmC	dark red	5	4	++++	++++
tktA	red	4	4	++++	++++
tktB	red	5	4	++++	++++
tpr	dark red	4	4	++++	++++
treC	red	4	4	++++	++++
truB	red	4-5	4	+++++	++++
ubiC		4	4	++++	+++++
uidC	dark red	5	4	++++	++++
ybaP	dark red	6	4	++++	++++
ycbR	red	5-6	4	++++	++++
yceP	light red	4	4	++++	++++
ycjU	dark red	5,5,5	4,4,4	++++,?,++++	++,?,++++
ydbH	light red	4	4	++++	++++
ydiZ	dark red	4	4	++++	++++
yfcL	dark red	5-6	4	++++	++++
yhaK	pink	5	4	++++	++++
yjbB	light red	5	4	++++	++++
yjiK	red	4	4	++++	++++
ylcG	red	4-5	4	++++	++++
ymjB	dark red	4	4	++++	++++
yniD	light red	4-5	4	++++	++++
yoaE	light red	4-5	4	++++	++++
ypfG	light red	4-5,4-5	4,4	++++	++++
yqeC	dark red	5-6	4	++++	++++
yqiK	dark red	5	4	++++	++++
zraR	dark red	5	4	++++	++++
ygiF	dark red	5,5	4,4	+++++,++++	+++++,+++
hns	red	5, 5	4, 3	+++++	++
galM	pink to orange	5,5	4,4	++++,++++	++++,+++++

Notes: Commas separate individual experiments.
SG – slow growth, NA – not applicable, ND,? – not done

Table 2.6. Results from PCR for strains with multiple CR phenotypes.

Gene	Color	PCR Results	Upstream Primer	COG	Notes
btuC	3,5	both correct	CGC TTA CAA AAG CTG AAA TGT C	H	
cysH	3,5	both correct	CGA AAC ATG GTG TCA GCG	H	
dcuC	3,5	both correct	CGA GAT ACA ACA ATC ATC TTA ACG	C	
dnaQ	1,2,3,4	all correct	GCA TCA AGA CGT TGC CAG	L	
fepE	2,6	both correct	CTG TTT GTT GCT GAA CGC	P,M	entertobactin and O antigen length
flhC	3,5	both correct	GTG AAA CCG CAT AAA AAT AAA GTT GG	K*	
fliG	2,5	both correct	GCT CAA TGT CGT TAA CTC G	N	
glcD	3,4	both correct	GCG TAC TGA ACA GAT GGA TC	C	
gshB	4,5	both correct	GCG TTA AAC TGG ATA GTG AAC G	H	
hepA	3,4	both correct	CGA TGA AGA AAA CCA AAA GCG	K	
hycI	3,5	both correct	GAA AAG GTG GTG TTC AGT CAA C	C	hydrogenase maturation
hypE	3,5	both correct	CGA TTG CCG ATG TGT TCT G	O	hydrogenase maturation
kbl	3,5	both correct	CTA TCT GGG GAG AGG AAA ATG G	H	
manY	3,5	both correct	CAC ACG TAG TTG ATG TTG C	G	
mfd	3,5	both correct	GCT GAA TGT GAT TGG CTC C	L*	
mppA	2,5	both correct	CGG CAT TCT CAT TAT TTA TAG ATA GG	E	
mutL	2,3,5	all correct	GTC TGA CCC CTA TTT AAG CC	L	
nuoH	5,6,1	all correct	CAA ATT CCA GGA CGA AGT GG	C	
nuoI	2,5,6	all correct	CCG CAA TTC TTT GGT TTT ATT ACC	C	
nuoL	2,5,6	all correct	GCC AAT CAG TGC TAA AGC	C	
pdxB	*1 mucoid,5*	both correct	*CGC CAG ATT ATG CTC AGC*	H*	*Mucoid white -> 5. Can't isolate only mucoid*
purN	2,5	both correct	GGT AAG CCA TTA GCC GAT C	F	
qseB	2,6	both correct	GCA TCT TTG AAC ACG TAG AGA TC	K	quorum sensing regulates *flhDC*
rbsK	2,4	both correct	GCG CAG AAT GAT GAA ATG G	G	
rbsR	2,5	both correct	CCA TCA AAA TAA GAC TAT CGT TGC	K	regulates *rbsDABCK*
rdgC	3,5	both correct	GTC TTT ATC GAA TGG CTG ACC	L	pillin variation in *Neisseria gonorrhoeae*
rnk	2,5	both correct	GCA GTT AAG CTA ACG TGC	K	
rssB	*3,5*	*NONE*	*GAA TGG TGA TTC TCT GCC G*	T	*No correct inserts - Wrong primer/Strain?*
rsxC	2,5	both correct	GAA GCC ATC AGC TGT AAC G	C	
sbcD	3,5	both correct	GGT GTT TGA TGA ACT GGA TAC C	L	
sfcA	2,5	both correct	CGG CAA CCT AAT TTA GGG G	C	
ssnA	*2,5*	both correct	*GTA ACT ACG TTT GCA GCA AG*	R*	
sucB	2,6	both correct	GTA GTG ATG TGT TCT GGT AAG G	C	
tonB	2,5	both correct	CCC TGG ATC GTT ACT GTC	M	
uup	3,4-5	both correct	GCA ACC TGC GTC TGA ATG	R	
wcaD	3,5	both correct	GGC GAC AAA ATT GAA CTG C	M*	
ybaB	2,5	both correct	GCT ACC AAA ACT GGT CGA AC	S	
rlmF	3,6	both correct	CAG GTT ACT CAT CGG TTG C	J*	*was ybiN*
ybjN	3,5	both correct	GAT TTT CGT TCC AAT TTG CAT CG	R*	*3' to rimK*
ycbL	3,5	both correct	GGC TAT ATT CAG GAA GAA TTG GC	R	
ycbQ	3,5	both correct	GAG AAG GAA AGC GAG GAC	U	pillin ATPase
bhsA	1,5	both correct	GAG AAG TCG CTT CAT AAC CG	M*	*was ycfR*
yciU	2,5	both correct	CGA TGA TTT ACT TCA TGC GAT TTG	S*	
ydeP	3,5	both correct	CAG GAA ACA AGG TTT CAG C	C	
ydfD	3,5	both correct	CGG TAT CAG TTT TAC TCC GTG	S*	
ydiY	3,5	both correct	GGA TAA ATT TGC GGG GTA ATT G	M	
ydjI	1 mucoid,5	both correct	CTA TTG CAC GCT GTC AGC	G	

Table 2.6 (continued).

Gene	Color	PCR Results	Upstream Primer	COG	Notes
yebY	3,5	both correct	CGA TTC ACC TGA TTT GTG C	S*	
yehE	3,6	both correct	GGT GGT GCA GAG AAA CTG	S*	
yfjW	3,5	both correct	CAT CCA CGG AGA TCA TAA CG	R*	
yjbI	3,5	both correct	CTA AGA ATA TCC ATT ATC TCA ATG CC	S	
gsiA	2,5	both correct	GCG GAA TGA CCA ATA AAT TAC C	H*	*was yliA,was E**
ynfH	5,6	both correct	GCA ATG CAT AAG CGT GAA G	C*	
Still used					
cmk	1,2	both correct	GCG CTA TCA ATG CTA AAT ACT CC	F	
fabF	2,3	both correct	CGA AAA CCA TCG CGA AAG C	Q	
mdoG	3, 1 mucoid	both correct	GGT TCA TAT ATG GTT AAC TAA TCT CGG	M*	
pgm	1 mucoid,2 mucoid	both correct	GCC GGT CAA AAC GAT TAA AGA C	G	
trpD	2,3	both correct	CCA AAG TTG ACC GTT ATT CC	E	
trpE	2,3	both correct	CGT GAA ATT TCC TCT CTT GC	E	
yiaK	2,3	both correct	GCA ATA AGC GAT GGA CGG	C	
ymgE	3,3-4	both correct	CGA TGT TGT CTT CGG CTT G	S	
Only one correct					
fhlA	3,5	only 3's	GGT TGT GGT ATT GAA CTT TCA G	K	
sdhC	2,6	only 6	GCA TTA TAT GCT TTT CCT GGT AAT G	C	
sdhD	3,6(SG)	only 6(SG)	CAT GTG GGC GTT ATT CAT GAT AAG	C	

Distribution of function of strains with multiple CR phenotypes

COG	Number	Total # COG Category	% of Suppressors	% COG Category
C	12	281	19.67	4.27
E	3	390	4.92	0.77
F	2	83		
G	4	355	6.56	1.13
H	6	144	9.84	4.17
J	1	170		
K	5	288	8.2	1.74
L	5	237	8.2	2.11
M	5	219	8.2	2.28
N	1	151		
O	1	133		
P	1	253		
Q	1	95		
R	5	448	8.2	1.12
S	7	447	11.48	1.57
T	1	187		
U	1	447		

Table 2.7. Congo Red phenotypes of Keio strains on YESCA and CFA plates or CFA plates without Coommassie Brilliant Blue.

Set A	Niba Results	24H			48H			72H			Different
Name		YESCA CR	CFA CR	CFA 2XCR	YESCA CR	CFA CR	CFA 2XCR	YESCA CR	CFA CR	CFA 2XCR	
flhC lt	+	3	3	3	3	3	3	3	3	3	Yes
flhC WT	+	4-5	5	5	4	5	5	4-5	4-5	5	no
fliG lt	+	2	2	2	3	2	2	3	3	2-3	Yes
fliG WT	+	4-5	4-5	4-5	4-5	5	5	5	5	5	no
fliI	+	6	6	6	6	6	6	6	6	6	Yes***
fliT	+	3	3-4	3	3	3-4	3	2-3	3	3	Yes
btuC lt	Not Tested	2	2	2	2	2	2	2	2	2	
btuC WT	Not Tested	4-5	4-5	4-5	4	4-5	5	4-5	4	4	
btuB	variable	4-5	4-5	4-5	6	5	6	4-5	5	5	unclear
csgA	-	1	1	1	1	1	1	1	1	1	no
csgB	-	1	1	1	1	1	1	1	1	1	no
csgC	Not Tested	4-5	4-5	4-5	5	5	5	6	6	6	
csgD	-	1	1	1	1	1	1	1	1	1	no
csgE	-	1	1	1	1	1	1	1	1	1	no
csgF *	+	1	1	1	1	1	1	1	1	1	Yes*
csgG	-	1	1	1	1	1	1	1	1	1	no
surA	+	2-3	3	3	3-4	3-4	4-5	4	4	4	Yes****
dnaK	+	1	1-2	2	2	2	2	2-3	2	3	Yes
fruR	+	2	1-2	2	2	2	2	3	2	2	Yes
gcvA	+/-	2	2	2	2-3	2-3	2	3	3	2	no
greA	+	2-3	2-3	2-3	3	3	3	4	4	3	Yes

Notes

* *csgF* has red fibers underneath the colonies in the agar and can appear pinkish

All strains in Table S1 were tested for this phenotype and only *csgF* displayed it.

**Somewhat mucoid

*** Niba et al did not test for higher curli expression

**** May be due to plate differences(Yesca vs CFA) or dye differences (CR w CBB vs 2XCR) or time (48 vs 72 hours)

Table 2.7 (continued). Congo Red phenotypes of Keio strains on YESCA and CFA plates or CFA plates without CBB.

Set B	Niba Results		24H			48H			72H		Different
Name		YESCA CR	CFA CR	CFA 2XCR	YESCA CR	CFA CR	CFA 2XCR	YESCA CR	CFA CR	CFA 2XCR	
lpcA	+	1-2	1	1	1-2	1	1-2	2	1	1-2	Yes
gmhB	+	2	2	1	2	1	1	2	1	1	Yes
rfaC	Not tested	2	2	1	2	1	2	2	1	2	
rfaD	+	2	2	2	2-3	2	3	2-3	1-2	2	Yes
rfaE	+	1-2	1	1	2	1	1-2	2	1	1-2	Yes
rfaF	+	1	1	1	1	1	1	1	1	1	Yes
rfaG	+/-	1	1	1	1	1	1	1	1	1	no
rfaH	+	2-3	3	3	3	3	3	3	3-4	4-5	Yes****
rfaP	+	2-3	3	3	3	4	3	3	4	3	Yes
galU	+	1	1	1	1	1	1	1	1	1	Yes
hfq	+/-	1	1	1	1-2	2	2	2-3	3	3	no
hsrA/yieO	-	3-4	4-5	4	5	5	3-4	4	4	4	Yes****
yieP	Not tested	1	1	1	1**	1**	1**	1**	1**	1**	
lon	-	1 mucoid	1 mucoid	1 mucoid	1 mucoid	1 mucoid	1 mucoid	1 mucoid	1 mucoid	1 mucoid	no
mdoH pink	+	3	4-5	4	3-4	4	3-4	3	5	4	Yes****
mlrA	-	1	1	1	1	1	1	1	1	1	no
nagA	+	2	3	3	2-3	2-3	2	2	3	3	Yes
nifU/iscU	+/-	3	3	3	3-4	4	3	3-4	4	4	no
ompR	-	1	1	1	2	2	2	2	2	2	no
rpoS	-	1	1	1	1	1	1	1-2	1-2	1-2	no
nlpD	-	1	1	1	1	2	1-2	1	2	2	no

Notes

* *csgF* has red fibers underneath the colonies in the agar and can appear pinkish
All strains in Table S1 were tested for this phenotype and only *csgF* displayed it.
**Somewhat mucoid
*** Niba et al did not test for higher curli expression
**** May be due to plate differences(Yesca vs CFA) or dye differences (CR w CBB vs 2XCR) or time (48 vs 72 hours)

Table 2.7 (continued). Congo Red phenotypes of Keio strains on YESCA and CFA plates or CFA plates without CBB.

Set C	Niba Results	24H			48H			72H			
Name		YESCA CR	CFA CR	CFA 2XCR	YESCA CR	CFA CR	CFA 2XCR	YESCA CR	CFA CR	CFA 2XCR	
crp	+/-	6	6	6	6	5-6	5	4-5	4-5	4-5	Yes*** ****
cyaA	-	1-2 slow	1 slow	1 slow	1 slow	1 slow	1 slow	1 slow	1 slow	1 slow	no
sdhC	+	6	5	6	5-6	5	5	5-6	5	5	Yes***
tolA	+	3-4	4	4	4	4	4-5	4	5	5	Yes****
tolB	+/-	2	2	2	2	2	2	2	2	3	no
tolR	+/-	2	2	2	2	2	2	2	2	3	no
ycfM	+	2-3	3	3	3	3	3	3	3	3	Yes
yciB	+	3	4	4	5	5	5	4	4	4-5	no
yciC	Not tested	3	3-4	3-4	4	5	4	4	4	4	
yciM	+/-	2	2	2	2-3	2	2-3	2-3	2-3	3	no
ydaM	+/-	1	1	1	1	1	1	1	1	1	no
ynjC	+	3	3-4	3	3	4	3-4	3	4	4-5	Yes****
ynjD	Not tested	6	6	6	6	6	6	6	6	6	
ptsI	+	1-2	2	2	1	1-2	1-2	1	1	1	Yes
ihfB	-	1	1	1	1-2	1-2	2	2	2	2	no
mdoH mucoid	+	2**	2-3**	2-3**	2**	2-3**	2-3**	2-3	3	3	Yes
lpxL	Not tested	3-4	3-4	4	5	5	5	4-5	5	5	
lpxM	Not tested	3-4	3-4	4	5	5	5	5	5	5	
lpxP	Not tested	4-5	3-4	5	5	5	5	5	5	5	
crcA/pagP	Not tested	4-5	3-4	5	5	5-6	5	5	5	5	
arnT	Not tested	3-4	3-4	4	4-5	4-5	5	5	5	4-5	

Notes

* *csgF* has red fibers underneath the colonies in the agar and can appear pinkish

All strains in Table S1 were tested for this phenotype and only *csgF* displayed it.

**Somewhat mucoid

*** Niba et al did not test for higher curli expression

**** May be due to plate differences(Yesca vs CFA) or dye differences (CR w CBB vs 2XCR) or time (48 vs 72 hours)

Table 2.8. Two hundred longest intergenic regions in *E. coli* K-12.

Rank	Name	Length	Orientation	L_END	R_END	Cs
1	nanC_fimB	1455	**Divergent**	4537525	4538979	97.8
2	ygcE_ygcF	1372	Convergent	2901397	2902768	62.53
3	iap_ygbF	950	Convergent	2875641	2876590	61.98
4	lrhA_yfbQ	919	**Divergent**	2404664	2405582	51.83
5	dppA_proK	910	Codirectional-	3705729	3706638	79.87
6	betT_yahA	874	Codirectional+	330721	331594	7.13
7	yiiD_yiiE	858	Codirectional+	4076462	4077319	87.86
8	mngB_cydA	846	Codirectional+	769835	770680	16.59
9	hdeD_gadE	798	Codirectional+	3655591	3656388	78.79
10	araF_ftnB	796	**Divergent**	1984153	1984948	42.76
11	matA_ykgL	775	**Divergent**	310561	311335	6.69
12	ycgV_ychF	768	Codirectional-	1255176	1255943	27.05
13	**csgD_csgB**	**754**	**Divergent**	**1102420**	**1103173**	**23.76**
14	essQ_cspB	753	Codirectional-	1638610	1639362	35.32
15	ydfJ'_ydfK	753	Divergent	1630310	1631062	35.14
16	ypjC'_ileY	750	Codirectional-	2783034	2783783	59.98
17	yncH_rhsE'	749	Codirectional+	1525177	1525925	32.87
18	ychE_oppA	737	Codirectional+	1298469	1299205	27.99
19	ycdU_serX	735	Convergent	1096053	1096787	23.62
20	eco_mqo	714	Convergent	2302416	2303129	49.62
21	gltA_sdhC	708	Divergent	753692	754399	16.24
22	ynaE_ttcC'	700	Codirectional-	1432282	1432981	30.87
23	iraM_ycgX	699	Codirectional-	1211227	1211925	26.11
24	dinQ_arsR	694	Divergent	3645857	3646550	78.58
25	yfcV_sixA	680	Codirectional-	2453669	2454348	52.88
26	yjiC_iraD	672	Divergent	4554344	4555015	98.16
27	yobF_yebO	669	Codirectional-	1905616	1906284	41.07
28	stpA_ygaW	668	Divergent	2796518	2797185	60.27
29	leuL_leuO	659	Divergent	83709	84367	1.8
30	wza_yegH	658	Divergent	2135268	2135925	46.02
31	yjdN_yjdM	657	Codirectional-	4323765	4324421	93.19
32	gltP_yjcO	641	Convergent	4293818	4294458	92.55
33	nuoA_lrhA	630	Codirectional-	2403095	2403724	51.79
34	arsC_yhiS'	628	Codirectional+	3648686	3649313	78.64
35	yjjP_yjjQ	618	Divergent	4600882	4601499	99.16
36	atpI_rsmG	616	Codirectional-	3920464	3921079	84.5
37	yaiS_tauA	615	Divergent	383841	384455	8.27
38	purH_rrsE	614	Divergent	4205556	4206169	90.64
39	yhaC_rnpB	613	Convergent	3267625	3268237	70.43
40	malP_malT	611	Divergent	3550496	3551106	76.52
41	nth_tppB	610	Codirectional+	1710183	1710792	36.86
42	dnaA_rpmH	606	Divergent	3881753	3882358	83.66
43	hns_tdk	604	Divergent	1292146	1292749	27.85
44	ompF_asnS	602	Codirectional-	986206	986807	21.26
45	fruR_mraZ	601	Codirectional+	89033	89633	1.92

Table 2.8 (continued). Two hundred longest intergenic regions in *E. coli* K-12.

Rank	Name	Length	Orientation	L_END	R_END	Cs
46	ppc_argE	597	Codirectional-	4151122	4151718	89.47
47	yedS'_hchA	591	Codirectional+	2033268	2033858	43.82
48	dcuC_pagP	588	Divergent	655192	655779	14.12
49	pgaA_ycdT	586	Divergent	1091513	1092098	23.53
50	yjhR'_yjhS	582	Convergent	4534055	4534636	97.72
51	yhcC_gltB	581	Divergent	3352073	3352653	72.25
52	mhpE_mhpT	577	Codirectional+	374106	374682	8.06
53	glnS_ybfM	576	Codirectional+	706981	707556	15.24
54	nmpC'_essD	572	Divergent	576049	576620	12.42
55	dcuB_dcuR	570	Codirectional-	4346768	4347337	93.69
56	yodB_serU	569	Convergent	2040923	2041491	43.99
57	gltD_gltF	559	Codirectional+	3358639	3359197	72.39
58	ileY_ygaQ'	559	Divergent	2783860	2784418	60
59	ygdH_sdaC	556	Codirectional+	2925695	2926250	63.06
60	yjhV'_fecE	556	Convergent	4508157	4508712	97.17
61	ydhZ_pykF	556	Divergent	1753166	1753721	37.79
62	ykgP'_eaeH'	551	Divergent	313030	313580	6.75
63	fucA_fucP	546	Divergent	2931711	2932256	63.19
64	tomB_acrB	545	Codirectional-	479933	480477	10.34
65	tfaX'_appY	545	Codirectional+	582359	582903	12.55
66	yiaW_aldB	544	Codirectional-	3752452	3752995	80.88
67	trxB_lrp	544	Divergent	931274	931817	20.07
68	putP_efeU'	542	Codirectional+	1080037	1080578	23.28
69	IS4_yjhB	542	Divergent	4501539	4502080	97.02
70	aroP_pdhR	540	Divergent	121552	122091	2.62
71	narI_rttR	539	Convergent	1285750	1286288	27.71
72	yibI_mtlA	536	Divergent	3769768	3770303	81.25
73	yneO'_lsrK	530	Codirectional-	1596111	1596640	34.4
74	ykgD_ykgE	526	Codirectional+	320306	320831	6.9
75	lysC_pgi	524	Divergent	4231257	4231780	91.2
76	yfjW_ypjI'	523	Codirectional+	2773044	2773566	59.77
77	hemB_yaiT'	523	Divergent	388952	389474	8.38
78	pntA_ydgH	523	Divergent	1675928	1676450	36.12
79	thrS_arpB'	523	Divergent	1800595	1801117	38.81
80	minC_ycgJ	519	Divergent	1225304	1225822	26.41
81	yhjB_yhjC	519	Divergent	3669918	3670436	79.1
82	narK_narG	515	Codirectional+	1278572	1279086	27.56
83	sseA_ryfA	515	Codirectional+	2651362	2651876	57.15
84	frc_yfdX	512	Codirectional-	2491277	2491788	53.7
85	tyrB_yjbS	501	Convergent	4266331	4266831	91.95
86	yjgZ_IS4	500	Convergent	4499613	4500112	96.98
87	pgi_yjbE	498	Codirectional+	4233431	4233928	91.24
88	yjiS_yjiT'	498	Codirectional+	4569939	4570436	98.5
89	ybjE_aqpZ	494	Codirectional-	914081	914574	19.7
90	icd_ymfD	493	Convergent	1195597	1196089	25.77

Table 2.8 (continued). Two hundred longest intergenic regions in *E. coli* K-12.

Rank	Name	Length	Orientation	L_END	R_END	Cs
91	zupT_ribB	489	Convergent	3181346	3181834	68.57
92	yghJ_glcA	484	Codirectional-	3117135	3117618	67.18
93	sokC_nhaA	482	Codirectional+	17007	17488	0.37
94	IS1H_uspC	482	Divergent	1977295	1977776	42.62
95	fimE_fimA	481	Codirectional+	4540657	4541137	97.87
96	yieP_rrsC	480	Divergent	3939351	3939830	84.91
97	mdtM_yjiP'	478	Divergent	4566543	4567020	98.42
98	ydaN_dbpA	477	Codirectional+	1407058	1407534	30.33
99	fimB_fimE	477	Codirectional+	4539583	4540059	97.84
100	adhE_ychE	476	Divergent	1297345	1297820	27.96
101	ygeV_ygeW	475	Divergent	3003809	3004283	64.74
102	yffP_yffQ	474	Codirectional+	2561140	2561613	55.2
103	ykgR_ykgP'	473	Codirectional-	312467	312939	6.73
104	rrsD_yrdA	473	Divergent	3426785	3427257	73.86
105	caiT_fixA	471	Divergent	41932	42402	0.9
106	folD_sfmA	470	Divergent	556965	557434	12
107	yffL_yffM	469	Codirectional+	2558921	2559389	55.15
108	yjgX'_yjgZ	468	Divergent	4498815	4499282	96.96
109	ybcK_ybcL	464	Codirectional+	569652	570115	12.28
110	xisE_ymfI	464	Divergent	1200256	1200719	25.87
111	ddlA_iraP	462	Divergent	400148	400609	8.62
112	yoaE_manX	462	Divergent	1899610	1900071	40.94
113	cyaY_yifL	462	Divergent	3992083	3992544	86.04
114	cyoA_ampG	459	Codirectional-	450835	451293	9.72
115	yqeG_yqeH	459	Codirectional+	2985099	2985557	64.34
116	yhbX_leuU	458	Codirectional-	3319636	3320093	71.55
117	uvrY_yecF	458	Divergent	1993384	1993841	42.96
118	ymfE_lit	457	Divergent	1197461	1197917	25.81
119	yhgF_feoA	456	Codirectional+	3537729	3538184	76.25
120	dapB_carA	455	Codirectional+	29196	29650	0.63
121	ydhY_ydhZ	454	Codirectional-	1752502	1752955	37.77
122	ygeR_xdhA	453	Divergent	2997914	2998366	64.61
123	yeeN_asnW	452	Convergent	2055599	2056050	44.3
124	ybdK_hokE	452	Divergent	606607	607058	13.07
125	iscR_trmJ	451	Codirectional-	2660154	2660604	57.33
126	rhaD_rhaA	450	Codirectional-	4092296	4092745	88.2
127	gcvT_visC	447	Codirectional-	3048690	3049136	65.71
128	astC_xthA	445	Divergent	1830007	1830451	39.44
129	rrsG_clpB	442	Codirectional-	2729180	2729621	58.82
130	yeeW'_yoeD'	441	Convergent	2076132	2076572	44.75
131	kilR_sieB	441	Divergent	1416254	1416694	30.52
132	rfaQ_waaA	441	Divergent	3806122	3806562	82.03
133	prpB_prpC	439	Codirectional+	348797	349235	7.52
134	flxA_intK'	436	Convergent	1644762	1645197	35.45
135	mipA_yeaG	435	Divergent	1864497	1864931	40.19

Table 2.8 (continued). Two hundred longest intergenic regions in *E. coli* K-12.

Rank	Name	Length	Orientation	L_END	R_END	Cs
136	*ygfT_ygfU*	435	Divergent	3028954	3029388	65.28
137	*gudP_yqcA*	434	Codirectional-	2920123	2920556	62.94
138	*mdh_argR*	434	Divergent	3382291	3382724	72.9
139	*yiaF_yiaG*	433	Divergent	3717068	3717500	80.11
140	*lysQ_nadA*	432	Codirectional+	780876	781307	16.83
141	*ybbP_rhsD*	430	Codirectional+	522055	522484	11.25
142	*yhdV_yhdW'*	430	Codirectional+	3416634	3417063	73.64
143	*tldD_yhdP*	429	Codirectional-	3390051	3390479	73.07
144	*yjcB_yjcC*	429	Divergent	4273065	4273493	92.1
145	*yobD_yebN*	428	Codirectional+	1903284	1903711	41.02
146	*serT_hyaA*	426	Divergent	1030936	1031361	22.22
147	*ygjK_fadH*	425	Codirectional+	3229262	3229686	69.6
148	*aqpZ_ybjD*	425	Divergent	915271	915695	19.73
149	*yfaL_ypaB*	424	Codirectional-	2342192	2342615	50.48
150	*aslA_glmZ*	424	Divergent	3984031	3984454	85.87
151	*glpF_zapB*	424	Divergent	4116114	4116537	88.72
152	*ynfM_asr*	423	Codirectional+	1668977	1669399	35.97
153	*metK_galP*	423	Codirectional+	3085883	3086305	66.51
154	*ompT_pauD'*	423	Divergent	584857	585279	12.61
155	*livK_yhhK*	423	Divergent	3595584	3596006	77.5
156	*psuK_fruA*	422	Codirectional-	2257319	2257740	48.65
157	*putA_putP*	422	Divergent	1078106	1078527	23.24
158	*yahG_yahI*	421	Codirectional+	338968	339388	7.31
159	*yoaB_yoaC*	421	Codirectional+	1891736	1892156	40.77
160	*cspF_quuQ*	421	Convergent	1640092	1640512	35.35
161	*yjjV_yjjW*	419	Convergent	4612284	4612702	99.41
162	*yabP'_rluA*	417	Convergent	59270	59686	1.28
163	*IS1D_cspH*	417	Convergent	1049769	1050185	22.63
164	*yagT_yagU*	417	Divergent	301798	302214	6.5
165	*aer_patA*	417	Divergent	3217099	3217515	69.34
166	*emrK_evgA*	415	Divergent	2481362	2481776	53.48
167	*casA_ygcB*	414	Codirectional-	2882161	2882574	62.12
168	*yahL_yahM*	412	Codirectional+	344216	344627	7.42
169	*rplA_rplJ*	412	Codirectional+	4177607	4178018	90.04
170	*ryjA_yjcD*	412	Divergent	4276090	4276501	92.16
171	*lpxT_spr*	411	Codirectional+	2267590	2268000	48.87
172	*mdtJ_tqsA*	411	Divergent	1671526	1671936	36.03
173	*yhjD_yhjE*	410	Codirectional+	3672399	3672808	79.15
174	*yjbM_dusA*	407	Codirectional+	4259330	4259736	91.8
175	*rpoE_nadB*	407	Divergent	2708035	2708441	58.37
176	*phoR_brnQ*	406	Codirectional+	418409	418814	9.02
177	*fdnI_yddM*	406	Convergent	1550016	1550421	33.41
178	*focA_ycaO*	405	Codirectional-	953690	954094	20.56
179	*yegR_yegS*	405	Divergent	2166331	2166735	46.69
180	*yrdE'_accB*	404	Codirectional+	3403054	3403457	73.35

Table 2.8 (continued). Two hundred longest intergenic regions in *E. coli* K-12.

Rank	Name	Length	Orientation	L_END	R_END	Cs
181	*glnH_dps*	403	Codirectional-	847228	847630	18.26
182	*yhjA_treF*	403	Divergent	3667212	3667614	79.04
183	*ydeN_ydeO*	401	Codirectional-	1580549	1580949	34.07
184	*yjbS_aphA*	401	Divergent	4267036	4267436	91.97
185	*aroK_hofQ*	400	Codirectional-	3517087	3517486	75.8
186	*agaI_yraH*	400	Codirectional+	3285048	3285447	70.8
187	*yjdM_yjdA*	400	Divergent	4324758	4325157	93.21
188	*ycfP_ndh*	399	Codirectional+	1164909	1165307	25.11
189	*ymgF_ycgH'*	399	Codirectional+	1218425	1218823	26.26
190	*yjjY_yjtD*	399	Codirectional+	4638566	4638964	99.98
191	*ugpB_livF*	398	Codirectional-	3590349	3590746	77.38
192	*sgcX_yjhY'*	398	Codirectional-	4529675	4530072	97.63
193	*alx_sstT*	398	Codirectional+	3237568	3237965	69.78
194	*envR_acrE*	398	Divergent	3411488	3411885	73.53
195	*speC_yqgA*	397	Divergent	3107178	3107574	66.97
196	*asnB_nagD*	396	Codirectional-	698401	698796	15.05
197	*yicT'_setC*	396	Codirectional+	3834580	3834975	82.65
198	*ybhK_moaA*	396	Divergent	815871	816266	17.58
199	*sspA_rpsI*	394	Codirectional-	3375443	3375836	72.75
200	*yjeJ_yjeK*	394	Codirectional-	4372258	4372651	94.24

Overall Length		Overall Divergent Length	
Mean	112.452	Mean	236.088
Median	67.000	Median	209.000
STDEVP	137.654	STDEVP	148.200

Table 2.9. CsgD regulon excluding *csgBAC.*

CsgD overexpression decreased levels of *cpsA, cpsB, cpsG, fecR, fhuE, gatA, gatC,* GatY, GatZ*, glnS, infA, metA, ompF, ompT, pyrB, pyrI, pepD, thyA,* TnaA, and *yagS* (14, 15).

CsgD overexpression increased levels of *adrA (yaiC),* Dps*, glyA,* GadA*, gsk, hmd, iraP,* OmpW, *osmB*, PflB*, recT*, WrbA, *yaiB,ydjC (chbG), yhiE, yjbR, yjgW, ymdA, yoaD,* and *ytfI,* levels (14, 15, 19, 20, 39).

CsgD also regulates bapA, a large membrane protein required for biofilm formation in *Salmonella enterica* serovar Enteritidis (65). Finally CsgD differentially regulates the divergent *yihVW* and *yihU-PyshA* operons, which are involved in O-antigen production in *Salmonella enterica* serovar Enteritidis (37).

Table 2.10. Comparison of curli associated genes with swimming and swarming motility associated genes (See Figure 2.13 for diagram).

Numbers in parenthesis indicate gene count for respective section of Venn diagram.

Without Table 2 (more defective swarming mutants (46))**:**
a (192): *aaeR, aceE, aegA, appC, argC, aroA, aroB, aroC, aroD, aroH, aroK, aroP, asnC, aspA, aspC, carA, carB, ccmA, clpA, clpP, clpX, cmr, cpxP, cpxR, csgA, csgB, csgD, csgE, csgF, csgG, cusB, cyaA, cysC, cysE, cysG, cysI, cytR, dam, ddpD, dksA, dnaG, eamA, envZ, essQ, fabF, fadJ, fbp, fhlA, flgM, fliT, fruR, gcvA, glmM, glvG, glxK, gnd, greA, hdfR, hfq, hybD, ihfA, ihfB, ilvG, kdpD, lipB, lon, lpp, lsrF, mdoG, mdoH, metC, miaA, mlrA, mltA, mtlR, nagA, nanE, nanK, narQ, nhaA, nlpD, nudC, nudH, nudL, nuoA, nuoB, nuoC, nuoE, nuoG, nuoJ, nuoK, nuoM, nuoN, nusB, ompC, ompF, pcnB, pdxH, pepT, perR, php, poxA, proW, purA, purD, purE, purF, purH, purL, purM, purR, putA, puuR, pyrB, pyrC, pyrD, pyrE, qseC, racC, rbfA, rcsB, rcsF, rimK, rimM, rnhA, rpoN, rpoS, rpoZ, rpsF, rpsT, rseA, rstA, rstB, rumA, sdhD, sdiA, slp, speB, srlR, srmB, sspA, sucC, surA, thyA, tolA, tomB, torI, tpiA, treR, trpA, trpB, trpC, trpD, trpE, ubiG, uspE, xapR, ybaM, ybcI, ybeZ, ybgC, ycfM, ycgV, yciC, yciM, ydaF, ydaM, ydcI, ydeJ, ydiT, yecH, yedQ, yedV, yeeN, yeiL, yeiR, yfeW, yfgC, yhbJ, yheV, yiaK, yicC, yieP, yiiS, yjjW, ykgH, ymgE, ynaK, yncG, ynjD, yoeE, yqaB*

b(47): *aroE, cpxA, crp, cysB, dnaT, dps, efp, fepB, fepD, fepG, glnA, gmhB, gmr, guaB, lipA, lpd, mdoC, nagC, ndk, nuoF, ompR, pal, ptsH, ptsI, purC, purK, pyrF, rfaF, rfaG, rfaP, rfe, rffA, rffC, rffT, sdhA, sdhB, sdhC, skp, sucA, sucD, tolC, tolQ, ubiE, ubiF, wzxE, ydaT, ygfZ*

c(26): *atpA, atpB, atpC, atpE, atpF, atpG, atpH, cmk, dnaK, fabH, fliI, galU, lpcA, lptB, pgm, priA, rfaC, rfaD, rfaE, rfaH, rodZ, rplA, rsgA, tolB, tolR, ubiH*

In order above:
COGs of b: E, T, T, k, l, P, j, P, P, P, E, M, T, F, H, C, M, G, F, C, T, M, G, G, F, F, F, M, M, M, M, M, M, M, C, C, C, M, C, C, N, N, H, H, M, S, R
COGs of c: C, C, C, C, C, C, C, F, O, I, N, M, M, M, G, l, M, M, M, M, M, J, J, N, N, H

Table 2.10. Comparison of curli associated genes with swimming and swarming motility associated genes (See Figure 2.13 for diagram).

Without Table 2 (continued):
d(52): *atpD, cheA, cheB, cheR, cheW, cheY, cheZ, dsbA, fis, flgA, flgB, flgC, flgD, flgE, flgF, flgG, flgI, flgJ, flgK, flgL, flgN, flhA, flhB, flhC, fliA, fliC, fliD, fliF, fliG, fliJ, fliK, fliM, fliN, fliO, fliP, fliQ, fliS, folB, folP, hflD, motA, motB, mtn, rcsC, rho, rluD, rrmJ, sufC, tap, ubiX, ycjW, yhcA*

e(169)*:* *ackA, acnA, ahpF, aldA, argG, argP, argR, arnF, arpB, asmA, asnB, atoS, bamB, barA, chaC, csdA, cusR, cvpA, cvrA, cydD, cyoA, dapF, deaD, dgkA, dsbB, entB, entE, entF, envC, envR, etk, eutS, evgS, fadK, fadL, fdrA, fdx, feoA, feoB, fepA, fes, fimA, fimB, fimC, fimD, fimF, fimH, fliR, fucU, galE, galM, gapC, gatY, gpmI, grxB, hemE, hflC, hipB, hscA, hspQ, htpG, htpX, hyaA, hyaB, idaB, iscS, iscU, lplT, macA, marA, mdtH, metL, mglB, mgsA, mpaA, mrcB, murR, narH, ninE, nlpI, ogt, ompA, oppC, osmB, osmC, pabC, pdxJ, pdxY, pgi, pldB, ppdC, ppiB, ppiD, ppk, prmB, prpD, puuC, rep, rfaB, rfaI, rfaJ, rfaQ, rfaS, rfaY, rfaZ, rffD, rffE, rffH, rffM, rof, rpiA, rssB, rsxG, sapF, sucB, sulA, tpx, treA, trkA, trmE, trpL, truA, trxB, tsx, ubiC, uxuR, wcaE, wcaH, xerC, xerD, yaiW, ybaK, ybdH, ybdO, ybeD, ybiA, yccK, ycdY, yciG, ycjZ, ydbA, ydcX, ydcZ, ydeE, ydfT, ydgA, yeeF, yehP, yeiA, yfeH, yfeR, yfgJ, yfjN, yfjP, ygfH, yhcH, yheL, yheM, yheN, yigG, ymfA, yncH, yncN, yneE, ynfG, yqaA, yqaD, yrbE, zwf*

With Table 2 (includes less defective swarming mutants (46))**:**
a(241): *aaeR, aceE, aegA, appC, argC, aroA, aroB, aroC, aroD, aroH, aroK, aroP, asnC, aspA, aspC, bhsA, btuC, carA, carB, ccmA, clpA, clpP, clpX, cmr, cpxP, cpxR, csgA, csgB, csgD, csgE, csgF, csgG, cusB, cyaA, cysC, cysE, cysG, cysH, cysI, cytR, dam, dcuC, ddpD, dksA, dnaG, dnaQ, eamA, envZ, essQ, fabF, fadJ, fbp, fepE, fhlA, flgM, fliT, fruR, gcvA, glcD, glmM, glvG, glxK, gnd, greA, gshB, gsiA, hdfR, hepA, hfq, hybD, hycI, hypE, ihfA, ihfB, ilvG, kbl, kdpD, lipB, lon, lpp, lsrF, manY, mdoG, mdoH, metC, mfd, miaA, mlrA, mltA, mppA, mtlR, mutL, nagA, nanE, nanK, narQ, nhaA, nlpD, nudC, nudH, nudL, nuoA, nuoB, nuoC, nuoE, nuoG, nuoH, nuoI, nuoJ, nuoK, nuoL, nuoM, nuoN, nusB, ompC, ompF, pcnB, pdxB, pdxH, pepT, perR, php, poxA, proW, purA, purD, purE, purF, purH, purL, purM, purN, purR, putA, puuR, pyrB, pyrC, pyrD, pyrE, qseB, qseC, racC, rbfA, rbsK, rbsR, rcsB, rcsF, rdgC, rimK, rimM, rlmF, rnhA, rnk, rpoN, rpoS, rpoZ, rpsF, rpsT, rseA, rstA, rstB, rsxC, rumA, sbcD, sdhD, sdiA, sfcA, slp, speB, srlR, srmB, ssnA, sspA, sucC, surA, thyA, tolA, tomB, tonB, torI, tpiA, treR, trpA, trpB, trpC, trpD, trpE, ubiG, uspE, uup, wcaD, xapR, ybaB, ybaM, ybcI, ybeZ, ybgC, ybjN, ycbL, ycbQ, ycfM, ycgV, yciC, yciM, yciU, ydaF, ydaM, ydcI, ydeJ, ydeP, ydfD, ydiT, ydiY, ydjI, yebY, yecH, yedQ, yedV, yeeN, yehE, yeiL, yeiR, yfeW, yfgC, yfjW, yhbJ, yheV, yiaK, yicC, yieP, yiiS, yjbI, yjjW, ykgH, ymgE, ynaK, yncG, ynfH, ynjD, yoeE, yqaB*

b(48): *aroE, cpxA, crp, cysB, dnaT, dps, efp, fepB, fepD, fepG, glnA, gmhB, gmr, guaB, lipA, lpd, mdoC, nagC, ndk, nuoF, ompR, pal, ptsH, ptsI, purC, purK, pyrF, rfaF, rfaG, rfaP, rfe, rffA, rffC, rffT, sdhA, sdhB, sdhC, skp, sucA, sucB, sucD, tolC, tolQ, ubiE, ubiF, wzxE, ydaT, ygfZ*

Table 2.10. Comparison of curli associated genes with swimming and swarming motility associated genes (See Figure 2.13 for diagram).

With Table 2 (continued):
c(28): *atpA, atpB, atpC, atpE, atpF, atpG, atpH, cmk, dnaK, fabH, flhC, fliG, fliI, galU, lpcA, lptB, pgm, priA, rfaC, rfaD, rfaE, rfaH, rodZ, rplA, rsgA, tolB, tolR, ubiH*

d(50): *atpD, cheA, cheB, cheR, cheW, cheY, cheZ, dsbA, fis, flgA, flgB, flgC, flgD, flgE, flgF, flgG, flgI, flgJ, flgK, flgL, flgN, flhA, flhB, fliA, fliC, fliD, fliF, fliJ, fliK, fliM, fliN, fliO, fliP, fliQ, fliS, folB, folP, hflD, motA, motB, mtn, rcsC, rho, rluD, rrmJ, sufC, tap, ubiX, ycjW, yhcA*

e(168): *ackA, acnA, ahpF, aldA, argG, argP, argR, arnF, arpB, asmA, asnB, atoS, bamB, barA, chaC, csdA, cusR, cvpA, cvrA, cydD, cyoA, dapF, deaD, dgkA, dsbB, entB, entE, entF, envC, envR, etk, eutS, evgS, fadK, fadL, fdrA, fdx, feoA, feoB, fepA, fes, fimA, fimB, fimC, fimD, fimF, fimH, fliR, fucU, galE, galM, gapC, gatY, gpmI, grxB, hemE, hflC, hipB, hscA, hspQ, htpG, htpX, hyaA, hyaB, idaB, iscS, iscU, lplT, macA, marA, mdtH, metL, mglB, mgsA, mpaA, mrcB, murR, narH, ninE, nlpI, ogt, ompA, oppC, osmB, osmC, pabC, pdxJ, pdxY, pgi, pldB, ppdC, ppiB, ppiD, ppk, prmB, prpD, puuC, rep, rfaB, rfaI, rfaJ, rfaQ, rfaS, rfaY, rfaZ, rffD, rffE, rffH, rffM, rof, rpiA, rssB, rsxG, sapF, sulA, tpx, treA, trkA, trmE, trpL, truA, trxB, tsx, ubiC, uxuR, wcaE, wcaH, xerC, xerD, yaiW, ybaK, ybdH, ybdO, ybeD, ybiA, yccK, ycdY, yciG, ycjZ, ydbA, ydcX, ydcZ, ydeE, ydfT, ydgA, yeeF, yehP, yeiA, yfeH, yfeR, yfgJ, yfjN, yfjP, ygfH, yhcH, yheL, yheM, yheN, yigG, ymfA, yncH, yncN, yneE, ynfG, yqaA, yqaD, yrbE, zwf*

References

1. **Anriany, Y., S. N. Sahu, K. R. Wessels, L. M. McCann, and S. W. Joseph.** 2006. Alteration of the rugose phenotype in waaG and ddhC mutants of Salmonella enterica serovar Typhimurium DT104 is associated with inverse production of curli and cellulose. Appl Environ Microbiol **72:**5002-12.
2. **Arnqvist, A., A. Olsen, and S. Normark.** 1994. Sigma S-dependent growth-phase induction of the csgBA promoter in Escherichia coli can be achieved in vivo by sigma 70 in the absence of the nucleoid-associated protein H-NS. Mol Microbiol **13:**1021-32.
3. **Atlung, T., K. Knudsen, L. Heerfordt, and L. Brondsted.** 1997. Effects of sigmaS and the transcriptional activator AppY on induction of the Escherichia coli hya and cbdAB-appA operons in response to carbon and phosphate starvation. J Bacteriol **179:**2141-6.
4. **Baba, T., T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori.** 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol **2:**2006 0008.
5. **Barak, J. D., L. Gorski, P. Naraghi-Arani, and A. O. Charkowski.** 2005. Salmonella enterica virulence genes are required for bacterial attachment to plant tissue. Appl Environ Microbiol **71:**5685-91.
6. **Barnhart, M. M., and M. R. Chapman.** 2006. Curli biogenesis and function. Annu Rev Microbiol **60:**131-47.
7. **Barnhart, M. M., J. Lynem, and M. R. Chapman.** 2006. GlcNAc-6P levels modulate the expression of Curli fibers by Escherichia coli. J Bacteriol **188:**5212-9.
8. **Bian, Z., A. Brauner, Y. Li, and S. Normark.** 2000. Expression of and cytokine activation by Escherichia coli curli fibers in human sepsis. J Infect Dis **181:**602-12.
9. **Bian, Z., and S. Normark.** 1997. Nucleator function of CsgB for the assembly of adhesive surface organelles in Escherichia coli. Embo J **16:**5827-36.
10. **Bohringer, J., D. Fischer, G. Mosler, and R. Hengge-Aronis.** 1995. UDP-glucose is a potential intracellular signal molecule in the control of expression of sigma S and sigma S-dependent genes in Escherichia coli. J Bacteriol **177:**413-22.
11. **Boos, W., U. Ehmann, H. Forkl, W. Klein, M. Rimmele, and P. Postma.** 1990. Trehalose transport and metabolism in Escherichia coli. J Bacteriol **172:**3450-61.
12. **Bougdour, A., C. Lelong, and J. Geiselmann.** 2004. Crl, a low temperature-induced protein in Escherichia coli that binds directly to the stationary phase sigma subunit of RNA polymerase. J Biol Chem **279:**19540-50.
13. **Bougdour, A., S. Wickner, and S. Gottesman.** 2006. Modulating RssB activity: IraP, a novel regulator of sigma(S) stability in Escherichia coli. Genes Dev **20:**884-97.
14. **Brombacher, E., A. Baratto, C. Dorel, and P. Landini.** 2006. Gene expression regulation by the Curli activator CsgD protein: modulation of cellulose biosynthesis and control of negative determinants for microbial adhesion. J Bacteriol **188:**2027-37.

15. **Brombacher, E., C. Dorel, A. J. Zehnder, and P. Landini.** 2003. The curli biosynthesis regulator CsgD co-ordinates the expression of both positive and negative determinants for biofilm formation in Escherichia coli. Microbiology **149:**2847-57.
16. **Brown, L., D. Gentry, T. Elliott, and M. Cashel.** 2002. DksA affects ppGpp induction of RpoS at a translational level. J Bacteriol **184:**4455-65.
17. **Brown, P. K., C. M. Dozois, C. A. Nickerson, A. Zuppardo, J. Terlonge, and R. Curtiss, 3rd.** 2001. MlrA, a novel regulator of curli (AgF) and extracellular matrix synthesis by Escherichia coli and Salmonella enterica serovar Typhimurium. Mol Microbiol **41:**349-63.
18. **Chapman, M. R., L. S. Robinson, J. S. Pinkner, R. Roth, J. Heuser, M. Hammar, S. Normark, and S. J. Hultgren.** 2002. Role of Escherichia coli curli operons in directing amyloid fiber formation. Science **295:**851-5.
19. **Chirwa, N. T., and M. B. Herrington.** 2003. CsgD, a regulator of curli and cellulose synthesis, also regulates serine hydroxymethyltransferase synthesis in Escherichia coli K-12. Microbiology **149:**525-35.
20. **Chirwa, N. T., and M. B. Herrington.** 2004. Role of MetR and PurR in the activation of glyA by CsgD in Escherichia coli K-12. Can J Microbiol **50:**683-90.
21. **Coleman, W. G., Jr.** 1983. The rfaD gene codes for ADP-L-glycero-D-mannoheptose-6-epimerase. An enzyme required for lipopolysaccharide core biosynthesis. J Biol Chem **258:**1985-90.
22. **Collinson, S. K., P. C. Doig, J. L. Doran, S. Clouthier, T. J. Trust, and W. W. Kay.** 1993. Thin, aggregative fimbriae mediate binding of Salmonella enteritidis to fibronectin. J Bacteriol **175:**12-8.
23. **Collinson, S. K., L. Emody, K. H. Muller, T. J. Trust, and W. W. Kay.** 1991. Purification and characterization of thin, aggregative fimbriae from Salmonella enteritidis. J Bacteriol **173:**4773-81.
24. **Da Re, S., and J. M. Ghigo.** 2006. A CsgD-independent pathway for cellulose production and biofilm formation in Escherichia coli. J Bacteriol **188:**3073-87.
25. **Danese, P. N., G. R. Oliver, K. Barr, G. D. Bowman, P. D. Rick, and T. J. Silhavy.** 1998. Accumulation of the enterobacterial common antigen lipid II biosynthetic intermediate stimulates degP transcription in Escherichia coli. J Bacteriol **180:**5875-84.
26. **Datsenko, K. A., and B. L. Wanner.** 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A **97:**6640-5.
27. **Dorel, C., O. Vidal, C. Prigent-Combaret, I. Vallet, and P. Lejeune.** 1999. Involvement of the Cpx signal transduction pathway of E. coli in biofilm formation. FEMS Microbiol Lett **178:**169-75.
28. **Dyszel, J. L., J. A. Soares, M. C. Swearingen, A. Lindsay, J. N. Smith, and B. M. Ahmer.** E. coli K-12 and EHEC genes regulated by SdiA. PLoS One **5:**e8946.
29. **Ebel, W., G. J. Vaughn, H. K. Peters, 3rd, and J. E. Trempy.** 1997. Inactivation of mdoH leads to increased expression of colanic acid capsular polysaccharide in Escherichia coli. J Bacteriol **179:**6858-61.

30. **Epstein, E. A., M. A. Reizian, and M. R. Chapman.** 2009. Spatial clustering of the curlin secretion lipoprotein requires curli fiber assembly. J Bacteriol **191:**608-15.
31. **Ferrieres, L., and D. J. Clarke.** 2003. The RcsC sensor kinase is required for normal biofilm formation in Escherichia coli K-12 and controls the expression of a regulon in response to growth on a solid surface. Mol Microbiol **50:**1665-82.
32. **Genevaux, P., P. Bauda, M. S. DuBow, and B. Oudega.** 1999. Identification of Tn10 insertions in the rfaG, rfaP, and galU genes involved in lipopolysaccharide core biosynthesis that affect Escherichia coli adhesion. Arch Microbiol **172:**1-8.
33. **Georgellis, D., O. Kwon, and E. C. Lin.** 2001. Quinones as the redox signal for the arc two-component system of bacteria. Science **292:**2314-6.
34. **Gerstel, U., C. Park, and U. Romling.** 2003. Complex regulation of csgD promoter activity by global regulatory proteins. Mol Microbiol **49:**639-54.
35. **Gerstel, U., and U. Romling.** 2001. Oxygen tension and nutrient starvation are major signals that regulate agfD promoter activity and expression of the multicellular morphotype in Salmonella typhimurium. Environ Microbiol **3:**638-48.
36. **Gibson, D. L., A. P. White, C. M. Rajotte, and W. W. Kay.** 2007. AgfC and AgfE facilitate extracellular thin aggregative fimbriae synthesis in Salmonella enteritidis. Microbiology **153:**1131-40.
37. **Gibson, D. L., A. P. White, S. D. Snyder, S. Martin, C. Heiss, P. Azadi, M. Surette, and W. W. Kay.** 2006. Salmonella produces an O-antigen capsule regulated by AgfD and important for environmental persistence. J Bacteriol **188:**7722-30.
38. **Gophna, U., M. Barlev, R. Seijffers, T. A. Oelschlager, J. Hacker, and E. Z. Ron.** 2001. Curli fibers mediate internalization of Escherichia coli by eukaryotic cells. Infect Immun **69:**2659-65.
39. **Gualdi, L., L. Tagliabue, and P. Landini.** 2007. Biofilm formation-gene expression relay system in Escherichia coli: modulation of sigmaS-dependent gene expression by the CsgD regulatory protein via sigmaS protein stabilization. J Bacteriol **189:**8034-43.
40. **Hammar, M.** 1997 Assembly and adhesive properties of curli. . Ph.D. thesis at Karolinska Institute, Stockholm, Sweden.
41. **Hammar, M., A. Arnqvist, Z. Bian, A. Olsen, and S. Normark.** 1995. Expression of two csg operons is required for production of fibronectin- and congo red-binding curli polymers in Escherichia coli K-12. Mol Microbiol **18:**661-70.
42. **Hammar, M., Z. Bian, and S. Normark.** 1996. Nucleator-dependent intercellular assembly of adhesive curli organelles in Escherichia coli. Proc Natl Acad Sci U S A **93:**6562-6.
43. **Harshey, R. M.** 2003. Bacterial motility on a surface: many ways to a common goal. Annu Rev Microbiol **57:**249-73.
44. **Hengge-Aronis, R.** 2002. Signal transduction and regulatory mechanisms involved in control of the sigma(S) (RpoS) subunit of RNA polymerase. Microbiol Mol Biol Rev **66:**373-95, table of contents.

45. **Holmqvist, E., J. Reimegard, M. Sterk, N. Grantcharova, U. Romling, and E. G. Wagner.** 2010. Two antisense RNAs target the transcriptional regulator CsgD to inhibit curli synthesis. Embo J **29:**1840-50.
46. **Inoue, T., R. Shingaki, S. Hirose, K. Waki, H. Mori, and K. Fukui.** 2007. Genome-wide screening of genes required for swarming motility in Escherichia coli K-12. J Bacteriol **189:**950-7.
47. **Ishihama, A.** 2010. Prokaryotic genome regulation: multifactor promoters, multitarget regulators and hierarchic networks. FEMS Microbiol Rev.
48. **Jensen, K. F.** 1993. The Escherichia coli K-12 "wild types" W3110 and MG1655 have an rph frameshift mutation that leads to pyrimidine starvation due to low pyrE expression levels. J Bacteriol **175:**3401-7.
49. **Jensen, L. J., M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering.** 2009. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res **37:**D412-6.
50. **Jeter, C., and A. G. Matthysse.** 2005. Characterization of the binding of diarrheagenic strains of E. coli to plant surfaces and the role of curli in the interaction of the bacteria with alfalfa sprouts. Mol Plant Microbe Interact **18:**1235-42.
51. **Jiang, S. S., T. Y. Lin, W. B. Wang, M. C. Liu, P. R. Hsueh, and S. J. Liaw.** 2010. Characterization of UDP-glucose dehydrogenase and UDP-glucose pyrophosphorylase mutants of Proteus mirabilis: defectiveness in polymyxin B resistance, swarming, and virulence. Antimicrob Agents Chemother **54:**2000-9.
52. **Joloba, M. L., K. M. Clemmer, D. D. Sledjeski, and P. N. Rather.** 2004. Activation of the gab operon in an RpoS-dependent manner by mutations that truncate the inner core of lipopolysaccharide in Escherichia coli. J Bacteriol **186:**8542-6.
53. **Jubelin, G., A. Vianney, C. Beloin, J. M. Ghigo, J. C. Lazzaroni, P. Lejeune, and C. Dorel.** 2005. CpxR/OmpR interplay regulates curli gene expression in response to osmolarity in Escherichia coli. J Bacteriol **187:**2038-49.
54. **Kader, A., R. Simm, U. Gerstel, M. Morr, and U. Romling.** 2006. Hierarchical involvement of various GGDEF domain proteins in rdar morphotype development of Salmonella enterica serovar Typhimurium. Mol Microbiol **60:**602-16.
55. **Kalamorz, F., B. Reichenbach, W. Marz, B. Rak, and B. Gorke.** 2007. Feedback control of glucosamine-6-phosphate synthase GlmS expression depends on the small RNA GlmZ and involves the novel protein YhbJ in Escherichia coli. Mol Microbiol **65:**1518-33.
56. **Karp, P. D., I. M. Keseler, A. Shearer, M. Latendresse, M. Krummenacker, S. M. Paley, I. Paulsen, J. Collado-Vides, S. Gama-Castro, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penaloza-Spinola, C. Bonavides-Martinez, and J. Ingraham.** 2007. Multidimensional annotation of the Escherichia coli K-12 genome. Nucleic Acids Res **35:**7577-90.

57. **Kikuchi, T., Y. Mizunoe, A. Takade, S. Naito, and S. Yoshida.** 2005. Curli fibers are required for development of biofilm architecture in Escherichia coli K-12 and enhance bacterial adherence to human uroepithelial cells. Microbiol Immunol **49:**875-84.
58. **Kim, S. H., and Y. H. Kim.** 2004. Escherichia coli O157:H7 adherence to HEp-2 cells is implicated with curli expression and outer membrane integrity. J Vet Sci **5:**119-24.
59. **Klein, G., B. Lindner, W. Brabetz, H. Brade, and S. Raina.** 2009. Escherichia coli K-12 Suppressor-free Mutants Lacking Early Glycosyltransferases and Late Acyltransferases: minimal lipopolysaccharide structure and induction of envelope stress response. J Biol Chem **284:**15369-89.
60. **Kneidinger, B., C. Marolda, M. Graninger, A. Zamyatina, F. McArthur, P. Kosma, M. A. Valvano, and P. Messner.** 2002. Biosynthesis pathway of ADP-L-glycero-beta-D-manno-heptose in Escherichia coli. J Bacteriol **184:**363-9.
61. **Komeda, Y., T. Icho, and T. Iino.** 1977. Effects of galU mutation on flagellar formation in Escherichia coli. J Bacteriol **129:**908-15.
62. **Kostakioti, M., M. Hadjifrangiskou, J. S. Pinkner, and S. J. Hultgren.** 2009. QseC-mediated dephosphorylation of QseB is required for expression of genes associated with virulence in uropathogenic Escherichia coli. Mol Microbiol **73:**1020-31.
63. **Kuznetsova, E., M. Proudfoot, C. F. Gonzalez, G. Brown, M. V. Omelchenko, I. Borozan, L. Carmel, Y. I. Wolf, H. Mori, A. V. Savchenko, C. H. Arrowsmith, E. V. Koonin, A. M. Edwards, and A. F. Yakunin.** 2006. Genome-wide analysis of substrate specificities of the Escherichia coli haloacid dehalogenase-like phosphatase family. J Biol Chem **281:**36149-61.
64. **Lange, R., D. Fischer, and R. Hengge-Aronis.** 1995. Identification of transcriptional start sites and the role of ppGpp in the expression of rpoS, the structural gene for the sigma S subunit of RNA polymerase in Escherichia coli. J Bacteriol **177:**4676-80.
65. **Latasa, C., A. Roux, A. Toledo-Arana, J. M. Ghigo, C. Gamazo, J. R. Penades, and I. Lasa.** 2005. BapA, a large secreted protein required for biofilm formation and host colonization of Salmonella enterica serovar Enteritidis. Mol Microbiol **58:**1322-39.
66. **Lee, J., T. Maeda, S. H. Hong, and T. K. Wood.** 2009. Reconfiguring the quorum-sensing regulator SdiA of Escherichia coli to control biofilm formation via indole and N-acylhomoserine lactones. Appl Environ Microbiol **75:**1703-16.
67. **Lipman, D. J., A. Souvorov, E. V. Koonin, A. R. Panchenko, and T. A. Tatusova.** 2002. The relationship of protein conservation and sequence length. BMC Evol Biol **2:**20.
68. **Loewen, P. C., B. Hu, J. Strutinsky, and R. Sparling.** 1998. Regulation in the rpoS regulon of Escherichia coli. Can J Microbiol **44:**707-17.

69. **Loferer, H., M. Hammar, and S. Normark.** 1997. Availability of the fibre subunit CsgA and the nucleator protein CsgB during assembly of fibronectin-binding curli is limited by the intracellular concentration of the novel lipoprotein CsgG. Mol Microbiol **26:**11-23.
70. **Ma, Q., and T. K. Wood.** 2009. OmpA influences Escherichia coli biofilm formation by repressing cellulose production through the CpxRA two-component system. Environ Microbiol **11:**2735-46.
71. **Malpica, R., B. Franco, C. Rodriguez, O. Kwon, and D. Georgellis.** 2004. Identification of a quinone-sensitive redox switch in the ArcB sensor kinase. Proc Natl Acad Sci U S A **101:**13318-23.
72. **Marolda, C. L., P. Lahiry, E. Vines, S. Saldias, and M. A. Valvano.** 2006. Micromethods for the characterization of lipid A-core and O-antigen lipopolysaccharide. Methods Mol Biol **347:**237-52.
73. **Mengin-Lecreulx, D., and J. van Heijenoort.** 1996. Characterization of the essential gene glmM encoding phosphoglucosamine mutase in Escherichia coli. J Biol Chem **271:**32-9.
74. **Mika, F., and R. Hengge.** 2005. A two-component phosphotransfer network involving ArcB, ArcA, and RssB coordinates synthesis and proteolysis of sigmaS (RpoS) in E. coli. Genes Dev **19:**2770-81.
75. **Misra, R. V., R. S. Horler, W. Reindl, Goryanin, II, and G. H. Thomas.** 2005. EchoBASE: an integrated post-genomic database for Escherichia coli. Nucleic Acids Res **33:**D329-33.
76. **Missiakas, D., J. M. Betton, and S. Raina.** 1996. New components of protein folding in extracytoplasmic compartments of Escherichia coli SurA, FkpA and Skp/OmpH. Mol Microbiol **21:**871-84.
77. **Nanchen, A., A. Schicker, O. Revelles, and U. Sauer.** 2008. Cyclic AMP-dependent catabolite repression is the dominant control mechanism of metabolic fluxes under glucose limitation in Escherichia coli. J Bacteriol **190:**2323-30.
78. **Neidhardt, F. C., R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low Jr., B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger.** 1996. Escherichia coli and Salmonella: Cellular and Molecular Biology, 2nd ed, vol. ASM Press, Washington, D.C.
79. **Nenninger, A. A., L. S. Robinson, and S. J. Hultgren.** 2009. Localized and efficient curli nucleation requires the chaperone-like amyloid assembly protein CsgF. Proc Natl Acad Sci U S A **106:**900-5.
80. **Niba, E. T., Y. Naka, M. Nagase, H. Mori, and M. Kitakawa.** 2007. A genome-wide approach to identify the genes involved in biofilm formation in E. coli. DNA Res **14:**237-46.
81. **Ogasawara, H., A. Hasegawa, E. Kanda, T. Miki, K. Yamamoto, and A. Ishihama.** 2007. Genomic SELEX search for target promoters under the control of the PhoQP-RstBA signal relay cascade. J Bacteriol **189:**4791-9.

82. **Ogasawara, H., K. Yamada, A. Kori, K. Yamamoto, and A. Ishihama.** 2010. Regulation of the E. coli csgD Promoter: Interplay between Five Transcription Factors. Microbiology.
83. **Olsen, A., A. Arnqvist, M. Hammar, S. Sukupolvi, and S. Normark.** 1993. The RpoS sigma factor relieves H-NS-mediated transcriptional repression of csgA, the subunit gene of fibronectin-binding curli in Escherichia coli. Mol Microbiol **7:**523-36.
84. **Parker, C. T., A. W. Kloser, C. A. Schnaitman, M. A. Stein, S. Gottesman, and B. W. Gibson.** 1992. Role of the rfaG and rfaP genes in determining the lipopolysaccharide core structure and cell surface properties of Escherichia coli K-12. J Bacteriol **174:**2525-38.
85. **Pedersen, A. G., L. J. Jensen, S. Brunak, H. H. Staerfeldt, and D. W. Ussery.** 2000. A DNA structural atlas for Escherichia coli. J Mol Biol **299:**907-30.
86. **Perrenoud, A., and U. Sauer.** 2005. Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in Escherichia coli. J Bacteriol **187:**3171-9.
87. **Pesavento, C., G. Becker, N. Sommerfeldt, A. Possling, N. Tschowri, A. Mehlis, and R. Hengge.** 2008. Inverse regulatory coordination of motility and curli-mediated adhesion in Escherichia coli. Genes Dev **22:**2434-46.
88. **Pirooznia, M., V. Nagarajan, and Y. Deng.** 2007. GeneVenn - A web application for comparing gene lists using Venn diagrams. Bioinformation **1:**420-2.
89. **Pradel, E., and C. A. Schnaitman.** 1991. Effect of rfaH (sfrB) and temperature on expression of rfa genes of Escherichia coli K-12. J Bacteriol **173:**6428-31.
90. **Pratt, L. A., and T. J. Silhavy.** 1998. Crl stimulates RpoS activity during stationary phase. Mol Microbiol **29:**1225-36.
91. **Raetz, C. R., and C. Whitfield.** 2002. Lipopolysaccharide endotoxins. Annu Rev Biochem **71:**635-700.
92. **Rahman, A., K. Barr, and P. D. Rick.** 2001. Identification of the structural gene for the TDP-Fuc4NAc:lipid II Fuc4NAc transferase involved in synthesis of enterobacterial common antigen in Escherichia coli K-12. J Bacteriol **183:**6509-16.
93. **Rahman, M., M. R. Hasan, T. Oba, and K. Shimizu.** 2006. Effect of rpoS gene knockout on the metabolism of Escherichia coli during exponential growth phase and early stationary phase based on gene expressions, enzyme activities and intracellular metabolite concentrations. Biotechnol Bioeng **94:**585-95.
94. **Reichenbach, B., A. Maes, F. Kalamorz, E. Hajnsdorf, and B. Gorke.** 2008. The small RNA GlmY acts upstream of the sRNA GlmZ in the activation of glmS expression and is subject to regulation by polyadenylation in Escherichia coli. Nucleic Acids Res **36:**2570-80.
95. **Rhodius, V. A., W. C. Suh, G. Nonaka, J. West, and C. A. Gross.** 2006. Conserved and variable functions of the sigmaE stress response in related genomes. PLoS Biol **4:**e2.

96. **Rick, P. D., K. Barr, K. Sankaran, J. Kajimura, J. S. Rush, and C. J. Waechter.** 2003. Evidence that the wzxE gene of Escherichia coli K-12 encodes a protein involved in the transbilayer movement of a trisaccharide-lipid intermediate in the assembly of enterobacterial common antigen. J Biol Chem **278:**16534-42.
97. **Robbe-Saule, V., M. D. Lopes, A. Kolb, and F. Norel.** 2007. Physiological effects of Crl in Salmonella are modulated by sigmaS level and promoter specificity. J Bacteriol **189:**2976-87.
98. **Robinson, L. S., E. M. Ashman, S. J. Hultgren, and M. R. Chapman.** 2006. Secretion of curli fibre subunits is mediated by the outer membrane-localized CsgG protein. Mol Microbiol **59:**870-81.
99. **Romling, U., Z. Bian, M. Hammar, W. D. Sierralta, and S. Normark.** 1998. Curli fibers are highly conserved between Salmonella typhimurium and Escherichia coli with respect to operon structure and regulation. J Bacteriol **180:**722-31.
100. **Romling, U., W. D. Sierralta, K. Eriksson, and S. Normark.** 1998. Multicellular and aggregative behaviour of Salmonella typhimurium strains is controlled by mutations in the agfD promoter. Mol Microbiol **28:**249-64.
101. **Rosenberg, M.** 1984. Ammonium sulphate enhances adherence of *Escherichia coli* J-5 to hydrocarbon and polystyrene, p. 41-45. FEMS Microbiol. Lett., vol. 25.
102. **Rosenberg, M.** 2006. Microbial adhesion to hydrocarbons: twenty-five years of doing MATH. FEMS Microbiol Lett **262:**129-34.
103. **Rosenberg, M., D. Gutnick, and E. Rosenberg.** 1980. Adherence of bacteria to hydrocarbons: a simple method for measuring cell surface hydrophobicity. FEMS Microbiol. Lett. **9:**29-33.
104. **Rubino, S. D., H. Nyunoya, and C. J. Lusty.** 1987. In vivo synthesis of carbamyl phosphate from NH3 by the large subunit of Escherichia coli carbamyl phosphate synthetase. J Biol Chem **262:**4382-6.
105. **Rudd, K. E.** 2000. EcoGene: a genome sequence database for Escherichia coli K-12. Nucleic Acids Res **28:**60-4.
106. **Ryu, J. H., and L. R. Beuchat.** 2005. Biofilm formation by Escherichia coli O157:H7 on stainless steel: effect of exopolysaccharide and Curli production on its resistance to chlorine. Appl Environ Microbiol **71:**247-54.
107. **Saier, M. H., Jr., and T. M. Ramseier.** 1996. The catabolite repressor/activator (Cra) protein of enteric bacteria. J Bacteriol **178:**3411-7.
108. **Saldana, Z., J. Xicohtencatl-Cortes, F. Avelino, A. D. Phillips, J. B. Kaper, J. L. Puente, and J. A. Giron.** 2009. Synergistic role of curli and cellulose in cell adherence and biofilm formation of attaching and effacing Escherichia coli and identification of Fis as a negative regulator of curli. Environ Microbiol **11:**992-1006.
109. **Schnaitman, C. A., and J. D. Klena.** 1993. Genetics of lipopolysaccharide biosynthesis in enteric bacteria. Microbiol Rev **57:**655-82.
110. **Serio, T. R., A. G. Cashikar, A. S. Kowal, G. J. Sawicki, J. J. Moslehi, L. Serpell, M. F. Arnsdorf, and S. L. Lindquist.** 2000. Nucleated conformational conversion and the replication of conformational information by a prion determinant. Science **289:**1317-21.

111. **Sevcik, M., A. Sebkova, J. Volf, and I. Rychlik.** 2001. Transcription of arcA and rpoS during growth of Salmonella typhimurium under aerobic and microaerobic conditions. Microbiology **147:**701-8.
112. **Shiomi, D., M. Sakai, and H. Niki.** 2008. Determination of bacterial rod shape by a novel cytoskeletal membrane protein. Embo J **27:**3081-91.
113. **Simm, R., A. Lusch, A. Kader, M. Andersson, and U. Romling.** 2007. Role of EAL-containing proteins in multicellular behavior of Salmonella enterica serovar Typhimurium. J Bacteriol **189:**3613-23.
114. **Solano, C., B. Garcia, C. Latasa, A. Toledo-Arana, V. Zorraquino, J. Valle, J. Casals, E. Pedroso, and I. Lasa.** 2009. Genetic reductionist approach for dissecting individual roles of GGDEF proteins within the c-di-GMP signaling network in Salmonella. Proc Natl Acad Sci U S A **106:**7997-8002.
115. **Sommerfeldt, N., A. Possling, G. Becker, C. Pesavento, N. Tschowri, and R. Hengge.** 2009. Gene expression patterns and differential input into curli fimbriae regulation of all GGDEF/EAL domain proteins in Escherichia coli. Microbiology **155:**1318-31.
116. **Sondheimer, N., and S. Lindquist.** 2000. Rnq1: an epigenetic modifier of protein function in yeast. Mol Cell **5:**163-72.
117. **Stenutz, R., A. Weintraub, and G. Widmalm.** 2006. The structures of Escherichia coli O-polysaccharide antigens. FEMS Microbiol Rev **30:**382-403.
118. **String.** http://string.embl.de/.
119. **Tam, C., and D. Missiakas.** 2005. Changes in lipopolysaccharide structure induce the sigma(E)-dependent response of Escherichia coli. Mol Microbiol **55:**1403-12.
120. **Tamayo, R., J. T. Pratt, and A. Camilli.** 2007. Roles of cyclic diguanylate in the regulation of bacterial pathogenesis. Annu Rev Microbiol **61:**131-48.
121. **Tschowri, N., S. Busse, and R. Hengge.** 2009. The BLUF-EAL protein YcgF acts as a direct anti-repressor in a blue-light response of Escherichia coli. Genes Dev **23:**522-34.
122. **Uhlich, G. A., P. H. Cooke, and E. B. Solomon.** 2006. Analyses of the red-dry-rough phenotype of an Escherichia coli O157:H7 strain and its role in biofilm formation and resistance to antibacterial agents. Appl Environ Microbiol **72:**2564-72.
123. **Uhlich, G. A., N. W. t. Gunther, D. O. Bayles, and D. A. Mosier.** 2009. The CsgA and Lpp proteins of an Escherichia coli O157:H7 strain affect HEp-2 cell invasion, motility, and biofilm formation. Infect Immun **77:**1543-52.
124. **Urban, J. H., and J. Vogel.** 2008. Two seemingly homologous noncoding RNAs act hierarchically to activate glmS mRNA translation. PLoS Biol **6:**e64.
125. **Vianney, A., G. Jubelin, S. Renault, C. Dorel, P. Lejeune, and J. C. Lazzaroni.** 2005. Escherichia coli tol and rcs genes participate in the complex network affecting curli synthesis. Microbiology **151:**2487-97.
126. **Vidal, O., R. Longin, C. Prigent-Combaret, C. Dorel, M. Hooreman, and P. Lejeune.** 1998. Isolation of an Escherichia coli K-12 mutant strain able to form biofilms on inert surfaces: involvement of a new ompR allele that increases curli expression. J Bacteriol **180:**2442-9.

127. **Vines, E. D., C. L. Marolda, A. Balachandran, and M. A. Valvano.** 2005. Defective O-antigen polymerization in tolA and pal mutants of Escherichia coli in response to extracytoplasmic stress. J Bacteriol **187:**3359-68.
128. **Wang, X., J. F. Preston, 3rd, and T. Romeo.** 2004. The pgaABCD locus of Escherichia coli promotes the synthesis of a polysaccharide adhesin required for biofilm formation. J Bacteriol **186:**2724-34.
129. **Wang, X., M. Rochon, A. Lamprokostopoulou, H. Lunsdorf, M. Nimtz, and U. Romling.** 2006. Impact of biofilm matrix components on interaction of commensal Escherichia coli with the gastrointestinal cell line HT-29. Cell Mol Life Sci **63:**2352-63.
130. **Warren, M. J., E. L. Bolt, C. A. Roessner, A. I. Scott, J. B. Spencer, and S. C. Woodcock.** 1994. Gene dissection demonstrates that the Escherichia coli cysG gene encodes a multifunctional protein. Biochem J **302 (Pt 3):**837-44.
131. **Weber, H., C. Pesavento, A. Possling, G. Tischendorf, and R. Hengge.** 2006. Cyclic-di-GMP-mediated signalling within the sigma network of Escherichia coli. Mol Microbiol **62:**1014-34.
132. **Weiner, J. H., and L. Li.** 2008. Proteome of the Escherichia coli envelope and technological challenges in membrane proteome analysis. Biochim Biophys Acta **1778:**1698-713.
133. **Weissborn, A. C., Q. Liu, M. K. Rumley, and E. P. Kennedy.** 1994. UTP: alpha-D-glucose-1-phosphate uridylyltransferase of Escherichia coli: isolation and DNA sequence of the galU gene and purification of the enzyme. J Bacteriol **176:**2611-8.
134. **White, A. P., D. L. Gibson, S. K. Collinson, P. A. Banser, and W. W. Kay.** 2003. Extracellular polysaccharides associated with thin aggregative fimbriae of Salmonella enterica serovar enteritidis. J Bacteriol **185:**5398-407.
135. **White, A. P., D. L. Gibson, W. Kim, W. W. Kay, and M. G. Surette.** 2006. Thin aggregative fimbriae and cellulose enhance long-term survival and persistence of Salmonella. J Bacteriol **188:**3219-27.
136. **Zhang, X. S., R. Garcia-Contreras, and T. K. Wood.** 2007. YcfR (BhsA) influences Escherichia coli biofilm formation through stress response and surface hydrophobicity. J Bacteriol **189:**3051-62.
137. **Zhao, G., P. Ceci, A. Ilari, L. Giangiacomo, T. M. Laue, E. Chiancone, and N. D. Chasteen.** 2002. Iron and hydrogen peroxide detoxification properties of DNA-binding protein from starved cells. A ferritin-like DNA-binding protein of Escherichia coli. J Biol Chem **277:**27689-96.
138. **Zheng, D., C. Constantinidou, J. L. Hobman, and S. D. Minchin.** 2004. Identification of the CRP regulon using in vitro and in vivo transcriptional profiling. Nucleic Acids Res **32:**5874-93.
139. **Zogaj, X., M. Nimtz, M. Rohde, W. Bokranz, and U. Romling.** 2001. The multicellular morphotypes of Salmonella typhimurium and Escherichia coli produce cellulose as the second component of the extracellular matrix. Mol Microbiol **39:**1452-63.

Chapter III

Economical Evolution: Microbes Optimize the Synthetic Cost of Extracellular Proteins

Abstract

Protein evolution is not simply a race towards improved function. Because organisms compete for limited resources, fitness is also affected by the relative economy of an organism's proteome. Indeed, many abundant proteins contain relatively high percentages of amino acids that are metabolically less taxing for the cell to make, thus reducing cellular cost. However, not all abundant proteins are economical and many economical proteins are not particularly abundant. Here we examined protein composition and found that the relative synthetic cost of amino acids constrains the composition of microbial extracellular proteins. In *Escherichia coli* extracellular proteins contain, on average, fewer energetically-expensive amino acids independent of their abundance, length, function, or structure. Economic pressures have strategically shaped the amino acid composition of multi-component surface appendages, such as flagella, curli, Type I pili; and extracellular enzymes including Type III effector proteins and secreted serine proteases. Furthermore, *in silico* analysis of *Pseudomonas syringae, Mycobacterium tuberculosis, Saccharomyces cerevisiae,* and over 25 other microbes spanning a wide range of GC-content revealed a broad bias towards more economical amino acids in extracellular proteins. The synthesis of any protein, especially those rich in expensive aromatic amino acids, represents a significant investment. Because extracellular proteins are lost to the environment and not recycled like other cellular proteins, they present a greater burden on the cell as their amino acids cannot be re-utilized during translation. Out analysis indicates that evolution has optimized extracellular proteins to reduce their synthetic burden on the cell.

Importance

Microbes secrete proteins to perform essential interactions with their environment such as motility, pathogenesis, biofilm formation and resource acquisition. However, because microbes generally lack protein import systems, secretion is often a one way street. Consequently, secreted proteins are less likely to be recycled by the cell due to environmental loss. We demonstrate evolution has in turn selected these extracellular proteins for increased economy at the level of their amino acid composition. Compared to their cellular counterparts, extracellular proteins have fewer synthetically expensive amino acids and more inexpensive amino acids. The resulting bias lessens the loss of cellular resources due to secretion. Furthermore, this economical bias was observed regardless of the abundance, length, structure, or function of extracellular proteins. Thus, it appears economy may account for the compositional bias seen in many extracellular proteins and deliver further insight into the forces driving their evolution.

Introduction

By enveloping cellular life, membranes separate proteomes into two distinct groups: cellular and extracellular. While free living bacteria secrete extracellular proteins through several dedicated pathways, there are no known systems by which extracellular proteins are imported (121, 125, 129, 130). Consequently, extracellular proteins are less likely to be recycled by the cell or passed down during cell division. Many extracellular proteins contribute to pathogenesis and have been noted for their unique compositional biases (48, 100, 159), which are significant enough to be predictive (27, 39, 51, 52, 90, 100, 120, 132, 136, 168). However, identifying and exploiting these biases has received more attention than determining which pressures led to them (9, 109).

Evolution selects for phenotypic changes that increase organismal fitness. At the molecular level, amino acid substitutions that enhance, diversify, or maintain beneficial functions or phenotypes are favored. However, not all substitutions are predicted to

affect protein folding or function (75, 148). Nonetheless such neutral substitutions, particularly in abundant proteins, can affect the metabolic load on an organism, and thus be subject to natural selection (24, 123). Accordingly, microbes that thrive in nutrient restrictive environments have proteins with fewer sulfurs, carbons, or nitrogens (16, 44, 81, 86, 91). Similarly, transient nutrient starvation results in expression of proteins with less of a limiting element (23, 46). Furthermore, many amino acid biosynthetic enzymes contain less of the amino acid they produce (7, 113).

Protein composition is also shaped by the energy required to synthesize individual amino acids (3, 63, 81, 119, 135, 149). The total synthetic cost of an amino acid includes both the ATPs/GTPs used in biosynthesis and the energy lost to central metabolism from the consumption of precursors (3, 35, 63, 160). The synthetic costs of amino acids vary over six fold in *Escherichia coli*: Gly costs 11.7 high energy phosphate bonds (~P) or ATPs, whereas, Trp costs 74.3 (Table 3.1, 3.2) (3). Numerous studies have found that abundant proteins are often composed of amino acids that require fewer ATPs to produce (3, 63, 81, 119, 135). Here, we demonstrate that protein composition and economy is more tightly coupled to location. Compared to cytoplasmic, periplasmic, or membrane proteins, extracellular proteins contain a significantly higher amount of economic amino acids.

Results and Discussion

Protein location and cost in *Escherichia coli*

We calculated the average synthetic cost (ASC) of each protein in *E. coli* K-12 (Fig. 3.1A, Table 3.3A) using the amino acid synthetic cost of chemoheterotrophic bacteria (3, 63). Strikingly, 11 of the 100 most economical proteins (lowest ASCs) were extracellular, even though extracellular proteins comprise only 0.37% of total proteins – a 30 fold enrichment (Fig. 3.1B, Table 3.4). Extracellular proteins required 2.9 fewer ATPs per residue than an average protein (Table 3.5A: U-test, $P=1.96\times10^{-9}$). Thus for a typical protein in *E. coli*, these biases would save ~900 ATPs. ASC was nearly predictive

for location in *E. coli*: not a single extracellular protein had an ASC above the global average.

Periplasmic and outer membrane proteins were enriched three and five fold respectively amongst economical proteins; inner membrane proteins were more likely to contain expensive residues. Surprisingly, outer membrane proteins were significantly more economical than inner membrane proteins due to an increased number of expensive amino acids in integral membrane proteins (Tables 3.6, 3.7, Fig. 3.6**:** 1.8 ATPs per aa, U-test, $P=4.21\times10^{-31}$). ASCs of outer membrane β-barrel and membrane anchored proteins were similar to cytoplasmic proteins; however, outer membrane lipoproteins, many of which have soluble periplasmic domains, had significantly lower ASCs than cytoplasmic proteins (Fig. 3.6, Tables 3.6, 3.7).

Inner membrane lipoproteins also had significantly lower ASCs than cytoplasmic proteins (Fig 3.6, Table 3.6). Many inner and outer membrane lipoporteins have domains that are localized within the periplasmic space, and expectedly both types of lipoproteins have a mean ASC similar to periplasmic proteins (Fig 3.6, Table 3.7). However, the inner membrane lipoproteins are slightly cheaper than their outer membrane counterparts (Table 3.7). Besides a slightly lower hydrophobic content than outer membrane lipoproteins (Table 3.7), one possible reason for the lower cost of inner membrane lipoproteins may lie in amino aicd biases proximal to the N-terminal cysteine by which lipoproteins are attached to phospholipids (154). Inner membrane lipoproteins, have biases in positions proximal to this cysteine which defines their retention from Lol system transport to the outer membrane; less expensive amino acids are typically preferred in position 3, while several of the more expensive amino acids are not (154). Consequently, I compared the ASC along the length of the ten most economical inner membrane liporproteins to similar costing outer membrane lipoproteins. The average cost of the first 25 and 50 amino acids of inner membrane lipoproteins were 4.3 and 4.0% less, respectively than in outer membrane lipoproteins with similar ASCs. These biases in inner membrane liporprotines may reflect that additional constraints for efficient sorting by the Lol transport system.

Protein economics of extracellular appendages

One of the most abundant extracellular proteins in *E. coli* is the major subunit of the curli fiber CsgA (13) (Fig. 3.8B). Similar to some curli-specific gene products, CsgA is rich in Gly and Ser. CsgA is composed of 19.2% Gly (global mean 7.2%) and 28.5% Gly+Ser (Fig. 3.1C, 3.7 and Table 3.8: global mean 12.8%), the 6th highest amount of any *E. coli* protein. Intriguingly, the curli regulator CsgD increases expression of a gene for the biosynthetic enzyme GlyA which interconverts Gly and Ser (30). CsgD may increase GlyA to balance Gly and Ser pools resulting in efficient curli production. More appropriately, as ancient (66), relatively simple amino acids, Gly and Ser are two of the least expensive to produce (3, 35, 160) (Table 3.1). Consequently, CsgA has the 9th lowest ASC in *E. coli*, utilizing 4.17 fewer ATPs per residue than average (Table 3.9). The major subunits of flagella and Type 1 pili have the 6th and 11th lowest ASCs, respectively (See Fig. 3.8; Table 3.9).

The relative economy of extracellular proteins is not due to enrichment of the same amino acids. Collectively, extracellular proteins in *E. coli* contain more of the inexpensive amino acids Ala, Asn, Gln, Ser, and Thr and fewer of the expensive residues Arg, His, Met, Phe, and Trp (Table 3.7). When examining amino acid usage in CsgA, FimA, and FliC, we found all three contain fewer aromatic residues. However, each major fiber subunit had a unique combination of inexpensive amino acids (Table 3.10). Enrichment of Gly and Asn is responsible for 65.8% of CsgA's energy savings (Table 3.8 and Fig 3.9); whereas, 54.9% of FimA's savings is due to enrichment of Ala and Thr. In contrast, FliC is not particularly rich in any one amino acid: 31.7% of its savings is due to enrichment of Asn and Thr. Instead, FliC contains reduced amounts of aromatic amino acids relative even to CsgA or FimA. Thus extracellular proteins are not simply rich in a specific subset of economical amino acids. Rather they contain many combinations of inexpensive amino acids and typically lack expensive ring-structured amino acids.

The bacterial flagellum is one of most complex and well-studied cellular structures in bacteria (87). With multiple proteins in every cellular compartment, the flagellum is an excellent system to analyze the connection between protein location and

ASC within a single organelle. Indeed, the cost of flagellar proteins decreases the farther they are from the cytoplasm (darker reds in Fig. 3.2). Extracellular flagellar proteins have significantly lower ASCs than cytoplasmic flagellar proteins (Table 3.11, Fig. 3.10A: 2.6 ATPs per aa, T-test, $P=2.62\times10^{-6}$). Furthermore, Curli and Type I pili proteins show economic trends similar to flagellar proteins (Table 3.9, Fig. 3.11). Because ASC might be influenced by protein length, abundance, or function (3, 135), we tried to correlate the ASC of flagellar proteins with these criteria. However, we did not find a significant trend when comparing the ASCs of flagellar proteins with their length or abundance (87) (Table 3.12, Fig. 3.10: Spearman r_s=-0.189,-0.255; P=0.232, 0.209). Additionally, the function of extracellular flagellar proteins includes structural, assembly, and regulatory roles; thus function does not appear to affect their relative economy. The cytoplasmic regulatory proteins of flagella are relatively expensive (Table 3.9). Intriguingly, FlgM which is a secreted anti-sigma factor (68) is quite economical. Among regulatory proteins in K-12 (Uniprot GO:65007), FlgM is by far the most economical, requiring 3.91 fewer ATPs per residue than an average regulatory protein.

In a more encompassing analysis, we reexamined the correlation between ASC and length, abundance, or function in the *E. coli* proteome. Although cytoplasmic and periplasmic proteins had significant negative correlations between abundance and cost, outer membrane β-barrel and integral membrane proteins did not (Table 3.13). Additionally, while protein length and cost were weakly correlated overall (r_s=-0.05, P=0.0009), there was no significant correlation in outer membrane, periplasmic, or extracellular proteins (Table 3.14). Finally, many extracellular proteins are fibrous in *E. coli*; therefore, we examined the ASC of several different fibrous protein polymers. As expected, extracellular protein polymers contained fewer expensive amino acids than their intracellular counterparts (Table 3.15: 1.7 ATPs per aa, T-test, $P=4.3\times10^{-4}$). Collectively, these results suggest location has a more significant role on the amino acid composition of proteins than previously appreciated.

How significant are the energy savings garnered by FliC and FimA? Previous studies have suggested a single neutral substitution can affect cell growth and thus be

subject to negative selection (24, 81). Indeed, one Gly to Trp substitution in FliC would increase the total cellular ATP requirement of *E. coli* by 0.031%. We compared the ASC of FliC and FimA to an average cellular protein to calculate how much energy *E. coli* saves by making these proteins with less expensive amino acids (See Table 3.16). The biases in FliC save the cell $4.4x10^7$ ATPs per flagellum. If converted to H^+ these savings correspond to the energy required to run the flagellum at 100Hz for 24 minutes. In a typical *E. coli* cell, FliC or FimA savings ($2.2x10^8$ ATPs for 5 flagella or 300 fimbriae) represents a 1.10% reduction in overall cellular cost. Accordingly, flagellar mutants rapidly overtake WT strains due to lower metabolic loads (106), constitutive flagellar mutants (*flgM* and *fliD)* grow slower due to excess FliC production (78), and *flgG* mutants but not *motAB* mutants outcompete WT bacteria on plates (43).

Alternative costs

The association between amino acid costs and protein abundance has been explored using different parameters including amino acid mass and atomic composition (16, 23, 24, 35, 63, 81, 135, 160). A composite of atomic content, mass has been proposed as a complementary approach to calculate relative costs (135). The synthetic cost and mass of amino acids are highly correlated (Fig. 3.3A, Table 3.2: R=0.803, $P=2.02\times10^{-5}$). Predictably, we found extracellular proteins have smaller amino acids than cytoplasmic proteins (Fig. 3.3B, Table 3.17A: 6.7 Daltons per aa, U-test, $P=3.4x10^{-9}$). Subsequently, we looked at their carbon and nitrogen content. Diversion of carbon precursors from central metabolism is the largest component of the synthetic cost of amino acids (3, 35, 160). While the carbon content of amino acids significantly correlates with their synthetic costs, nitrogen content does not (Table 3.2). However, extracellular proteins have significantly lower contents of both carbon and nitrogen compared to cytoplasmic proteins (Fig. 3.12, Table 3.17A: 8.5%, 6.4% less per aa, U-test, $P=3.0\times10^{-9}$, 5.0×10^{-8}). Amino acids in extracellular proteins also have on average lower sulfur content and Gibbs free energy (8) (Fig. 3.12, Table 3.17A: 46.4%, 8.3% less per aa, U-test, $P=1.5\times10^{-6}$, 4.1×10^{-9}). Thus no matter how costs were calculated, the amino acids

in extracellular proteins are, on average, more economical than those in cellular proteins.

Surpisingly, inner membrane proteins have less nitrogen content and mass than cytoplasmic proteins. These results conflict with the higher ASC and carbon content of inner membrane proteins and may reflect functional constraints. Inner membrane proteins and integral membrane proteins, in particular, have less charged residues including Arg, Lys, Asp, and Glu likely due to the hydrophobic nature of phospholipids (Table 3.7). Arg and Lys are nitrogen rich amino acids. The reduced content of Arg and Lys along with slight reductions in Asn and Gln residues results in a lower mean nitrogen content of inner membrane proteins. The mass difference in inner membrane proteins is likely due to the increased content of Asn, Asp, Glu, Gln, and Arg and slightly increased contents of Ala and Gly in inner membrane proteins. In Fig 3.3A these amino acids are further away from the linear fit of amino acid mass to synthetic costs. Consequently, these amino acids are respectively more and less massive relative to their synthetic cost in comparison to other amino acids. The composition of these residues in inner membrane proteins results in their contradictory lower mean amino acid mass.

Protein function and structure

The extracellular environment represents a unique folding environment which is relatively devoid of cellular chaperones and thus may affect amino acid preferences in extracellular proteins. To explore this possibility we looked at ASCs in Type III secretion effectors which many pathogenic bacteria secrete directly into eukaryotic cells (34). Although Type III effectors are extracellular proteins, they function within the host in an environment that is similar to the bacterial cytoplasm. Despite their potential functional constraints, the Type III effectors of *Pseudomonas syringae* (called HOPs) had significantly lower ASCs than cytoplasmic proteins (Fig. 3.4A, Tables 3.18A, 3.19, 3.20: 1.48 ATPs per aa, U-test, $P=6.22\times10^{-15}$). Type III effectors of other animal and plant pathogens also had significantly lower ASCs (Fig. 3.13, Tables 3.18B, 3.19: 1.27 ATPs per aa, U-test, $P=4.4\times10^{-19}$).

To further investigate if localized protein economy is independent of function, we inspected bacterial serine proteases. On average, extracellular serine proteases cost 0.72 ATPs less per residue than cellular serine proteases (Table 3.21: U-test, $P=2.7\times10^{-9}$). In *Escherichia* and *Bacillus* species, savings were 1.15 and 1.18 ATPs per residue, respectively (T-test, $P=1.6\times10^{-8}$, 4.6×10^{-5}).

Additionally, we examined Type V secretion proteins, or autotransporters, to explore the effects of localized protein economy at the domain level. Autotransporters are bacterial proteins that contain a β-domain which integrates into the outer membrane and assists the translocation of an α-domain to the cell surface where it is often cleaved autolytically or by other proteases (65). While β-domains have high similarity, α-domains are quite diverse and function as enzymes (protease, esterase, lipase) or assist cell aggregation, actin polymerization, or other virulent activities (65). I compared the cost of cleaved α-domains to their respective β-domains in autotransproters with defined α-domains in Uniprot from Henderson et al 2004 (65). While most β-domains had average ASCs, all the α-domains except IgA1 protease had lower ASCs (See Ag43 in Fig. 1.7). On average, the α-domains examined had an 8.7% lower ASC than their respective β-domain (Table 3.18C). Autotransporters thus demonstrate that extracellular protein economy is also refected on the domain level with secreted and cleaved domains having lower cost than retained transport domains.

Oxidation and proteolysis are more likely in the harsh environment where extracellular proteins function. Protein oxidation levels are influenced by both the structure and amino acid composition of proteins as well as the particular oxidant to which they are exposed (H_2O_2, HOCl, NO) (25, 62, 142, 166). Commonly oxidized residues include Met, Cys, and the aromatic amino acids. With the exception of Tyr, there are fewer of these amino acids in extracellular proteins; however, the majority of the cost savings in extracellular proteins is not due to these biases (Tables 3.7, 3.10, 3.20). Further exploration of how protein inactivation may have shaped extracellular proteins will require more knowledge of which amino acids are surface exposed or found within sensitive structures such as loops and α-helices (151). Besides oxidation,

extracellular proteins are also exposed to extracellular peptidases. Within the GI tract, the extracellular proteins of *E. coli* encounter trypsin, chymotrypsin, and elastase which cleave after basic, aromatic, and simple amino acids (31). However, the effects these peptidases would have on ASC are contradictory; avoidance of cleavage would theoretically select for more or less expensive amino aicds depending on the protease. Therefore, selection to prevent proteolysis is unlikely to be the primary cause of compositional economy in extracellular proteins.

The unique folding environment of the extracellular space may also constrain protein structure. Consequently, we examined the predicted secondary structure and intrinsic disorder of non-membrane proteins in *E. coli* and *P. syringae*. Extracellular proteins in *E. coli* had significantly less helical structure and significantly more strand content than cytoplasmic proteins (See Table 3.22A). However, there was no correlation between the ASC of *E. coli* proteins and their helix, strand, or coil content (Table 3.22). Furthermore, strands are by far the most expensive secondary structure (Table 3.22). Such analysis led us to examine which structures had the most savings relative to cytoplasmic proteins. While all three secondary structures were less expensive in extracellular proteins, coiled regions had the most economical substitutions (Table 3.22). Finally, extracellular proteins had greater amounts of disordered regions compared to cytoplasmic proteins; however these differences are not significant in *E. coli* (Table 3.23: T-test, P=0.324). Disordered regions also are a small percentage of the overall structure of extracellular proteins and thus do not significantly alter the relative economy of these proteins (Table 3.23). We found similar results in *P. syringae*, except for Type III effectors which had higher contents of disordered regions and lower strand contents than other extracellular proteins (Tables 3.22B, 3.23). Disordered regions in Type III effectors may assist their function within eukaryotes where such structural disorder is more common and is often associated with protein-protein interactions (41, 138). Collectively, these results suggest structural differences are not responsible for the economy of many extracellular proteins and that more economical substitutions occur more frequently in less structured regions.

Ubiquity of extracellular protein economy

A wide variety of other extracellular proteins also contain, on average, fewer expensive amino acids. For example the elastases and exotoxins of *Pseudomonas aeruginosa*, the S-layer and holdfast proteins of *Caulobacter cresentus*, and the major capsule protein Caf1p of *Yersinia pestis* are among the most economical proteins in those organisms (Table 3.18D). More distantly related organisms, such as the Gram-positive pathogen *Mycobacterium tuberculosis* and the budding yeast *Saccharomyces cerevisiae,* showed similar patterns. Both cell wall and extracellular proteins of *M. tuberculosis* have significantly lower ASCs, partially due to the many cell surface antigens in the PE_GRS protein family (Fig. 3.4B, Tables 3.3C, 3.17C). Likewise, the cell wall and extracellular proteins of yeast have significantly lower ASCs under both respiratory (Fig. 3.4C, 3.14 and Table 3.3B) and fermentative growth conditions (Fig. 3.4D, Table 3.17B).

To see just how broadly protein location affected ASC, we initially examined the ASCs of the extracellular proteins in all 717 Gram-negative organisms in PSORTDB. To our knowledge, this is the most extensive examination of protein synthetic cost in bacteria. Overall, the amino acids in extracellular proteins cost 1.3 ATPs less per residue than cytoplasmic proteins (U-test, $Z=64.1$, $P<<1\times10^{-325}$). However, typical savings may be much greater. Combining these proteomes over represents certain species such as *E. coli*, includes obligate anaerobes and amino acid auxotrophs, excludes known extracellular proteins (See **Methods**), and overlooks the effects of GC-content on carbon content and amino acid composition (15, 32, 139).

Currently, the relative cost of proteins in organisms lacking one or more amino acid synthesis pathways is difficult to assess. Abundant proteins in two different *Chlamydia* species have been shown to contain either more or less of the amino acids for which they are auxotrophic, perhaps due to metabolic or nutritional differences (63). When comparing the cost of extracellular proteins in *Gammaproteobacteria*, we found several insect endosymbionts had relatively expensive extracellular flagellar proteins (Table 3.24). This increased cost comes from amino acid biases that only partially reflect

their metabolism or nutrition (127, 165, 169). For example, *Buchnera aphidicola* has more His, Ile, and Lys and less Gly, Thr, and Val in its extracellular flagellar proteins despite its capacity to synthesize these amino acids. Other factors including transport efficiency, host metabolic interdependency, or GC-skew may be more relevant (1, 32). More knowledge of how these factors affect amino acid composition is needed to properly study protein economy in auxotrophic organisms.

Protein composition is also affected by an organism's GC-content (139). GC-rich codons tend to code for less expensive amino acids (135) (Table 3.2). Accordingly, proteins produced by GC-rich organisms are, on average, less expensive to synthesize than proteins produced by organisms with lower GC-contents (Fig. 3.5A). For example, the mean cost of *M. tuberculosis* proteins is less than in *E. coli* due to high GC-content affecting amino acid preferences (Compare Mtu and Eco in Fig. 3.5A) (33). To overcome these limitations we looked, individually, at a diverse collection of 25 Gram-negative aerobes representing a wide range of genomic GC-content. Each has retained the ability to synthesize the standard 20 amino acids (See Table 3.25). In all of these organisms, extracellular proteins had significantly lower ASCs and mean amino acid mass when compared to other cellular proteins (Fig. 3.5B, 3.15 and Table 3.26). Given the trends in Fig. 3.5B, a typical Gram-negative organism with 50% GC-content would save 2.05 ATPs per amino acid in its extracellular proteins, an 8.7% reduction in synthetic cost (Table 3.16). Assuming $5x10^5$ copies per cell, amino acid biases in extracellular proteins would reduce total cellular cost by 1.54%. Theoretically, in a direct competition, strains without these savings would be outnumbered nearly 15 fold within 250 generations.

Conclusion

Previous studies have explored the connection between amino acid cost and a variety of attributes (3, 23, 35, 36, 63, 81, 119, 135, 149, 160, 161). However, as evidenced in the flagella system of *E. coli*, cellular location can have a stronger influence on average cost of amino acids. We found the synthetic costs of extracellular proteins are significantly reduced in *E. coli, P. syringae, M. tuberculosis, S. cerevisiae*, and many

other organisms. Furthermore, this economic bias is observed despite the abundance, length, function, or structure of extracellular proteins. Understanding these compositional biases in extracellular proteins may improve current prediction methods. In Fig 3.5B, 92.3% of extracellular proteins have an ASC below the organism's mean ASC. Additionally, comprehending the economic selection of amino acids in extracellular proteins may elucidate new pressures upon and constraints of their evolution, particularly in horizontally-acquired genomic islands where disparate codon usage and GC-content gradually adapt to the host (57, 79, 80)

Microbes interact with their environment directly through external structures, leading to possible loss of surface proteins. Besides secretion, extracellular proteins are lost during fiber-shedding, outer membrane blebbing, and cell wall damage. This egress of extracellular protein is likely irreparable (121, 125, 129, 130); consequently, they are less likely to be recycled by the cell's chaperone and protease systems. Such loss increases the relative cost of extracellular proteins to the cell. Accordingly, excessive production of extracellular proteins results in decreased growth rate and competitive fitness (43, 78, 106, 114). Therefore, we propose that there is a strong selection for less expensive amino acids in extracellular proteins to counteract this loss of cellular resources.

Methods

Calculating Average Synthetic Cost (ASC) and other cost values.

To calculate protein cost including ASC (ATPs/aa), mean amino acid mass, mean Gibbs free energy (ΔG_r), and average atoms (Carbon, Nitrogen, Sulfur) per amino acid we used the following equation

$$\text{Protein Cost} = (\sum C_i \times F_i) / L$$

where C_i is the appropriate cost of the amino acid *i*, F_i is the frequency of *i*th amino acid, and L is total protein length.

The different cost values for amino acids used include the amino acid synthetic cost in ATPs, the amino acid mass in Daltons, the ΔG_r for an amino acid, or the number of carbons, nitrogens, or sulfurs in a given amino acid. For ASC, different synthetic costs were used depending on the organisms. For *E. coli* K-12, *P. syringae pv. tomato* str DC3000*, M. tuberculosis* H37Rv*,* and other bacteria the synthetic cost values for amino acids in chemoheterotrophic bacteria were used (3); for *S. cerevisiae* the respiratory and fermentative synthetic costs of amino acids in yeast were used (160).

For more information about cost values see Table 3.1, 3.2, and associated notes. Similar economic trends were seen using costs values other than synthetic cost including atomic composition, Gibbs free energy (8), and mean mass of amino acids (135) (See Fig. 3.3, 3.4, 3.12, Table 3.17). Statistical comparisons of cost values between locations were primarily Mann Whitney U-test; data for many groups failed normality test (detailed statistics in Tables 3.3A-C). P values were determined for large Z and t values using R (www.r-project.org).

Protein location

Protein sequence and locations for *E. coli* K-12 are from Echobase Echolocation (http://www.york.ac.uk/res/thomas/index.cfm); YdbA (EB1284) lacked sequence data. The location of FlgM (68) and FliK (95) were reassigned to extracellular and FlgJ to periplasmic (87). *S. cerevisiae* protein location and sequence were downloaded from the Comprehensive Yeast Genome Database (http://mips.helmholtz-

muenchen.de/genre/proj/yeast/). Protein sequence and locations for *M. tuberculosis, P. syringae,* and other bacteria were from PSORTDB (http://db.psort.org/). For *M. tuberculosis* Fmt, TrmD, Hns, HupB, and ribosomal proteins (33) were changed to cytoplasmic. All bacterial proteins with the GO term "secreted" (keyword 0964 in UniProt) were reassigned to extracellular, except the cell wall proteins of *M. tuberculosis*. Other changes are listed in below; modified locations in *E. coli* and *M. tuberculosis* are bold in Tables 3.3A,C.

Changes to Protein Location Annotations

Bacteria sequences and locations (exceptions *E. coli* – Echobase; *S. glossinidius, B. aphidicola* Sg, and *H. cicadicola* Dsem - NCBI) were from PSORTdb. The location of several proteins in PSORTdb were either annotated as 'Unknown' or given the wrong location altogether. To improve the location data, all ribosomal proteins were changed to cytoplasmic. We also searched the literature for known extracellular proteins and reannotated their locations. Those changes are as follows: CsgA,B and FimH for *S. typhimurium* LT2; MshA,B,O for *C. psychrerythraea* 34H (155); MshA,B,O and CsgA,B for *S. oneidensis* MR-1; CiaB for *W. succinogenes* DSM 1740 (12); AcuG, ComC, and ComP for *Acinetobacter sp.* ADP1 (BD413) (54, 82, 116); AprX and PopB for *P. aeruginosa* PAO1 (42, 61); TcpA, HylA (6), MshA,B (155), Hcp, and VgrG (22) for *V. cholera* El Tor; and HfaA for *C. cresentus* CB15 (152) were all changed to extracellular. In Gram-negative bacteria, PilA, FlgD,E,K,L (87), FlgM (68), FliK (95), and flagellins were changed to extracellular. For *F. johnsoniae* UW101, SprB (72) and homologs of the curli subunits CsgA,B were changed to extracellular. For *C. hutchinsonii* ATCC 33406, SprB was identified by blasting with FjoH_0979 of *F. johnsoniae* UW101. CHU_2225 (Uniprot Q11SX6_CYTH3) was identified as the most likely SprB homolog (E value of 1e-104) and was changed to extracellular.

Graphing of Protein Average Synthetic Cost (ASC)

To vertically spread the cost data within a location, protein length was used. The value $0.2\times LOG_{10}$ (Length) was added to a specific integer of each location. For example cytoplasmic proteins were $1+0.2\times LOG_{10}$ (Length), and inner membranes were $2+0.2\times LOG_{10}$ (Length). These numbers were graphed vs. proteins costs to generate the graphs in Fig. 3.1A, 3.3B, 3.4, 3.6, 3.8, 3.12, 3.13.

Electron Microscopy

For flagella, Type 1 pili, and curli TEMs, BW25113 was grown to late log in Lysogeny broth (LB: 10g Tryptone, 5g Yeast extract, 10g NaCl per liter) shaking at 250 rpm, stationary in LB at 37°C for 16 h, or on YESCA plates (10g Casamino acids, 1g Yeast extract, 20g Agar per liter) for 48h at 26°C respectively. All samples were resuspended in PBS, incubated with formvar-coated copper grids (Ernest F Fullam, Inc), briefly washed with H_20, and stained 2 minutes in 2% uranyl acetate. Grids were viewed using a Phillips CM10 microscope.

Calculating Percent Savings for a Specific Amino Acid Bias in a Given Protein

Total ATPs saved in a protein:

Protein chain length × (Organism's mean ASC – Protein ASC)

ATPs saved per amino acid:

(Number of a specific amino acid in a protein – (Organism's mean % of specific amino acid × Protein chain length)) × (Average cost of all amino acids – Cost of a specific amino acid)

Percent Savings of a specific amino acid:

(ATPs saved per amino acid / Total ATPs saved) × 100

Type III Effectors

Type III effectors were culled from a recent survey of plant and animal pathogens and are listed in Table 3.18B (10). For *P. syringae pv. tomato* str. DC3000

Type III effectors (also called Hops – Hrp outer protein) we used the Hop database, the T3SS helper database, and the Discontinued Hops database at the "*P. syringae* Hop Identification and Nomenclature Home Page" (http://pseudomonas-syringae.org/pst_func_gen2.htm). We included the following proteins: PSPTO_4101 (*hopAK1),* PSPTO_1022 (*hopAM1-1),* and PSPTO_5061 (*hopAN1)* which have plant phenotypes (56, 71, 131); PSPTO_1378 (*hrpH*) which was shown to be a T3SS substrate (111), and PSPTO_3292,3293,0907 which are translocated into plant cells (131). Confirmed Hop proteins were considered to be extracellular in later analysis. Only one copy of Hcp (*P. aeruginosa* PAO1*)* and HopAM1 (*P. syringae tomato*) were used in data analysis. See Table 3.18A for a complete table of *P. syringae* Hops and Hop helpers. Fig. 3.4A, 3.13, and Table 3.19 compare the ASC of Type III effectors to other cellular proteins.

Serine Proteases

Extracellular serine proteases were those annotated as "secreted" in Uniprot (http://www.uniprot.org). Cellular serine proteases were serine proteases not annotated as "secreted." 5,784 bacterial serine proteases were downloaded from Uniprot using the keywords "taxonomy:2 keyword:720" on August 18, 2009. Bacterial extracellular serine proteases (191 total) were downloaded using "taxonomy:2 keyword:720 keyword:964". Taxonomy 2 = "Bacteria"; keyword 720 = "Serine protease"; and keyword 964 = "Secreted".

Autotransporters

The autotransporters used were those described in Table 1 of Henderson et al 2004 (65). Uniprot (http://www.uniprot.org) was used to select autotransporters which had defined secreted α-domains. The ASC of both the α-domain and their respective β-domains were calculated using the cost values for chemoheterotrophic bacteria (3, 63).

Protein Abundance Values

Abundance values from Masuda et al (89) and Ishihama et al (70). Protein sequences and location were from Echobase; abundance values were merged with Echobase data using ECK and Uniprot numbers respectively.

Prediction of Secondary Structure and Intrinsic Disorder

The secondary structure of cytoplasmic, periplasmic, and extracellular proteins from *E. coli* K-12 and *P. syringae* were predicted using PSIPRED version 2.61 (92). Intrinsically disordered regions were predicted with DISOPRED2 version 2.4 (163) using a 2% false positive rate. Both programs were downloaded from (http://bioinfadmin.cs.ucl.ac.uk/downloads/).

Comparison of Amino Acid Composition of Flagellar Subunits

Extracellular assembly or structural components of *E. coli* flagella were individually blasted against 216 *Gammaproteobacteria*. Many flagella genes have similar regions, so sequences with E values below 10^{-10} were removed except for FlgG ($<10^{-20}$) and FliK ($<10^{-5}$) which have high identity to FlgG (83) and low ortholog identity respectively (164). Sequences whose lengths were 50% (+/-) their respective *E. coli* counterpart were removed. The remaining homologs were used in our analysis.

Figure Legends

Fig. 3.1. Protein location in *E. coli* is indicative of synthetic cost. (A) Each protein in *E. coli* is plotted based upon its average synthetic cost (ASC) and cellular location. Dotted line = mean ASC of all proteins in *E. coli*. Black bars = mean ASC of proteins in that location. U-tests were used to compare the protein ASCs of each location against cytoplasmic proteins. (B) Location of all proteins, the 100 most economical, and the 100 least economical proteins in *E. coli* as ranked by ASC. (C) Histogram of the %Gly+Ser for all proteins in *E. coli*. CsgA, the major component of of extracellular curli fibers, is indicated (green arrow); 28.5% of the amino acids in CsgA or either glycine or serine.

Fig. 3.2. Protein abundance and cost of flagellar proteins. Flagellum diagram showing colored economic percentiles of each protein. Proteins with lower ASCs (more economical) have higher percentiles and are dark red. ASC increases in order: dark red to pink, light to dark blue, gray. * Contain Sec secretion sequences. Number of proteins per flagella is given if known (87).

Fig. 3.3. Molecular mass of amino acids correlates with their synthetic cost; extracellular proteins have simpler amino acids. (A) The synthetic cost of amino acids positively correlates with their mass. Letter codes are used to indicate individual amino acids. (B) Mass has been used as an alternative cost for amino acids (135). Using mass, extracellular proteins of *E. coli* are significantly more economical; their amino acids are simpler and have less mass than average. Dotted line = mean average mass of amino acids of all proteins in *E. coli*. Black bars = mean average mass of amino acids of proteins in that location. U-tests were used to compare mass based costs.

Fig. 3.4. Relationship between protein location and cost extends to diverse organisms. (A) Cost and location of extracellular proteins and Type III Effectors – called HOPs for Hrp outer proteins (green) and HOP helpers (blue) from *P. syringae*. Smaller blue and green bars = mean ASC of HOPs and HOP helpers, respectively. (B) Cost and location of

M. tuberculosis proteins. PE_GRS (red), PPE (orange), and PE (green) family proteins are indicated. (C,D) Cost and location of *S. cerevisiae* proteins under respiratory and fermentative growth. Dotted lines = each organism's mean ASC. Black bars = the mean ASC of proteins in that location. U-tests were used to compare the protein ASCs of each location against an organism's cytoplasmic proteins.

Fig. 3.5. Per residue, proteins in GC-rich organism cost less to synthesize; however, extracellular proteins are still economical. (A) Chromosomal GC-content and mean ASC of chromosomally encoded proteins in 70 Gram-negative (red) and 30 Gram-positive (blue) bacteria. White + = thermophile, white center = psychophile, black center = host-associated organisms, black outline = chromosomal DNA < 1.3 MB of *P. ubique* the smallest genome of free living organisms (53). Slope=15.1, R=0.930. (B) Comparison of the mean ASC of extracellular proteins to total proteins in 25 Gram-negative bacteria. Each is capable of aerobic growth and synthesis of all 20 amino acids (See Table 3.25). Slope=-11.0,-17.0; R=0.696, 0.960 respectively. Eco-*Escherichia coli,* Cru-*Carsonella ruddii,* Smu-*Sulcia muelleri,* Wgl-*Wigglesworthia glossinidia,* Bap-*Buchnera aphidicola,* Abu-*Arcobacter butzleri*, Plu-*Photorhabdus luminescens,* Aae-*Aquifex aeolicus,* Th-*Thermotoga maritime/petrophila,* Hci-*Hodgkinia cicadicola,* Mtu-*M. tuberculosis,* Dra-*Deinococcus radiodurans,* Sco-*Streptomyces coelicolor,* Rsp-*Rhodobacter sphaeroides,* Pae-*Pseudomonas aeruginosa,* Psy-*Pseudomonas syringae pv. tomato,* Fjo-*Flavobacterium johnsoniae*.

Fig. 3.6. Cost and sublocations of *E. coli* proteins in inner and outer membranes. Each protein in *E. coli* is plotted based upon its average synthetic cost (ASC) and cellular location. Dotted line = overall mean average synthetic cost. Black bars = mean ASC of location. Mann Whitney U-tests were used to compare ASCs of proteins with integral membrane proteins. *E. coli* protein sublocations are also available at Echobase's Echolcation.

Fig. 3.7. Glycine and Serine composition of *E. coli* proteins with an emphasis on curli specific gene proteins. Histograms of percent of amino acids including (A) Glycine (B) Serine (C) Glycine plus Serine. Green arrows indicate curli proteins above the 95th percentile of respective percent amino acid composition. (D) Schematic of Glycine metabolism including GlyA and the glycine cleavage system.

Fig. 3.8. Protein abundance and cost of extracellular fibers. (A-C) Negative stained TEM micrographs of BW25113 cells showing flagella, curli, and Type 1 fimbriae, respectively. Bars equal to 500 nm. (D) Each protein in *E. coli* is plotted based upon its average synthetic cost (ASC) and cellular location. Dotted line = mean ASC of all proteins in *E. coli*. Black bars = mean ASC of proteins in that location.

Fig. 3.9. Amino acid compositions of *E. coli* K-12 proteins with an emphasis on curli specific gene proteins. Histograms of percent of amino acids including (A) Asparagine (B) Glutamine (C) Asparagine and Glutamine (D) A combination of Gly, Ser, Asn, and Gln. Green arrows indicate curli proteins above the 95th percentile of respective percent amino acid composition.

Fig. 3.10. The average synthetic cost (ASC) of flagellar proteins correlates with location but not abundance. (A) Comparison of flagellar protein ASC and location. Dotted line = overall mean ASC of flagellar proteins. Black bars = mean ASC of flagellar proteins in that location. P values from Student T-test (See Table 3.11). (B,C) Comparisons of ASC and abundance using known stoichiometry (87) (B) Log scale of abundance used to show FliC (20,000 per flagellum). (C) Linear scale of abundance; does not include FliC. Slope of linear fit is -2.82, R= 0.135.

Fig. 3.11. Protein economy of curli and Type 1 pili proteins. Colored diagram of the economic percentiles of each protein. Proteins with lower ASCs (more economical) have higher percentiles and are dark red. ASC increases in order: dark red to pink, light to dark blue, white. Proteins without a known function are marked with a "?".

Fig. 3.12. Extracellular proteins contain less carbon, nitrogen, oΔG_r per amino acid than other locations. (A) Correlation of synthetic cost and number of carbons per amino acid. (B) Extracellular proteins in *E. coli* have less carbons per amino acid relative to other locations. (C) The synthetic cost and number of nitrogens do not correlate. (D) Extracellular proteins in *E. coli* have less nitrogens per amino acid relative to other locations. (E) Correlation of synthetic cost and Gibbs free energy (ΔG_r) of amino acids in seawater at 18°C. (F) Extracellular proteins in *E. coli* have amino acids with lowerΔG_r relative to other locations. Dotted lines = overall mean number of carbons, nitrogens, or ΔG_r per amino acid in *E. coli* proteins. Black bars = mean value of that location. Mann Whitney U-tests were used to compare values between locations.

Fig. 3.13. Type III effectors from multiple pathogens are economical. Average synthetic cost (ASC) of Type III effectors from animal (blue) and plant (green) pathogens from (10) versus cost and location of *E. coli* proteins. Species are listed in Table 3.18B. Dotted line = overall mean ASC of *E. coli*. Black bars = mean ASC of protein in that location. Smaller blue and green bars = mean ASC of animal and plant Type III effectors respectively. Mann Whitney U-tests were used to compare the Type III effector ASC values with those of cytoplasmic proteins of *E. coli*.

Fig. 3.14. Cost and location of S. cerevisiae proteins. Location and sequences from CYGD (55). Dotted line represent overall mean average synthetic cost (ASC). Black bars represent mean ASC of location. ASC calculated using yeast respiratory values.

Fig. 3.15. **Although proteins in GC-rich organisms have reduced mass per residue, extracellular proteins have simpler amino acids.** More massive amino acids are generally more expensive to produce (See Tables 3.1, 3.2), and thus mass has been proposed as an complementary approach to calculating their cost (135) . (A) Organism GC-content and the average of mean amino acid mass of chromosomally encoded proteins in 70 Gram-negative (red) and 30 Gram-positive organisms (blue). White + = thermophile, white center = psychophile, black center = host-associated organisms, black outline = chromosomal DNA < 1.3 MB of *P. ubique* the smallest genome of free living organisms (53). Slope = -6.54, R = 0.938 (B) Mean average synthetic cost (ASC) vs average of mean amino acid mass of chromosomally encoded proteins in 100 bacteria . Slope = -2.23,R = 0.978. (C) Comparison of the average mean amino acid mass of extracellular proteins and total proteins of 25 Gram-negative bacteria. Each is capable of aerobic growth and synthesis of all 20 amino acids (See Table 3.25). Slope = -5.49, -6.56; R = 0.813, 0.956 respectively. Abu-*Arcobacter butzleri*, Fjo-*Flavobacterium johnsoniae*, Plu-*Photorhabdus luminescens,* Eco-*Escherichia coli,* Psy-*Pseudomonas syringae pv. tomato,* Pae-*Pseudomonas aeruginosa,* Rsp-*Rhodobacter sphaeroides.*

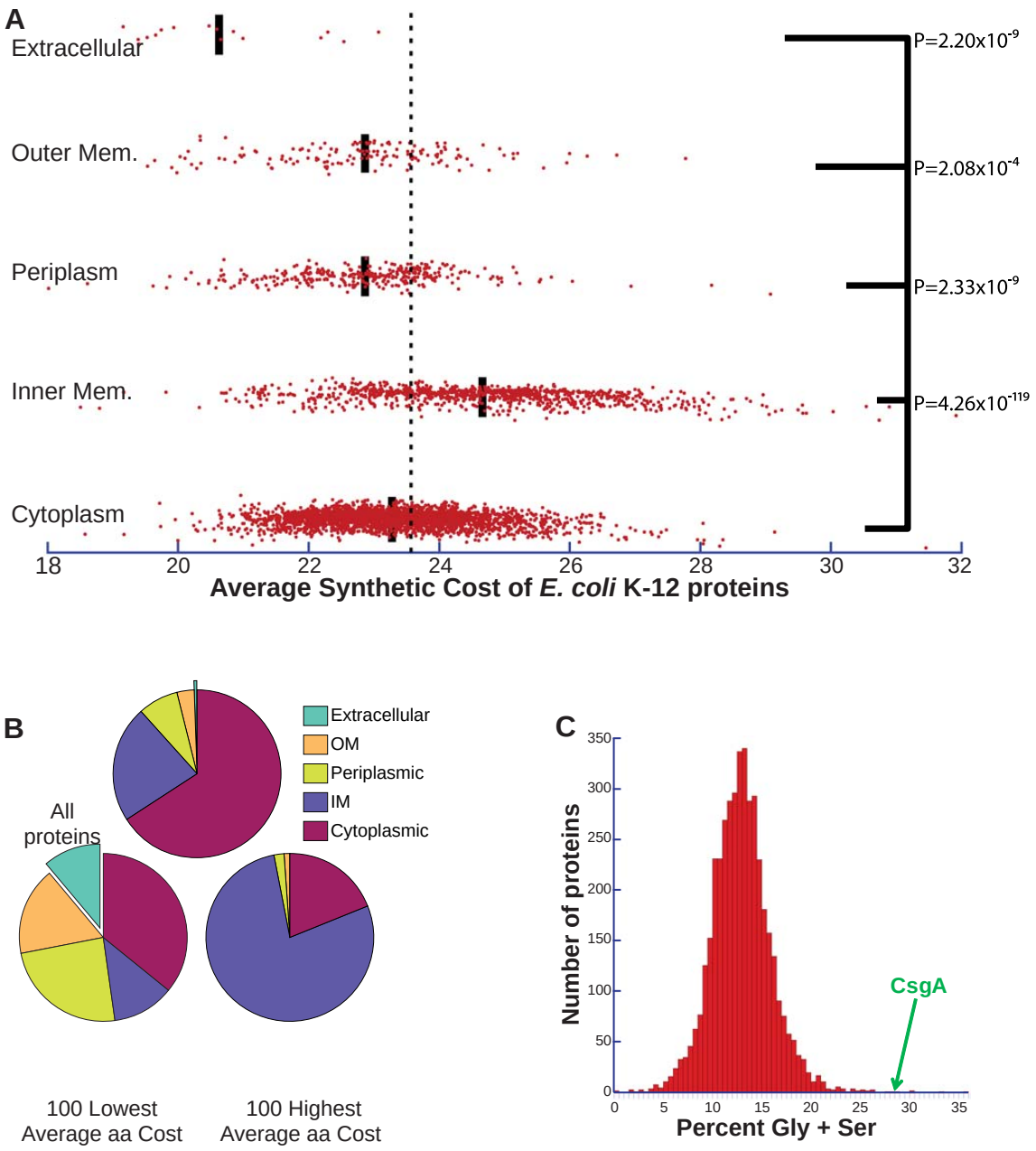


Figure 3.1. Protein location in *E. coli* is indicative of synthetic cost.

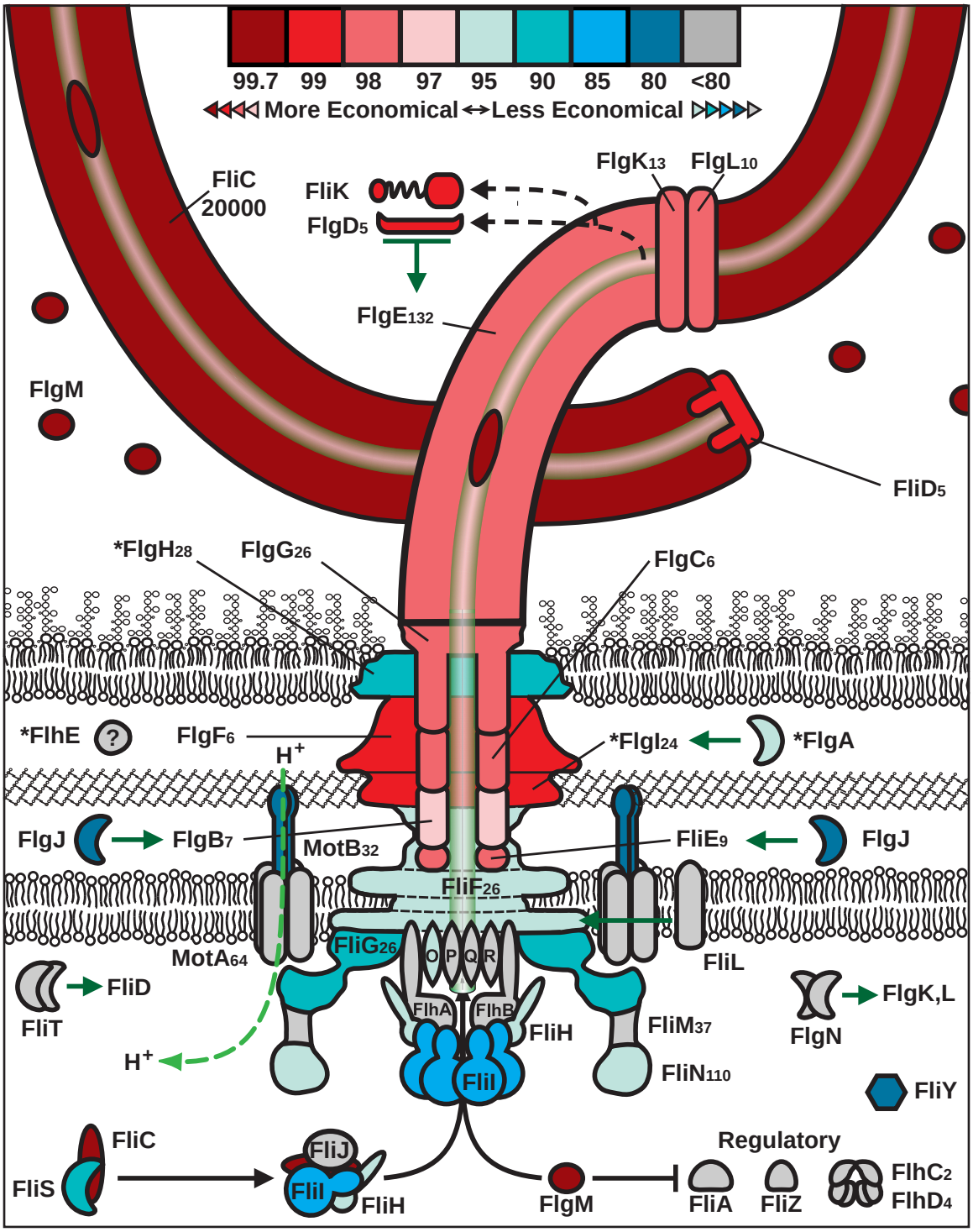


Figure 3.2. Protein abundance and cost of flagellar proteins.

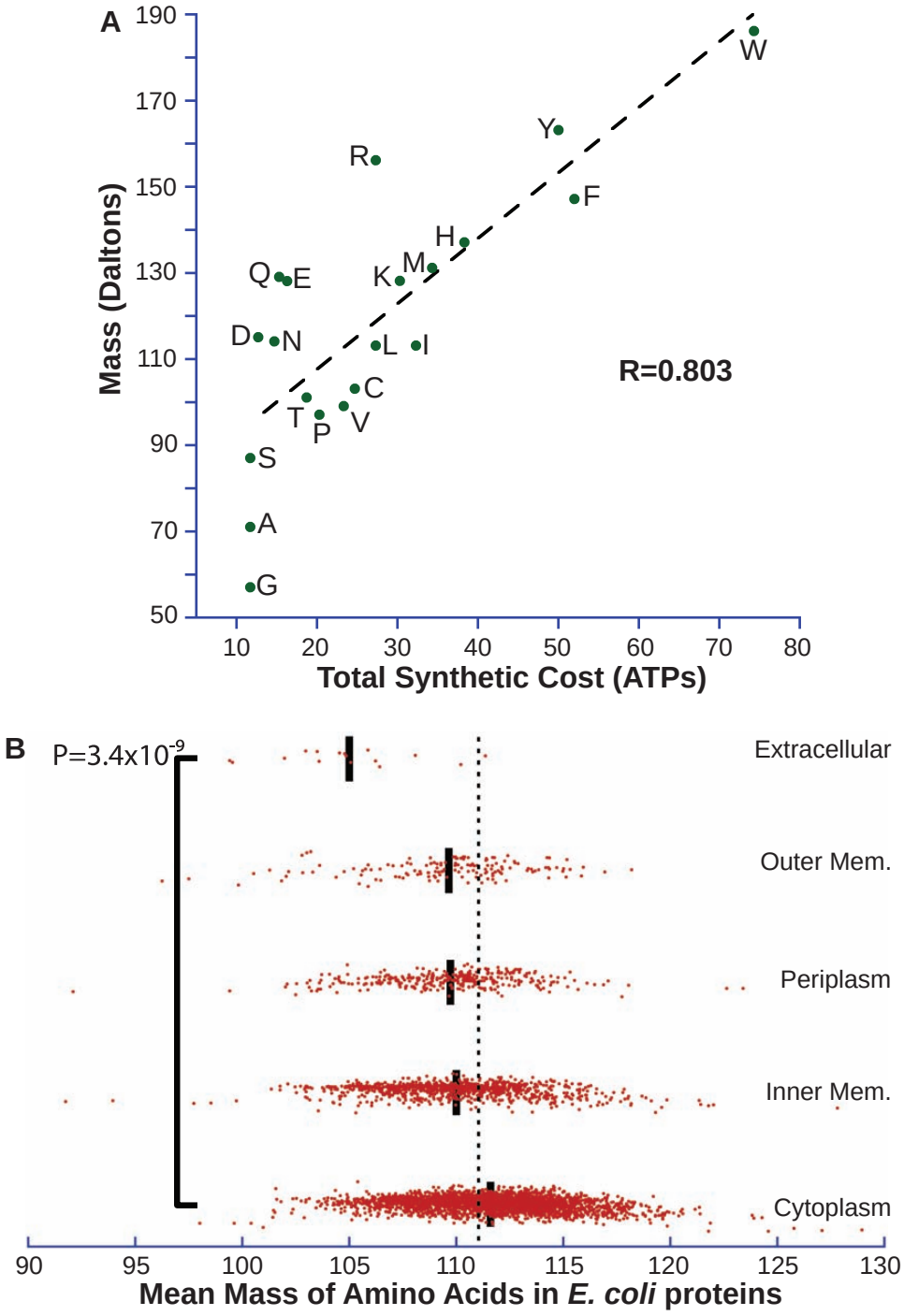


Figure 3.3. Molecular mass of amino acids correlates with their synthetic cost; extracellular proteins have simpler amino acids.

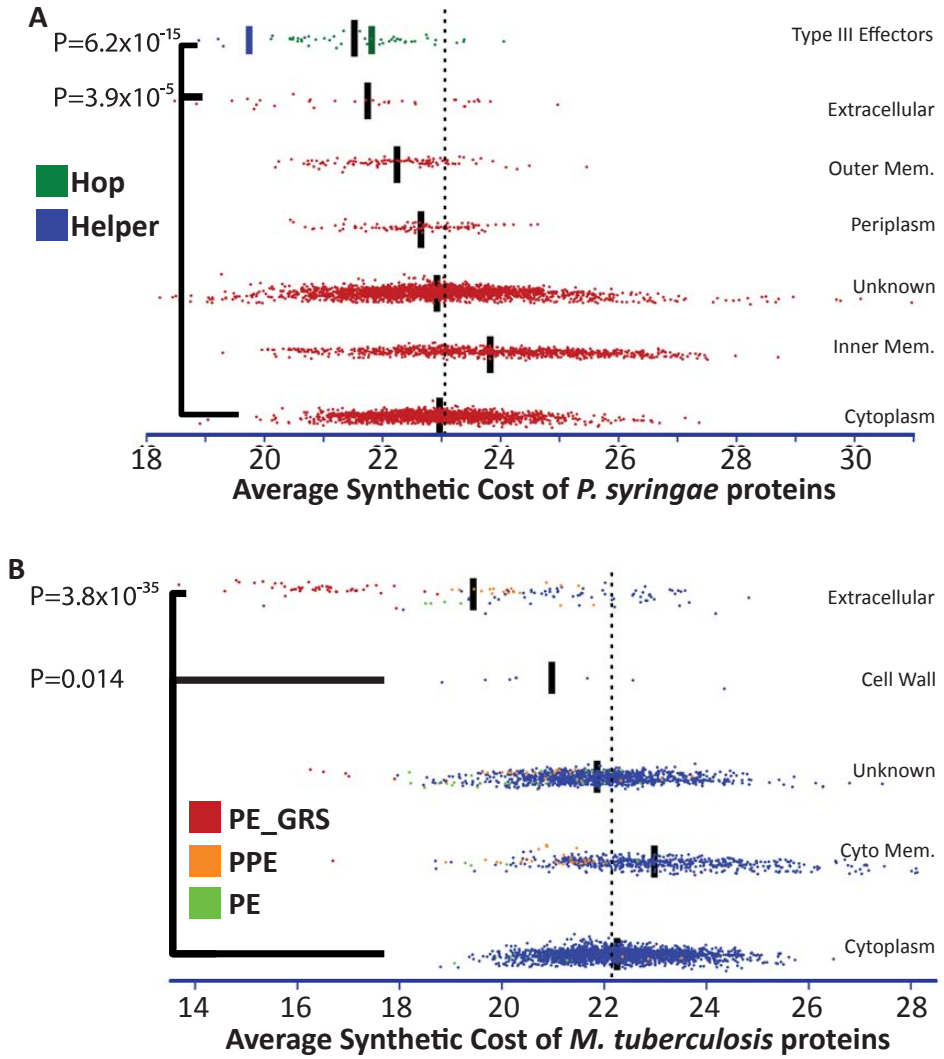


Figure 3.4. Relationship between protein location and cost extends to diverse organisms.

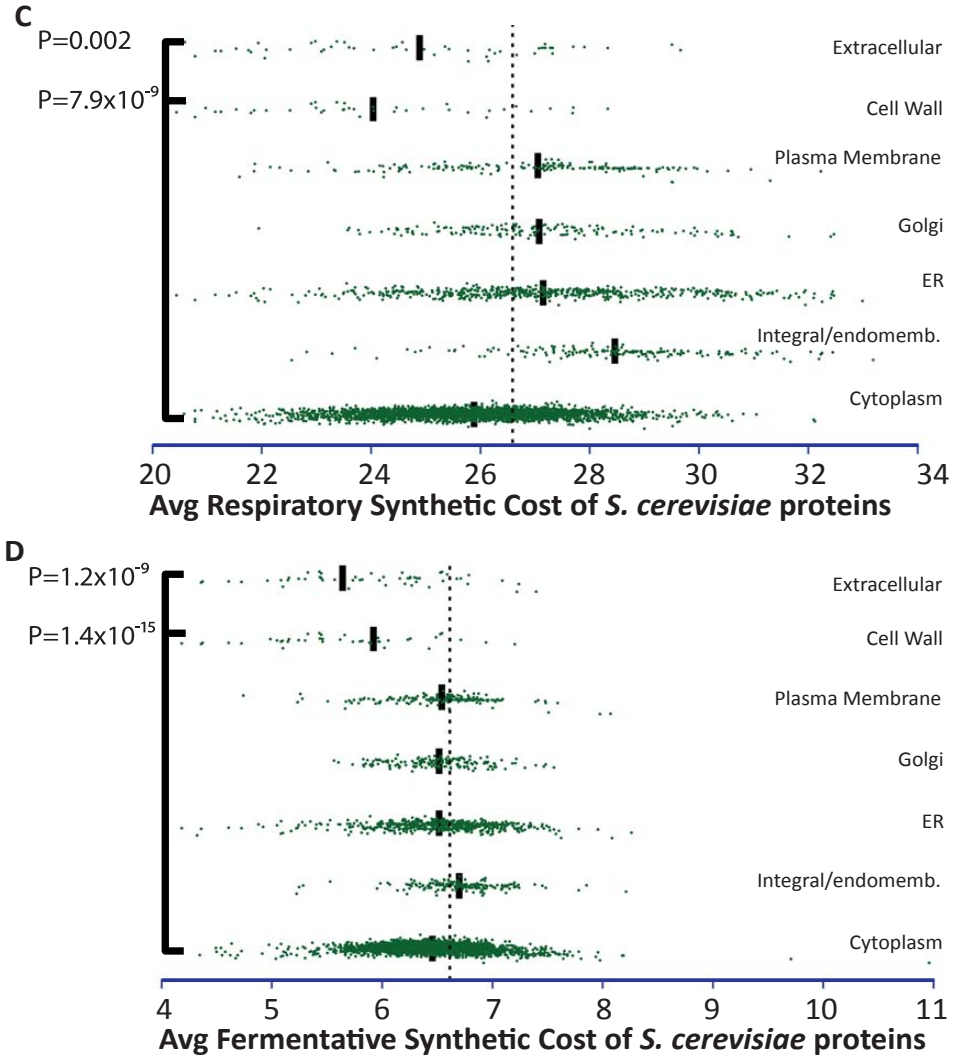


Figure 3.4 (continued). Relationship between protein location and cost extends to diverse organisms.

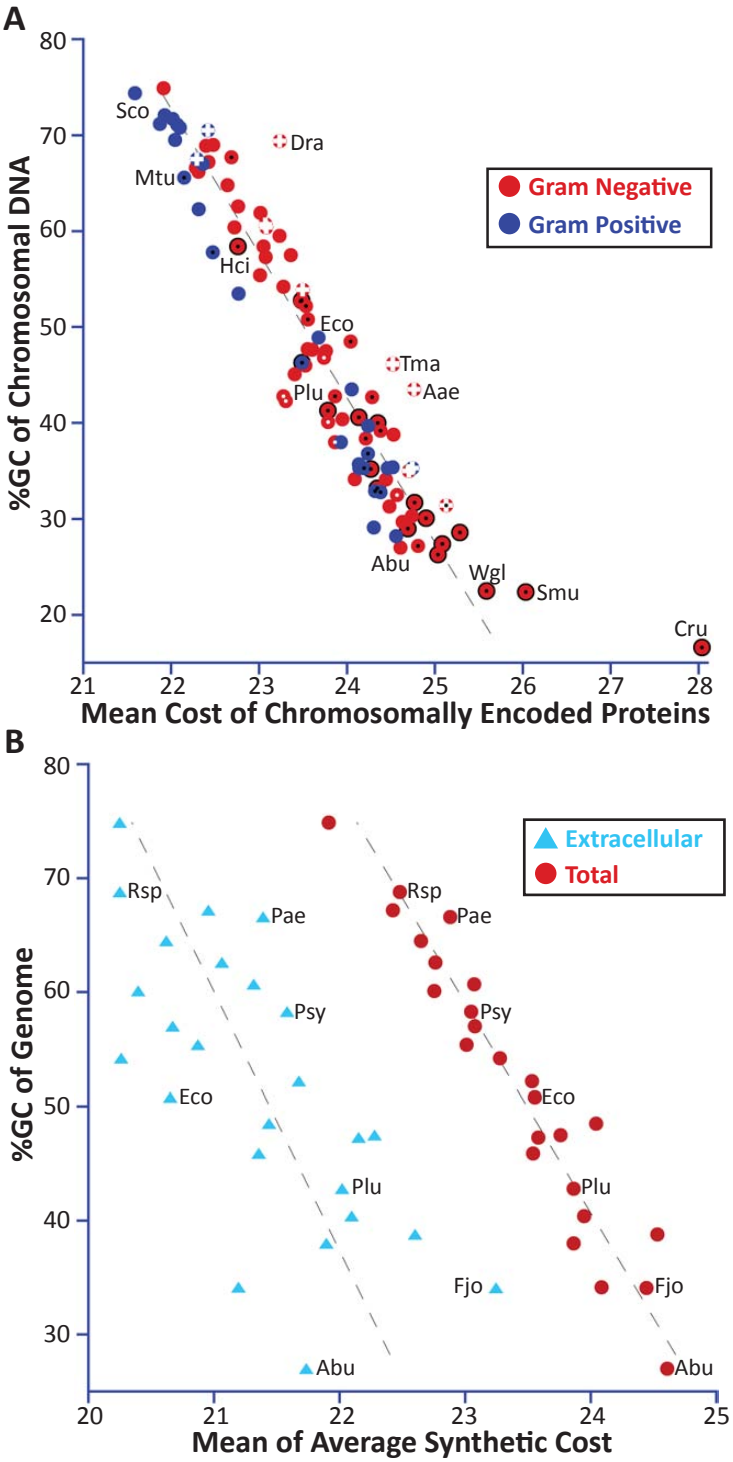


Figure 3.5. Per residue, proteins in GC-rich organism cost less to synthesize; however, extracellular proteins are still economical.

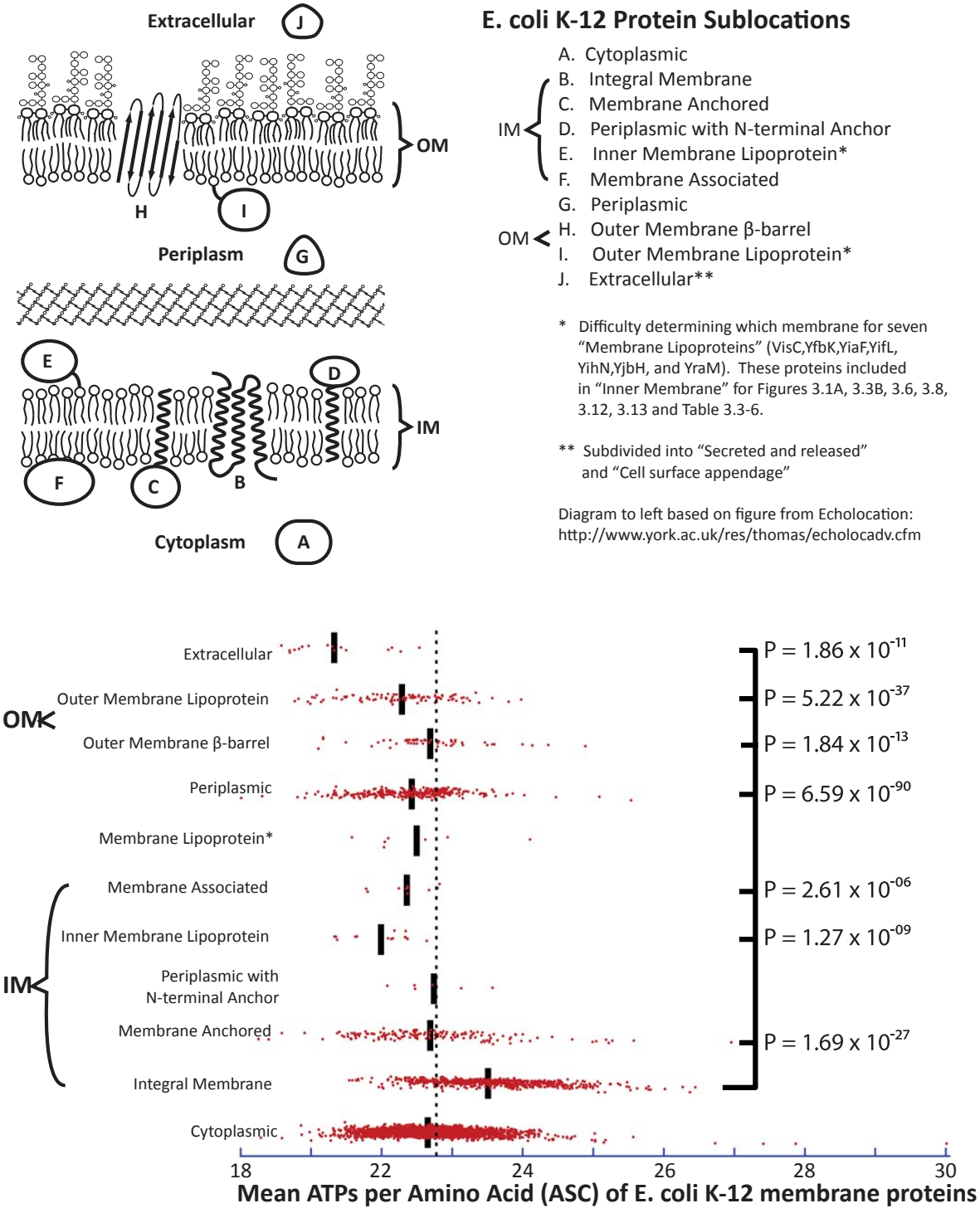
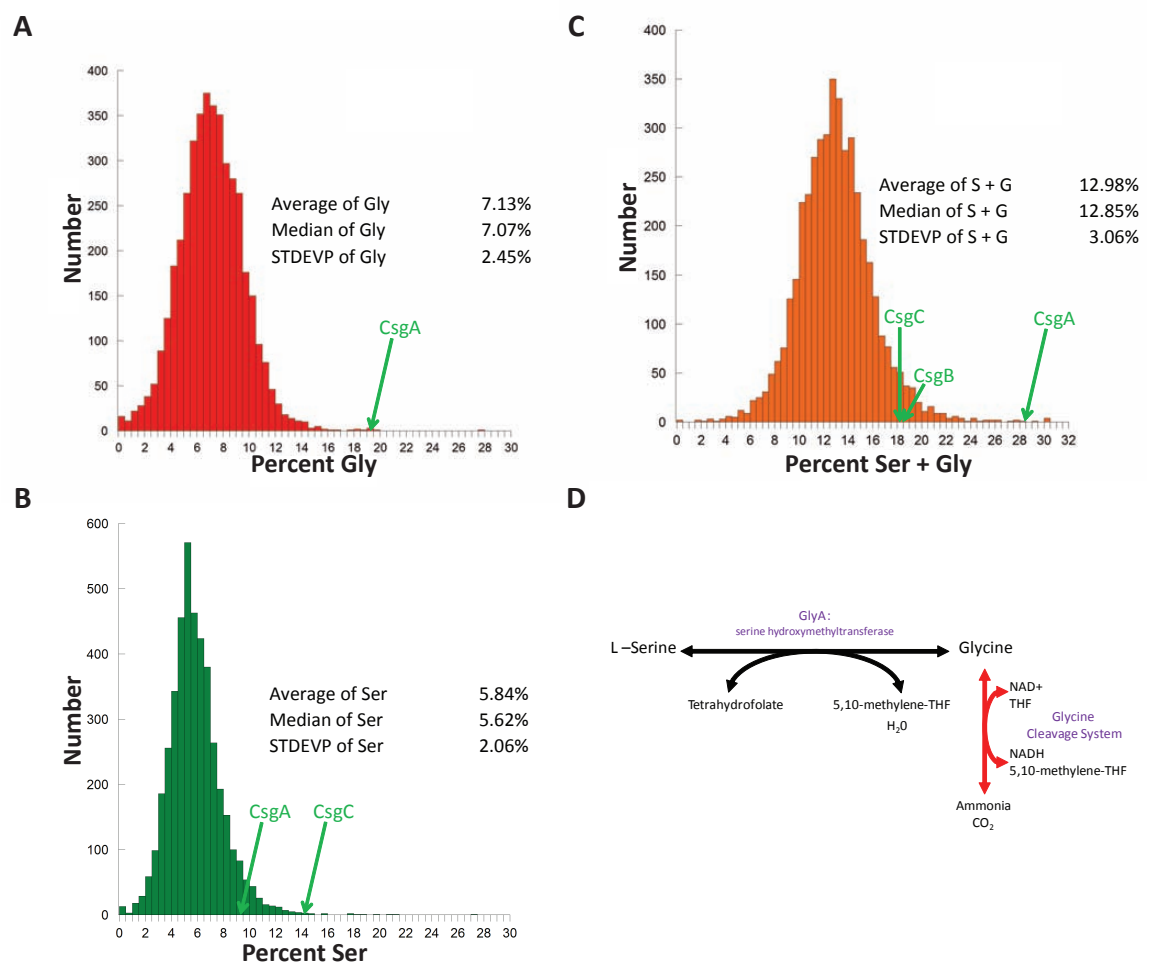


Figure 3.6. Cost and sublocations of E. coli proteins in inner and outer membranes.



Rank	% Gly	# Gly	# AA	Name
1	27.66	26	94	YmgG
2	19.51	16	82	YeaQ
3	19.23	15	78	YciG
4	19.21	29	151	CsgA
5	19.05	16	84	YmgE
6	18.84	13	69	YedS_2
7	18.06	13	72	OsmB
8	18.06	26	144	RplO
9	17.65	33	187	YfaZ
10	16.77	26	155	SlyB
11	16.29	29	178	Ssb
12	15.76	26	165	YahC
13	15.55	206	1325	YdeK
14	15.29	13	85	RpmA
15	15.20	26	171	YedS_1
16	15.15	35	231	AqpZ
17	15.12	13	86	YoeE
18	15.07	63	418	YdhQ
19	14.80	29	196	SlyD
20	14.57	159	1091	Flu (Ag43)

	% Gly	Rank Gly	% Ser	Rank Ser	%SG	Rank SG
CsgA	**19.2**	**4**	**9.3**	**228**	**28.5**	**6**
CsgB	9.9	483	8.6	361	**18.5**	**169**
CsgC	3.6	4003	**14.6**	**12**	**18.2**	**206**
CsgD	4.2	3854	8.3	414	12.5	2402
CsgE	4.7	3665	5.4	2350	10.1	3663
CsgF	8.7	1068	8.0	537	16.7	426
CsgG	6.9	2304	7.2	855	14.0	1411
Mean	7.13		5.84		12.98	
Median	7.07		5.62		12.85	
STDEVP	2.45		2.06		3.06	

Bold 95th Percentile

Figure 3.7. Glycine and Serine composition of *E. coli* proteins with an emphasis on curli specific gene proteins.

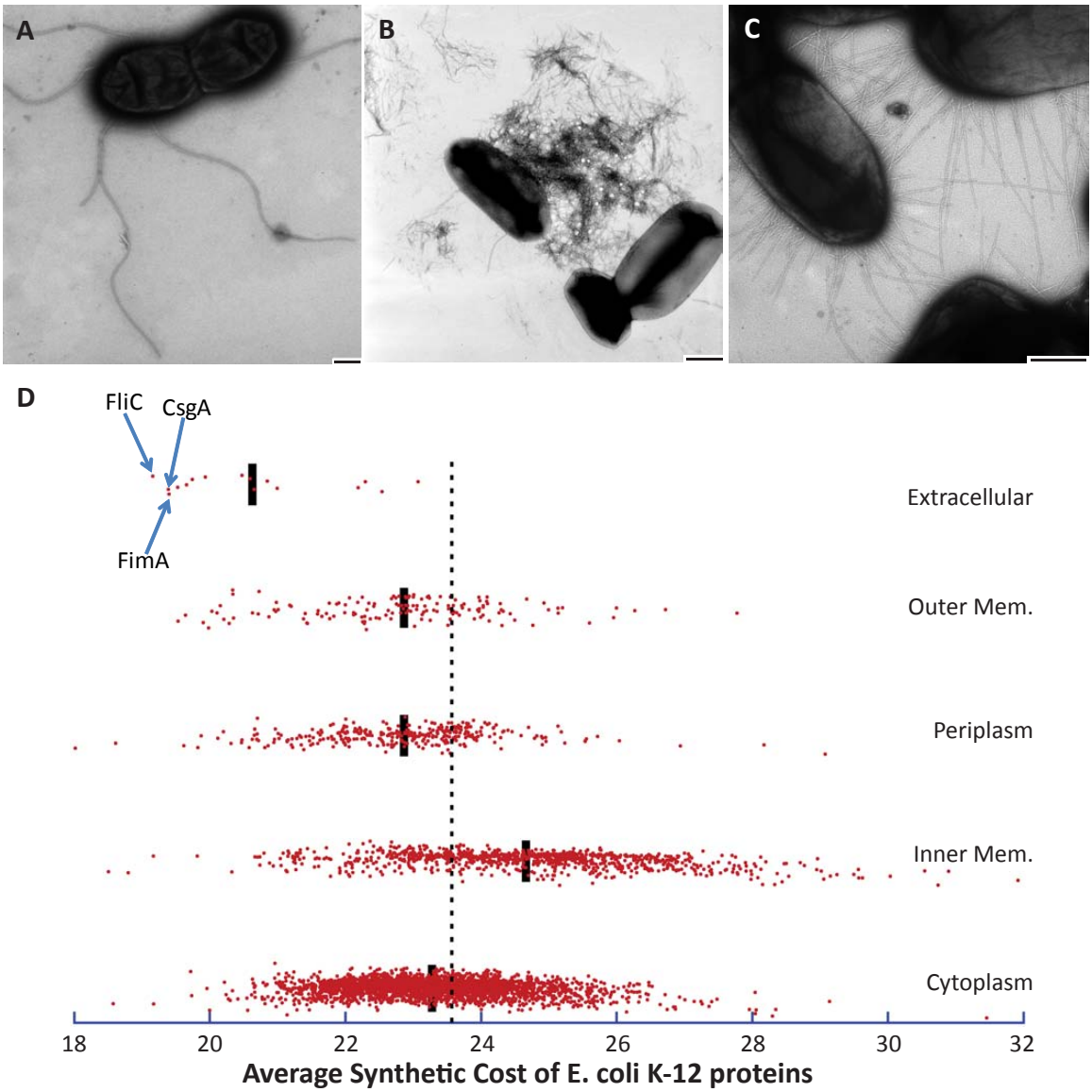
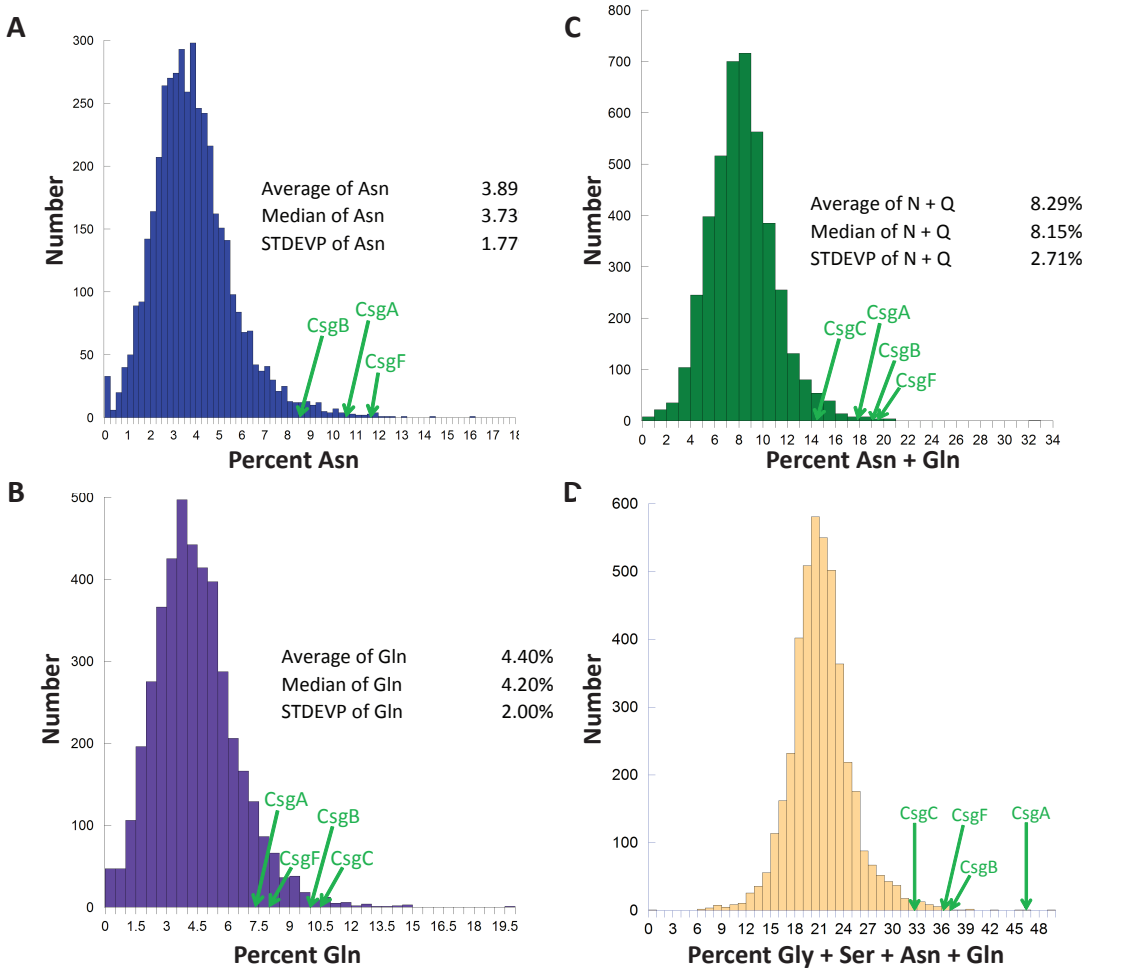


Figure 3.8. Protein abundance and cost of extracellular fibers.



	% Gly	Rank Gly	% Ser	Rank Ser	%GS	Rank GS	% Asn	Rank Asn	% Gln	Rank Gln	%NQ	Rank NQ	%GSNQ	Rank GSNQ
CsgA	**19.21**	6	**9.27***	227	**28.48**	6	**10.60**	**22**	7.28**	354	**17.89**	**18**	**46.36**	**2**
CsgB	9.93	492	8.61	361	**18.54**	165	**8.61**	**87**	**9.93**	**55**	**18.54**	**12**	**37.08**	**8**
CsgC	3.64	4028	**14.55**	12	**18.18**	198	3.64	2218	**10.91**	**33**	**14.55**	**102**	**32.73**	**47**
CsgD	4.17	3884	8.33	420	12.50	2377	**8.80**	**75**	4.17	2214	12.96	223	25.46	438
CsgE	4.65	3693	5.43	2358	10.08	3669	4.65	1160	3.88	2483	8.53	1907	18.61	3415
CsgF	8.70	1077	7.97	536	16.67	419	**11.59**	**13**	**7.97**	**218**	**19.57**	**7**	**36.23**	**13**
CsgG	6.86	2323	7.22	846	14.08	1405	5.78	531	5.78	940	11.55	460	25.63	411
Mean	7.12		5.81		12.93		3.87		4.42		8.29		21.22	
Median	7.06		5.61		12.84		3.68		4.24		8.16		21.01	
STDEVP	2.50		2.07		3.09		1.79		2.06		2.77		3.84	
$>P_{95}$	11.11		9.33		18.04		7.14		7.97		13.03		27.78	
$>P_{90}$	10.11		8.33		16.60		6.06		7.01		11.64		25.49	

* 94.7th Percentile

** 91.8th Percentile

Figure 3.9. Amino acid compositions of E. coli K-12 proteins with an emphasis on curli specific gene proteins.

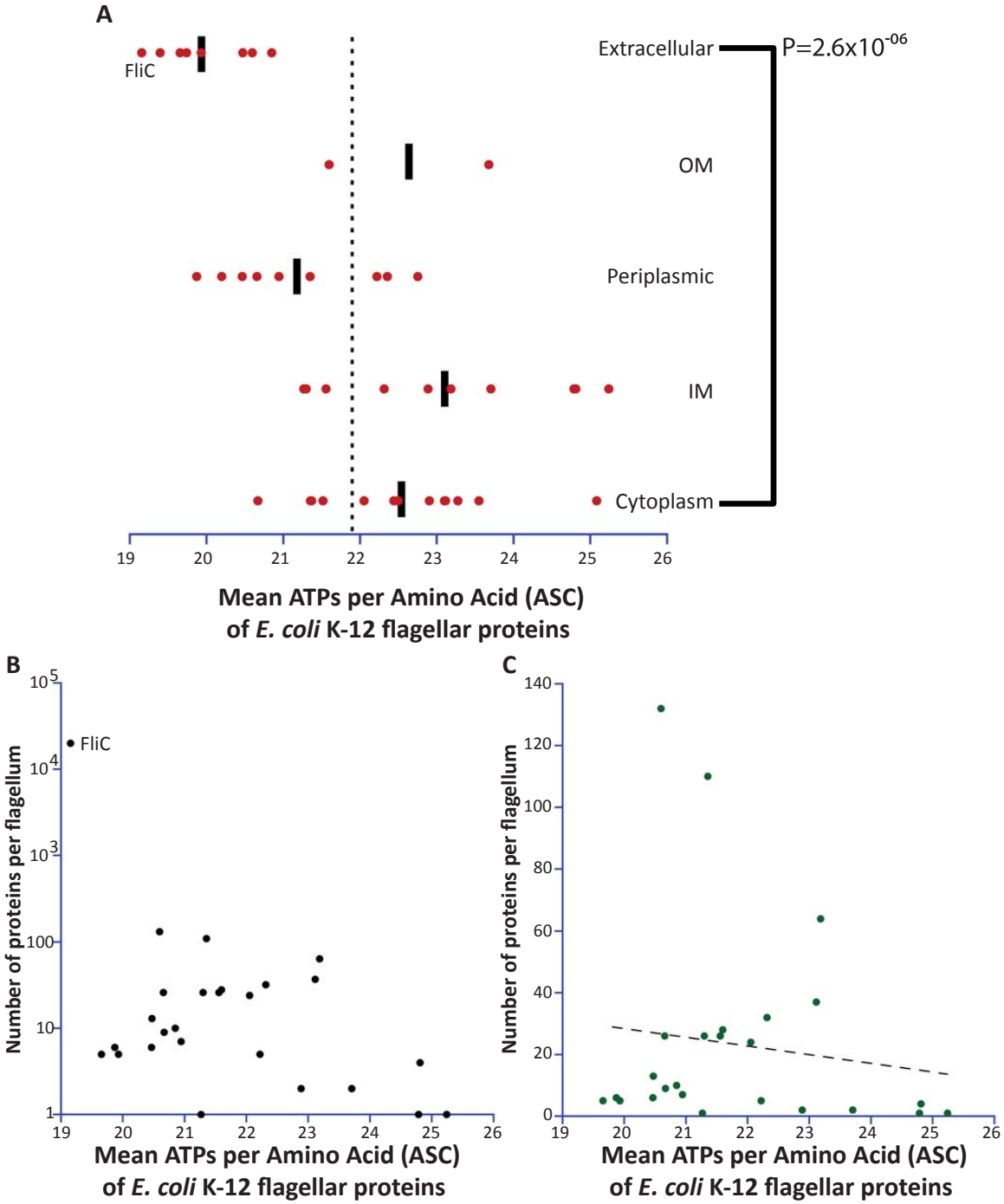


Figure 3.10. The average synthetic cost (ASC) of flagellar proteins correlates with location but not abundance.

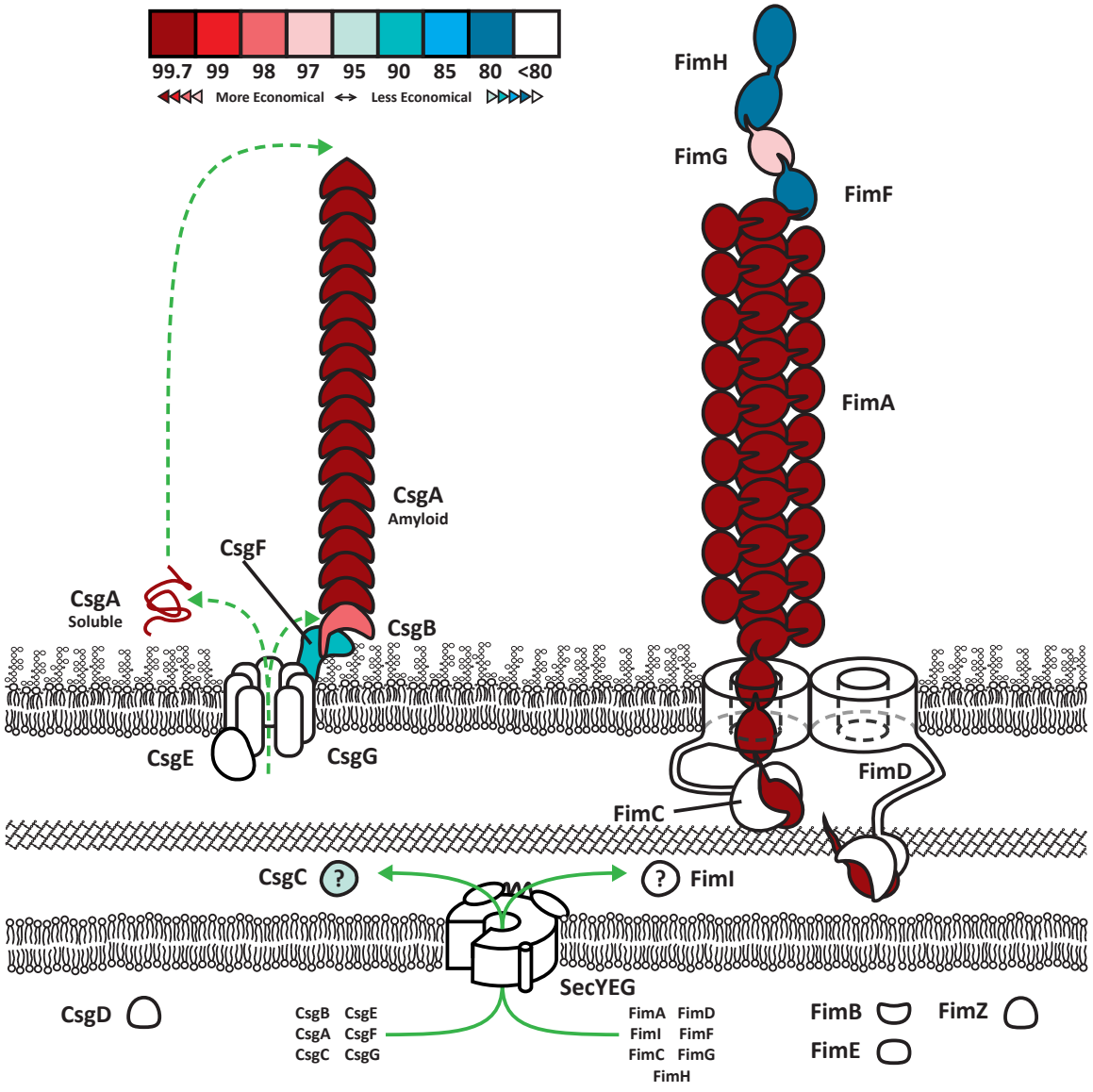


Figure 3.11. Protein economy of curli and Type 1 pili proteins.

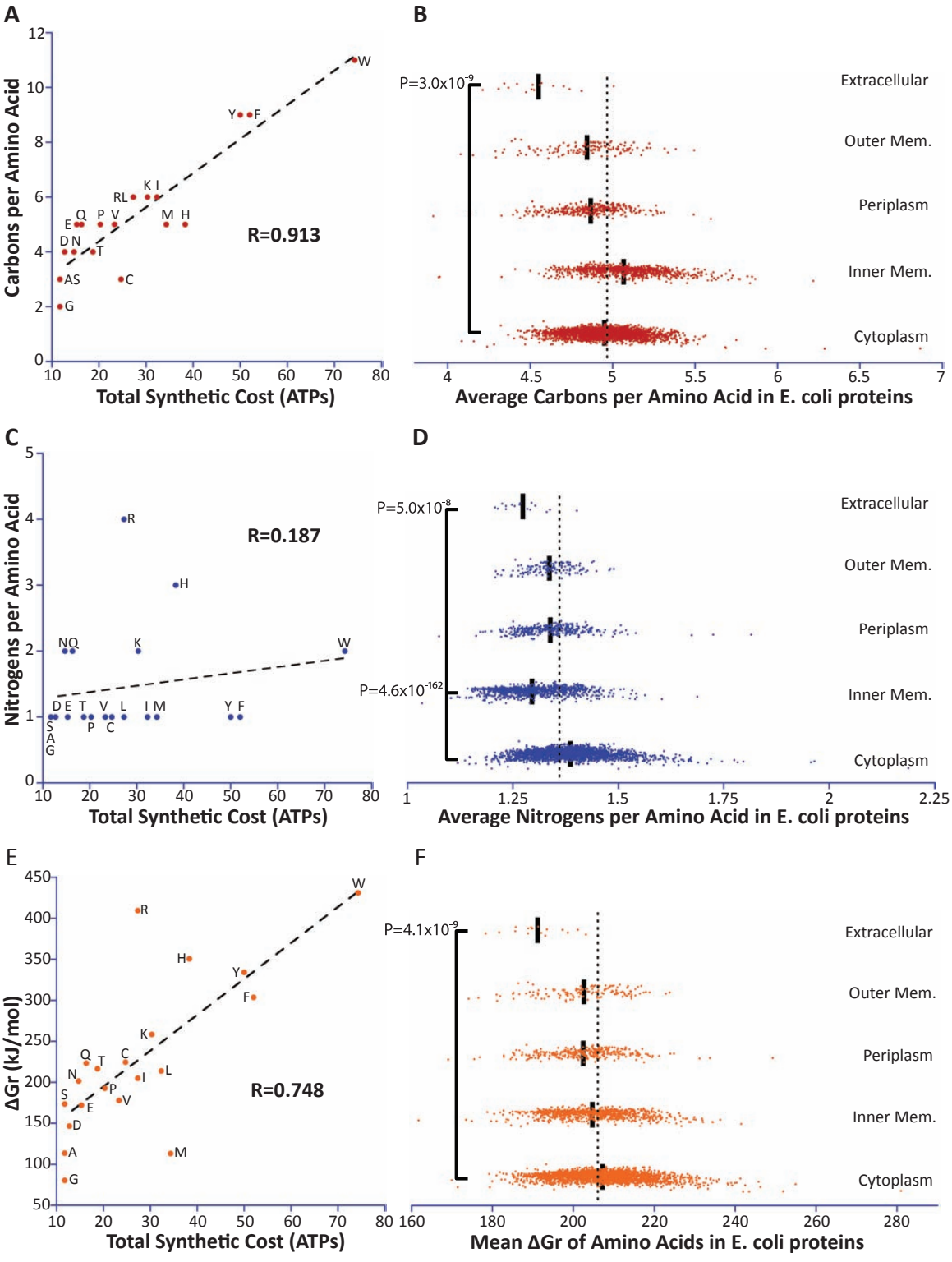


Figure 3.12. Extracellular proteins contain less carbon, nitrogen, or ΔGr per amino acid than other locations.

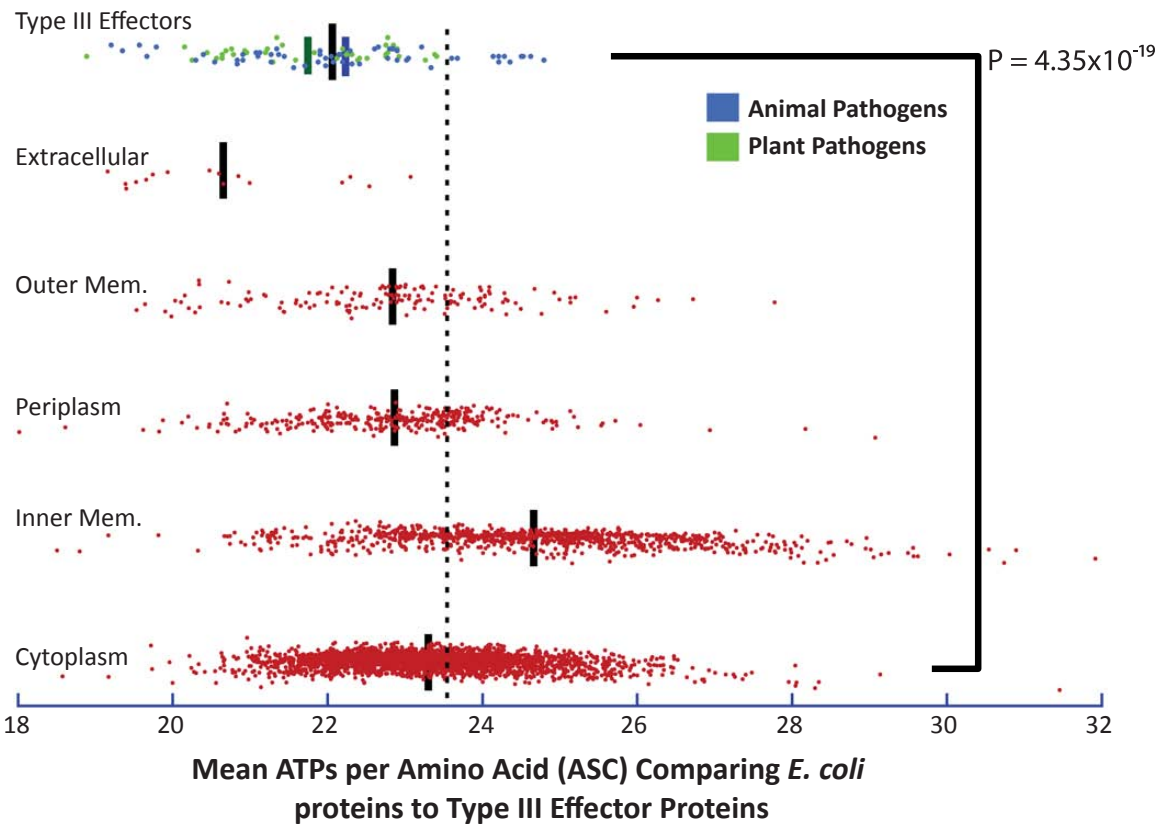


Figure 3.13. Type III effectors from multiple pathogens are economical.

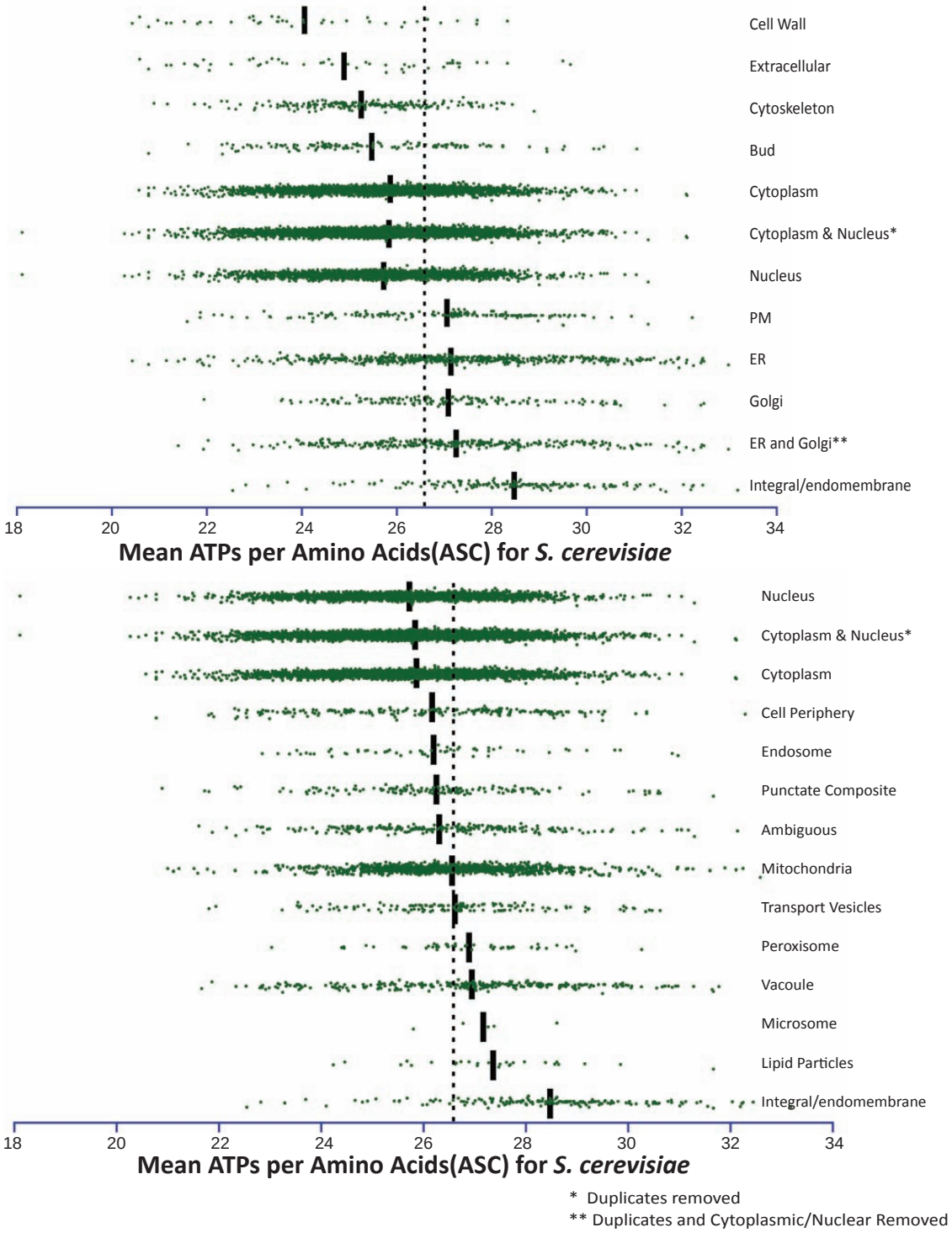


Figure 3.14. Cost and location of S. cerevisiae proteins.

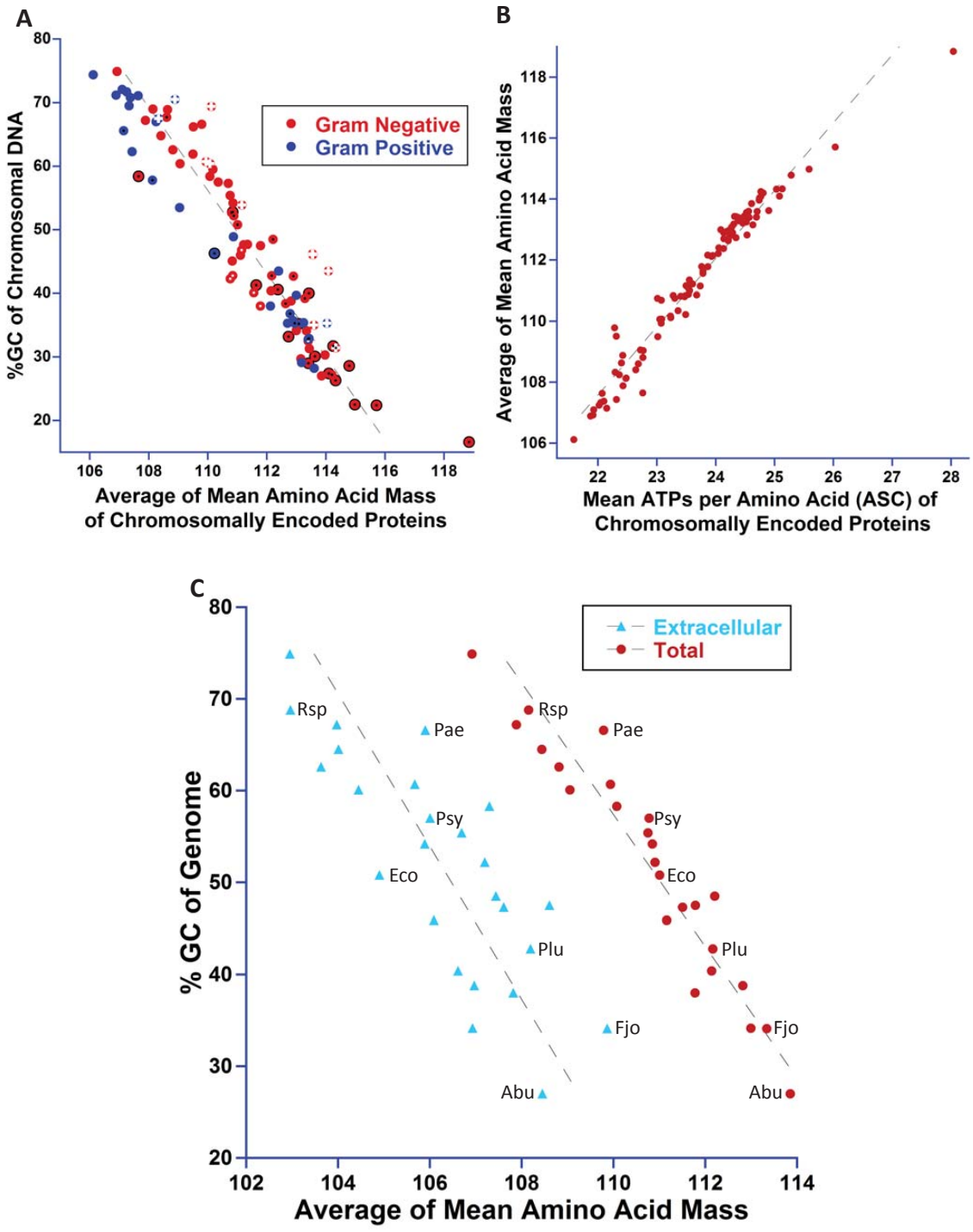


Figure 3.15. Per residue, proteins in GC-rich organisms have reduced mass; however, extracellular proteins have simpler amino acids.

Table 3.1: Synthetic cost of amino acids.

									Gibbs Free Energy (ΔGr)	
			Enzymes in *E. coli*	Codons		Molecular			Hydrothermal	Seawater
				#	%GC	Mass	Carbons	Nitrogens	(kJ mol^{-1}) 100°C	(kJ mol^{-1}) 18°C
Alanine	Ala	A	1	4	83.33	71.079	3	1	-12.12	113.66
Arginine	Arg	R	10	6	72.22	156.188	6	4	197.52	409.46
Asparagine	Asn	N	1	2	16.67	114.104	4	2	83.53	201.56
Aspartic Acid	Asp	D	1	2	50	115.089	4	1	32.78	146.74
Cysteine	Cys	C	9	2	50	103.145	3	1	60.24	224.67
Glutamic Acid	Glu	E	1	2	50	129.116	5	1	-1.43	172.13
Glutamine	Gln	Q	2	2	50	128.131	5	2	44.03	223.36
Glycine	Gly	G	4	4	83.33	57.052	2	1	14.89	80.49
Histidine	His	H	1	2	50	137.141	5	3	154.48	350.52
Isoleucine	Ile	I	11	3	11.11	113.16	6	1	-96.4	213.93
Leucine	Leu	L	7	6	38.89	113.16	6	1	-105.53	205.03
Lysine	Lys	K	10	2	16.67	128.17	6	2	-28.33	258.56
Methionine	Met	M	9	1	33.33	131.199	5	1	-174.71	113.22
Phenylalanine	Phe	F	9	2	16.67	147.177	9	1	-114.54	303.64
Proline	Pro	P	4	4	83.33	97.117	5	1	-38.75	192.83
Serine	Ser	S	3	6	50	87.078	3	1	69.47	173.73
Threonine	Thr	T	6	4	50	101.105	4	1	53.51	216.5
Tryptophan	Trp	W	12	1	66.67	186.213	11	2	-38.99	431.17
Tyrosine	Tyr	Y	9	2	16.67	163.176	9	1	-59.53	334.2
Valine	Val	V	4	4	50	99.133	5	1	-70.12	178
Average aa			5.7	3.05		118.887	5.3	1.45	-1.5	227.17
Source			(35)						(8)	(8)

Table 3.1: Synthetic cost of amino acids (continued).

	Bacteria					Yeast	
	Chemoheterophs		Photoautotrophs*	Li et al*	Craig&Weber*	Respiratory	Fermentative
	Cost (~P)	Savings (~P)	Cost (~P)	Cost (~P)	Cost (ATPs)	Cost (~P)	Cost (~P)
Alanine	11.7	15.66	11.7	11.7	12.5	14.5	2
Arginine	27.3	0.06	27.3	**38.3**	18.5	20.5	13
Asparagine	14.7	12.66	14.7	**12.7**	4	18.5	6
Aspartic Acid	12.7	14.66	12.7	**14.7**	1	15.5	3
Cysteine	24.7	2.66	24.7	24.7	24.5	26.5	13
Glutamic Acid	15.3	12.06	15.3	**34.3**	8.5	9.5	2
Glutamine	16.3	11.06	16.3	**15.3**	9.5	10.5	3
Glycine	11.7	15.66	11.7	11.7	14.5	14.5	1
Histidine	38.3	-10.94	**40.3**	**32.3**	33	29	5
Isoleucine	32.3	-4.94	32.3	**30.3**	20	38	14
Leucine	27.3	0.06	27.3	27.3	33	37	4
Lysine	30.3	-2.94	30.3	**27.3**	18.5	36	12
Methionine	34.3	-6.94	34.3	**20.3**	18.5	36.5	24
Phenylalanine	52	-24.64	**54**	52	63	61	10
Proline	20.3	7.06	20.3	**16.3**	12.5	14.5	7
Serine	11.7	15.66	11.7	11.7	15	14.5	1
Threonine	18.7	8.66	18.7	18.7	6	21.5	9
Tryptophan	74.3	-46.94	**76.3**	74.3	78.5	75.5	14
Tyrosine	50	-22.64	**52**	50	56.5	59	8
Valine	23.3	4.06	23.3	23.3	25	29	4
Average aa	27.36		27.76	27.36	23.625	29.075	7.75
Source	(3, 63)		(63)	(81)	(35)	(160)	(160)

* Numbers in **BOLD** differ from chemoheterotrophic bacteria.

Notes: Values in Akashi and Gojobori (3) are based on the average of overall synthetic cost under growth on three substrates: glucose, acetate, and malate. Cost varied slightly with different carbon sources but were highly correlated (glucose vs acetate r^2=0.968, glucose vs malate r^2=0.998)(3). Bragg and Wagner (23) use 16 ~P (respiratory) and 1 ~P (fermentative) for Lys instead of 36 ~P and 12~P from Wagner (160); Lys production in yeast uses α-ketoglutarate instead of oxaloacetate(23). Average cost (equal amino acids) for respiratory and fermentative growth using these numbers would be 28.075 and 7.2 respectively.

Li et al(81) has the same values for eukaryotes as Wagner (160) and sources Heizer et al (63) values for prokaryotes. However the values listed in their supplementary section do not match with Heizer et al (63): Arg is listed 38.3 not 27.3, Asp and Asn values were mixed up, Glu is listed 34.3 not 15.3, Gln is listed 15.3 not 16.3, His is listed 32.3 not 38.3, Ile is listed 30.3 not 32.3, Lys is listed 27.3 not 30.3, Met is listed 20.3 not 34.3, and Pro is listed 16.3 not 20.3. No explanation is given as to why these values were used.

Table 3.2: Correlation between different amino acid synthetic costs.

	Spearman		Kendall			Linear Correlation (Pearson)	
vs Chemoheterotrophic	**R**	**p value**	**Tau**	**p value**	**Gamma**	**R**	**p value**
Number of Enyzmes *E. coli*	0.6974	6.31E-04	0.5736	4.06E-04	0.6145	0.692	7.24E-04
Number of Codons	-0.4565	0.0431	-0.3427	0.0346	-0.4058	-0.4309	0.0579
Number of Carbons	0.8429	3.08E-06	0.7176	9.71E-06	0.7875	0.9125	2.07E-08
Number of Nitrogens	0.2536	0.2806	0.193	0.2341	0.2826	0.1871	0.4295
Average Codon %GC	-0.4438	0.05	-0.3536	0.0293	-0.4	-0.2855	0.2223
Average Codon %GC Weighted*	-0.4211	0.0645	-0.2979	0.0663	-0.3011	-0.3026	0.1947
Molecular Mass	0.7502	1.39E-04	0.6027	2.03E-04	0.6108	0.803	2.02E-05
Surface Seawater ΔG(kJ mol-1) 18°C	0.7149	3.97E-04	0.5639	5.09E-04	0.5699	0.748	1.49E-04
Hydrothermal Sol. ΔG (kJ mol-1) 100°C	-0.4588	0.0419	-0.2979	0.0663	-0.3011	-0.326	0.1607
Photoautotrophic		<0.000001				0.9996	→0, t=156
Yeast respiratory	0.8737	4.87E-07	0.7147	1.05E-05	0.7363	0.954	7.37E-11
Yeast fermentative	0.7724	6.59E-05	0.6183	1.38E-04	0.6354	0.5414	0.0137
Yeast respiratory Bragg	0.8427	3.12E-06	0.672	3.43E-05	0.6923	0.936	1.35E-09
Yeast fermentative Bragg	0.6788	0.001	0.546	7.64E-04	0.5642	0.5155	0.02
Li et al Prokaryotic	0.8615	1.06E-06	0.7566	3.10E-06	0.773	0.9256	5.08E-09
Craig&Weber *E. coli*	0.809	1.57E-05	0.6307	1.01E-04	0.6464	0.9337	1.84E-09
vs Yeast respiratory							
Number of Enyzmes *E. coli*	0.7223	3.23E-04	0.5525	6.60E-04	0.5976	0.7328	2.38E-04
Number of Codons	-0.3207	0.168	-0.2575	0.1124	-0.3088	-0.3748	0.1035
Number of Carbons	0.7103	4.49E-04	0.5913	2.67E-04	0.6561	0.8809	2.96E-07
Number of Nitrogens	0.0417	0.8616	0.0292	0.8569	0.0435	-0.0059	0.9805
Average Codon %GC	-0.5558	0.0109	-0.4436	0.0063	-0.5067	-0.4183	0.0664
Molecular Mass	0.5272	0.0169	0.3871	0.017	0.3956	0.6895	7.71E-04
Surface Seawater ΔG(kJ mol-1) 18°C	0.5459	0.0128	0.4129	0.0109	0.4208	0.6131	0.004
Hydrothermal Sol. ΔG (kJ mol-1) 100°C	-0.5595	0.0103	-0.3164	0.0511	-0.3224	-0.4797	0.0323
Photoautotrophic	0.8737	4.87E-07	0.7147	1.05E-05	0.7363	0.9523	1.02E-10
Yeast fermentative	0.7106	4.46E-04	0.5203	0.0013	0.5393	0.514	0.0204
Yeast respiratory Bragg	0.9682	2.85E-12	0.9	2.89E-08	0.9344	0.9702	1.56E-12
Yeast fermentative Bragg	0.6167	0.0038	0.4469	0.0059	0.4659	0.4693	0.0368
Li et al Prokaryotic	0.6778	1.02E-03	0.5547	6.28E-04	0.5714	0.859	1.24E-06
Craig&Weber *E. coli*	0.8063	1.76E-05	0.6541	5.53E-05	0.676	0.9286	3.54E-09
vs Molecular Mass							
Number of Enyzmes *E. coli*	0.3811	0.0973	0.3068	0.0586	0.3254	0.5097	0.0217
Number of Codons	-0.6168	0.0038	-0.5225	0.0013	-0.6115	-0.4507	0.0461
Number of Carbons	0.7633	9.03E-05	0.6364	8.74E-05	0.6914	0.8669	7.59E-07
Number of Nitrogens	0.5111	0.0213	0.4265	0.0086	0.6129	0.474	0.0347
Average Codon %GC	-0.3603	0.1187	-0.2823	0.0819	-0.3158	-0.3606	0.1184
Average Codon %GC Weighted*	-0.4438	0.05	-0.3325	0.0404	-0.3333	-0.3769	0.1014
Surface Seawater ΔG(kJ mol-1) 18°C	0.6702	0.0012	0.533	0.001	0.5344	0.8352	4.60E-06
Hydrothermal Sol. ΔG (kJ mol-1) 100°C	-0.1121	0.6381	-0.0792	0.6256	-0.0794	-0.0357	0.8813
Photoautotrophic	0.7502	1.39E-04	0.6027	2.03E-04	0.6108	0.8024	2.08E-05
Yeast fermentative	0.53	0.0162	0.3807	0.0189	0.388	0.4854	0.03
Yeast respiratory Bragg	0.503	0.0238	0.3656	0.0242	0.3736	0.6667	0.0013
Yeast fermentative Bragg	0.4664	0.0382	0.3289	0.0426	0.337	0.4488	0.0472
Li et al Prokaryotic	0.806	1.79E-05	0.6774	2.97E-05	0.6865	0.8434	3.00E-06
Craig&Weber *E. coli*	0.4414	0.0514	0.3102	0.0559	0.3152	0.6568	0.0017

* Based on codon bias of W3110: http://www.kazusa.or.jp/codon/

Table 3.2: Correlation between different amino acid synthetic costs (continued).

	Spearman		Kendall			Linear Correlation (Pearson)	
vs Average Codon %GC	**R**	**p value**	**Tau**	**p value**	**Gamma**	**R**	**p value**
Number of Enyzmes *E. coli*	-0.2779	0.2355	-0.2526	0.1195	-0.3043	-0.3195	0.1698
Number of Codons	0.3559	0.1236	0.2104	0.1947	0.2846	0.3803	0.0981
Average Codon %GC Weighted*	0.9375	1.11E-09	0.8504	1.58E-07	0.9477	0.984	6.18E-15
Number of Carbons	-0.4215	0.0642	-0.3365	0.0381	-0.4179	-0.3484	0.1323
Number of Nitrogens	0.0489	0.8377	0.0429	0.7914	0.0789	0.1211	0.6111
vs Number of Codons							
Number of Enyzmes *E. coli*	-0.1029	0.666	-0.0601	0.7109	-0.0758	-0.0982	0.6805

* Based on codon bias of W3110: http://www.kazusa.or.jp/codon/

Table 3.3: Location and ASC of proteins in (A) *E. coli* (B) *S. cerevisiae* (C) *M tuberculosis*.

Available online in a published work at *mBio* in an excel format.

Table 3.4: Localization of proteins with extreme average amino acid costs in *E. coli*.

	Overall Number	Overall Percent	Top 100	Top 200	Top 300	Bottom 100	Bottom 200	Bottom 300	Increase fold Top 100	Increase fold Bottom 100
Cytoplasmic	2859	65.96	36	95	167	19	38	68	0.546	0.288
IM	972	22.43	12	34	43	78	156	223	0.535	3.477
Periplasmic	337	7.80	24	37	50	2	3	4	3.077	0.256
OM	149	3.44	17	22	28	1	3	5	4.944	0.291
Extracellular	16	0.37	11	12	12	0	0	0	29.789	0
Total	4333	65.96	100	200	300	100	200	300		

Table 3.5A: Statistics comparing ASCs of proteins in different locations of *E. coli*.

	Mann Whitney U		Kolmogorov–Smirnov	
vs Cytoplasm	**Z**	**p value**	**Max Dif**	**p value**
IM	23.196	4.98E-119	-0.418	9.82E-112
Periplasm	6.029	1.65E-09	0.156	6.71E-07
OM	3.713	2.05E-04	0.144	4.82E-03
Extracellular	5.983	2.19E-09	0.674	3.57E-07
vs All				
Cytoplasm	6.223	4.89E-10	0.084	5.46E-11
IM	19.529	6.17E-85	-0.336	5.73E-79
Periplasm	8.242	1.69E-16	0.225	2.26E-14
OM	5.149	2.62E-07	0.174	2.66E-04
Extracellular	6.001	1.96E-09	0.663	5.70E-07
vs IM				
Periplasm	17.023	5.56E-65	0.549	6.61E-67
OM	11.598	4.21E-31	0.489	7.70E-28
Extracellular	6.447	1.14E-10	0.776	2.74E-09
vs OM				
Periplasm	0.146	0.8839	0.107	0.1803
Extracellular	4.868	1.13E-06	-0.567	8.89E-05

	Student T Test*				
vs Cytoplasm	**T value**	**95% CI**	**99.9% CI**	**Deg Free**	**p value**
IM	22.178	1.962	3.298	1314	7.87E-93
Periplasm	6.378	1.966	3.314	414	4.79E-10
OM	3.818	1.975	1.975	159	1.93E-04
Extracellular	8.461	2.131	4.073	15	4.29E-07

* Assumes unequal variance. Confidence intervals are 2 tailed
Mann Whitney U test results used instead of student t test
Most locations fail normality (even taking Square root of data)

Table 3.5B: Statistics comparing ASCs of proteins in different locations of *S. cerevisiae.*

	Mann Whitney U		Kolmogorov–Smirnov	
vs Cytoplasm	**Z**	**p value**	**Max Dif**	**p value**
Extracellular	3.08	0.0021	-0.271	6.09E-04
Cell Wall	5.78	7.46E-09	-0.432	9.93E-08
Cytoskeleton	5.654	1.56E-08	-0.196	7.05E-07
Bud	3.345	0.0008	-0.211	4.66E-06
Nucleus	3.609	0.0003	-0.056	9.77E-04
PM	9.011	2.04E-19	0.41	1.92E-26
ER	13.623	2.94E-42	0.311	1.33E-39
Golgi	7.736	1.02E-14	0.281	6.95E-11
ER or Golgi Unique**	11.7	1.28E-31	0.338	1.08E-29
Integral/Endo Membranes	16.56	1.35E-61	0.648	7.49E-61
Cell Periphery	2.78	0.0054	0.216	1.07E-08
Trans Vesicles	5.295	1.19E-07	0.241	2.21E-07
Mitochondria	11.708	1.16E-31	0.175	5.03E-21
Peroxisome	5.031	4.88E-07	0.378	5.19E-07
Endosome	0.857	0.3912	0.125	0.315
Vacoule	9.486	2.41E-21	0.329	3.85E-25
Microsome	2.113	0.0346	0.732	4.09E-03
Lipid Particles	4.712	2.45E-06	0.504	1.15E-06
Punctate Comp.	2.261	0.0237	0.128	0.0221
Ambigious	3.368	0.0008	0.145	1.73E-04
ER or Golgi Only***	13.398	6.25E-41	0.513	1.42E-34
Extracellular or Cell Wall	4.08	4.51E-05	-0.286	1.93E-05
Cyto & Nuc*	1.127	0.2598	-0.017	0.741
All*	14.862	5.85E-50	0.159	8.70E-45

* Duplicates removed or none present

** Duplicates, Cytoplasmic, Nuclear removed

*** Only ER, Golgi, or both

Table 3.5B: Statistics comparing ASCs of proteins in different locations of *S. cerevisiae* (continued).

	Mann Whitney U		Kolmogorov–Smirnov	
	Z	p value	Max Dif	p value
vs All*				
Extracellular	4.846	1.26E-06	-0.294	0.0001
Cell Wall	6.898	5.27E-12	-0.47	3.57E-09
Cytoskeleton	9.542	1.41E-21	-0.299	4.40E-16
Bud	6.606	3.94E-11	-0.279	1.36E-10
Cytoplasm	14.862	5.85E-50	-0.159	8.70E-45
Nucleus	16.904	4.19E-64	-0.183	6.98E-48
PM	4.39	1.13E-05	0.255	7.21E-11
ER	6.641	3.11E-11	0.156	2.18E-11
Golgi	3.324	0.0009	0.144	0.003
ER or Golgi Unique**	5.95	2.68E-09	0.182	1.82E-09
Integral/Endo Membranes	12.337	5.71E-35	0.493	1.66E-36
Cell Periphery	1.701	0.0889	-0.126	0.0023
Trans Vesicles	1.031	0.3024	0.104	0.092486
Mitochondria	1.17	0.242	0.084	4.60E-06
Peroxisome	2.031	0.0423	0.264	0.0011
Endosome	1.529	0.1262	-0.093	0.6795
Vacoule	3.988	6.66E-05	0.181	2.84E-08
Microsome	1.059	0.2897	0.584	0.0387
Lipid Particles	2.566	0.0103	0.363	0.0011
Punctate Comp.	1.559	0.1191	-0.084	0.2713
Ambigious	1.628	0.1035	-0.061	0.343
ER or Golgi Only***	9.218	3.02E-20	0.357	1.77E-17
Extracellular or Cell Wall	6.026	1.68E-09	-0.32	8.49E-07
Cyto & Nuc*	17.351	1.93E-67	-0.159	1.70E-53

* Duplicates removed or none present

** Duplicates, Cytoplasmic, Nuclear removed

*** Only ER, Golgi, or both

Table 3.5B: Statistics comparing ASCs of proteins in different locations of *S. cerevisiae* (continued).

*vs Cyto & Nuc**	Mann Whitney U		Kolmogorov–Smirnov	
	Z	p value	Max Dif	p value
Extracellular	2.9331	0.0034	-0.2616	0.001
Cell Wall	5.6425	1.68E-08	-0.4221	1.94E-07
Cytoskeleton	5.2603	1.44E-07	-0.1812	4.79E-06
Bud	3.0696	0.0021	-0.1976	2.01E-05
Cytoplasm	1.1269	0.2598	0.0169	0.7412
Nucleus	2.7344	0.0062	-0.0388	0.0329
PM	9.2554	2.13E-20	0.4125	4.14E-27
ER	14.2357	5.50E-46	0.3127	9.94E-42
Golgi	8.0091	1.16E-15	0.2868	1.77E-11
ER or Golgi Unique**	12.1017	1.04E-33	0.339	1.57E-30
Integral/Endo Membranes	16.7279	8.21E-63	0.6476	1.63E-61
Cell Periphery	3.0619	0.0022	0.219	4.43E-09
Trans Vesicles	5.5631	2.65E-08	0.2486	6.98E-08
Mitochondria	12.7576	2.83E-37	0.1906	1.99E-26
Peroxisome	5.1641	2.42E-07	0.3945	1.27E-07
Endosome	1.0502	0.2936	0.1283	0.2846
Vacoule	9.8263	8.68E-23	0.334	2.17E-26
Microsome	2.1375	0.0326	0.7372	0.0037
Lipid Particles	4.769	1.85E-06	0.5112	7.58E-07
Punctate Comp.	2.5557	0.0106	0.1258	0.0255
Ambigious	3.7317	0.0002	0.1513	6.27E-05
ER or Golgi Only***	13.5775	5.45E-42	0.5142	4.10E-35
Extracellular or Cell Wall	3.9124	9.14E-05	-0.2767	3.84E-05
All*	17.3513	1.93E-67	0.1593	1.70E-53

* Duplicates removed or none present

** Duplicates, Cytoplasmic, Nuclear removed

*** Only ER, Golgi, or both

Table 3.5B: Statistics comparing ASCs of proteins in different locations of *S. cerevisiae* (continued).

	Student T†				
vs Cytoplasm	**T value**	**95% CI**	**99.9% CI**	**Deg Free**	**p value**
Extracellular	3.1178	2.0049	3.48	54	0.0029
Cell Wall	5.7357	2.0154	3.5258	44	8.24E-07
Cytoskeleton	5.6785	1.97	3.332	237	3.96E-08
Bud	2.4981	1.975	3.3528	159	0.0135
Nucleus	3.6015	1.9605	3.2926	4592	0.0003
PM	7.982	1.972	3.3401	199	1.12E-13
ER	12.8933	1.9636	3.3055	654	4.72E-34
Golgi	7.9141	1.9741	3.349	169	3.13E-13
ER or Golgi Unique**	11.3575	1.9666	3.3178	360	9.33E-26
Integral/Endo Membranes	18.3235	1.9728	3.3436	186	1.59E-43
Cell Periphery	2.0926	1.9702	3.333	232	0.0375
Trans Vesicles	5.1461	1.9758	3.3561	151	8.16E-07
Mitochondria	12.3166	1.9612	3.2958	1859	1.45E-33
Peroxisome	5.4403	2.0057	3.4838	53	1.38E-06
Endosome	1.2543	2.0017	3.4663	58	0.2148
Vacoule	8.5034	1.9675	3.3218	314	7.60E-16
Microsome	2.8618	2.7764	8.6103	4	0.0458
Lipid Particles	4.9165	2.0555	3.7066	26	4.19E-05
Punctate Comp.	2.36	1.9762	3.3579	147	0.0196
Ambigious	3.6335	1.969	3.3279	263	0.0003
ER or Golgi Only***	13.7257	1.9747	3.3516	162	6.17E-29
Extracellular or Cell Wall	4.0728	1.9939	3.4329	71	0.0001
Cyto & Nuc*	1.1471	1.9603	3.2921	6186	0.2514
All*	18.959	1.9603	3.2918	7445	2.55E-78

† Assumes unequal variance. Confidence Intervals are two tailed.

* Duplicates removed or none present

** Duplicates, Cytoplasmic, Nuclear removed

*** Only ER, Golgi, or both

Table 3.5C: Statistics comparing ASC of proteins in different locations in *M tuberculosis.*

	Mann Whitney U		Kolmogorov–Smirnov	
vs Cytoplasm	**Z**	**p value**	**Max Dif**	**p value**
Cytoplasmic Membrane	10.7167	8.49E-27	-0.229	6.74E-22
Unknown	8.9557	3.37E-19	0.1446	1.93E-13
Extracellular	12.371	3.75E-35	0.5584	8.58E-37
Cell Wall	2.4495	0.0143	0.4744	0.0365
Extracellular and Cell Wall	12.5622	3.41E-36	0.5546	3.44E-38
vs All				
Cytoplasm	2.0838	0.0372	-0.0488	0.0028
Cytoplasmic Membrane	11.761	6.20E-32	-0.2285	3.28E-24
Unknown	7.5395	4.72E-14	0.1152	3.32E-10
Extracellular	11.8675	1.75E-32	0.5126	5.74E-32
Cell Wall	2.2392	0.0251	0.4366	0.067
Extracellular and Cell Wall	12.0425	2.12E-33	0.5093	2.94E-33
vs Cytoplasmic Membrane				
Unknown	14.5359	7.18E-48	0.3347	3.10E-38
Extracellular	13.4196	4.64E-41	0.6219	1.81E-39
Cell Wall	2.9964	0.0027	0.4716	0.0397
Extracellular and Cell Wall	13.6098	3.50E-42	0.6174	1.19E-40

	Student T Test*				
vs Cytoplasm	**T value**	**95% CI**	**99.9% CI**	**Deg Free**	**p value**
Cytoplasmic Membrane	10.5911	1.9631	3.3035	754	1.57E-24
Unknown	8.9146	1.9613	3.2958	1841	1.15E-18
Extracellular	12.2829	1.9768	3.3604	142	4.42E-24
Cell Wall	2.0153	2.3646	5.4079	7	0.0837
Extracellular and Cell Wall	12.3553	1.9759	3.3566	150	1.24E-24

* Assumes unequal variance. Confidence intervals are two tailed

Mann Whitney U test results used in text instead of student t test

Most locations fail normality

Table 3.6: Statistics comparing protein ASCs in different sublocations in *E. coli*.

	Mann Whitney U Test		Kolmogorov–Smirnov	
vs Cytoplasm	**Z**	**p value**	**Max Dif**	**p value**
Integral Membrane Protein	27.6379	3.90E-168	-0.51	7.80E-140
Membrane anchored	0.649	0.5164	-0.1178	0.0287
Periplasmic w N-term mem anchor	0.3666	0.7139	-0.383	0.2696
Inner Membrane Lipoprotein	4.1532	3.28E-05	0.5465	0.0002
Membrane associated	1.7771	0.0756	0.287	0.3271
Membrane Lipoprotein*	1.1243	0.2609	0.2708	0.6113
OM ß-barrel protein	0.305	0.7604	-0.1226	0.3914
OM Lipoprotein	4.892	9.98E-07	0.2395	3.51E-05
Cell surface appendage	5.6057	2.07E-08	0.7376	1.40E-06
Secreted and released	2.2686	0.0233	0.3138	0.7407
Extracellular**	5.9832	2.19E-09	0.6739	3.57E-07
vs All				
Integral Membrane Protein	24.0127	2.05E-127	-0.4287	1.32E-106
Membrane anchored	2.1015	0.0356	0.1275	0.0127
Periplasmic w N-term mem anchor	0.0509	0.9594	-0.3453	0.3899
Inner Membrane Lipoprotein	4.2749	1.91E-05	0.5735	9.75E-05
Membrane associated	2.1025	0.0355	0.3283	0.1873
Membrane Lipoprotein*	1.3172	0.1878	0.273	0.6005
OM ß-barrel protein	0.7611	0.4466	0.0973	0.6821
OM Lipoprotein	5.89	3.86E-09	0.2483	1.36E-05
Cell surface appendage	5.5665	2.60E-08	0.7417	1.17E-06
Secreted and released	2.3687	0.0179	0.357	0.583488
Extracellular**	6.0012	1.96E-09	0.663	5.70E-07
vs Integral Membrane Protein				
Membrane anchored	10.8649	1.69E-27	0.4402	5.02E-23
Periplasmic w N-term mem anchor	2.6038	0.0092	0.3674	0.3187
Inner Membrane Lipoprotein	6.0709	1.27E-09	0.8172	4.64E-09
Membrane associated	4.6997	2.61E-06	0.7301	1.85E-05
Membrane Lipoprotein*	2.9816	0.0029	0.4839	0.0514
Periplasmic	20.1056	6.59E-90	0.6431	2.86E-86
OM ß-barrel protein	7.3597	1.84E-13	0.5222	1.24E-12
OM Lipoprotein	12.4994	5.22E-37	0.6116	5.52E-29
Cell surface appendage	5.9058	3.51E-09	0.8948	2.19E-09
Secreted and released	3.2553	0.0011	0.6689	0.0296
Extracellular**	6.7169	1.86E-11	0.8564	3.58E-11

Table 3.6: Statistics comparing protein ASCs in different sublocations in *E. coli* (continued).

	Mann Whitney U Test		Kolmogorov–Smirnov	
vs OM Lipoprotein	**Z**	**p value**	**Max Dif**	**p value**
Periplasmic	1.579	0.114	-0.125	0.183
Inner Membrane Lipoprotein	1.695	0.090	0.347	0.082
Membrane Lipoprotein*	0.197	0.844	-0.313	0.467
vs Periplasmic				
Periplasmic w N-term mem anchor	1.171	0.242	-0.496	0.074
Inner Membrane Lipoprotein	2.841	0.005	0.430	0.009
Membrane Lipoprotein*	0.205	0.837	-0.286	0.551

Student T Test with Unequal Variance. CI 2 tailed

vs Cytoplasm	**T value**	**95% CI**	**99.9% CI**	**Deg Free**	**p value**
Integral Membrane Protein	28.897	1.962	3.300	1072	3.01E-136
Membrane anchored	0.359	1.975	3.351	164	0.7198
Periplasmic w N-term mem anchor	0.373	2.571	6.869	5	0.7244
Inner Membrane Lipoprotein	5.652	2.160	4.221	13	7.91E-05
Membrane associated	2.674	2.262	4.781	9	0.0255
Membrane Lipoprotein*	0.548	2.447	5.959	6	0.6038
OM ß-barrel protein	0.281	2.006	3.484	53	0.7799
OM Lipoprotein	5.249	1.984	3.391	100	8.61E-07
Cell surface appendage	9.614	2.201	4.437	11	1.09E-06
Secreted and released	2.257	3.182	12.924	3	0.1092
Extracellular**	8.461	2.132	4.073	15	4.29E-07
vs OM Lipoprotein					
Periplasmic	1.839	1.977	3.360	143	0.068
Inner Membrane Lipoprotein	2.109	2.064	3.745	24	0.0456
Membrane Lipoprotein*	0.625	2.365	5.408	7	0.5516
vs Periplasmic					
Periplasmic w N-term mem anchor	1.419	2.571	6.869	5	0.2152
Inner Membrane Lipoprotein	3.550	2.132	4.073	15	0.0029
Membrane Lipoprotein*	0.183	2.447	5.959	6	0.8612

*Difficulty determining which membrane for seven "Membrane Lipoproteins" VisC, YfbK, YiaF, YifL, YihN, YjbH, and YraM. These proteins included in "Inner Membrane" for Fig. 3.1A, 3.3B, 3.6,3.8, 3.12, 3.13 and Tables 3.3-6.

**Subdivided into "Secreted and released" and "Cell surface appendage.

Table 3.7: Amino acid percent of proteins in different compartments of *E. coli*.

	Mean ASC	%A	%R	%N	%D	%C	%E	%Q	%G	%H	%I	%L	%K	%M	%F	%P	%S	%T	%W	%Y	%V	Total aa	Number
Overall Mean	23.555	9.37	5.6	3.87	5.03	1.29	5.83	4.42	7.12	2.33	6.1	10.6	4.7	2.96	3.91	4.3	5.81	5.36	1.52	2.79	7.09	312.46	4333
Overall Median	23.398	9.23	5.45	3.68	5.15	1.08	5.9	4.24	7.06	2.17	5.87	10.4	4.27	2.78	3.62	4.24	5.61	5.18	1.32	2.61	6.99	275.00	4333
Overall STDEVP	1.5038	2.9	2.43	1.79	2.04	1.24	2.59	2.06	2.5	1.5	2.27	3.2	2.44	1.33	2.06	1.75	2.07	1.95	1.2	1.47	2.28	209.22	4333
Extracellular*	20.652	**12**	**2.5**	**6.9**	5.81	**0.6**	**3**	**6.5**	8.04	**0.7**	4.83	8.45	4.43	**1.7**	**2.6**	**2.7**	**9.1**	**9.9**	**0.6**	2.42	7.09	280.44	16
OM	22.852	9.04	4.59	**5.5**	5.66	1.14	**4.2**	5.16	8.2	**1.4**	4.97	8.64	4.88	2.79	3.23	4.16	**7.4**	**6.8**	1.51	**4**	6.69	334.95	149
Periplasm	22.854	10.1	4.24	**5.1**	5.75	1.07	4.69	5.1	7.02	**1.7**	5.06	8.84	**6.2**	2.64	3.44	4.58	6.65	6.44	1.44	3.09	6.96	309.42	337
IM	24.653	9.93	4.56	3.22	**3.3**	0.99	**3.6**	3.54	7.8	**1.7**	7.48	13.2	3.56	3.53	**5.3**	4.13	6.04	5.25	**2.2**	2.72	8.04	357.55	972
Cytoplasm	23.317	9.1	6.18	3.86	5.5	1.44	6.82	4.59	6.84	2.68	5.82	10	4.9	2.82	3.54	4.34	5.53	5.17	1.32	2.72	6.8	296.50	2859
Sublocations																							
Cell surface appendage	20.474	**12**	**2.2**	**8**	5.26	**0.5**	**2.3**	**6.3**	**9.3**	**0.7**	4.91	8.2	**3.5**	**1.6**	**2.7**	**2.2**	**9.4**	**11**	**0.6**	2.67	7.11	299.17	12
Secreted and released	21.187	**12**	**3.5**	3.78	**7.4**	**0.8**	5.27	**6.9**	**4.4**	**0.8**	4.61	9.19	**7.2**	**2**	**2.1**	4.43	**8.4**	**7.9**	**0.9**	**1.7**	7.01	**224.25**	4
OM Lipoprotein	22.564	9.58	4.63	4.78	5.07	1.58	**4.1**	**5.6**	7.47	**1.4**	5.2	9.26	5.3	3.26	**2.8**	4.62	**7.3**	6.65	1.2	3.08	7.11	**216.46**	96
OM ß-barrel protein	23.374	8.05	4.52	**6.9**	**6.7**	**0.4**	**4.3**	4.34	**9.5**	**1.5**	**4.6**	**7.5**	4.12	**2**	4.09	3.31	**7.5**	**7.1**	**2.1**	**5.6**	5.94	**549.57**	53
Integral Membrane Protein	25.01	9.98	4.29	3	**2.8**	0.99	**3.1**	**3.1**	8.04	**1.6**	**7.9**	**14**	**3.1**	**3.7**	**5.8**	4.05	6.09	5.19	**2.3**	2.74	8.21	369.58	777
Membrane anchored	23.374	9.69	5.9	3.98	4.95	**0.9**	5.45	5.09	6.77	1.79	5.88	10.8	5.23	2.82	3.56	4.28	5.81	5.49	1.67	2.52	7.39	284.58	158
Peri. w N-term anchor	23.48	9.06	5.22	4.06	5.74	**1.8**	6.53	5.48	7	**3.2**	5.23	8.29	**6.3**	3.21	2.95	**5.8**	5.05	4.2	1.73	3.07	6.06	271.33	6
IM Lipoprotein	21.987	9.47	**4**	4.76	**6.4**	**0.8**	6.14	**6**	7.11	**1.2**	**4.4**	9.65	**6.6**	**1.9**	**2.6**	4.38	6.44	6.32	**0.8**	3.09	7.85	**485.50**	14
Membrane associated	22.708	11.1	5.65	4.17	5.03	1.02	6.13	5.22	7.08	2.19	4.92	9.79	5.17	3.25	**2.7**	4.6	5.65	5.06	1.46	2.46	7.36	390.50	10
Membrane Lipoprotein	22.97	9.77	**3.9**	4.62	5.27	**0.9**	**3.7**	**6.8**	7.54	**1.2**	**4.3**	10.2	5.31	2.4	4.46	**5.9**	6.24	6.28	1.27	3.35	6.64	**439.29**	7
Cost Savings (ATP)		15.7	0.06	12.7	14.7	2.66	12.1	11.1	15.7	10.9	4.94	0.06	2.94	6.94	24.6	7.06	15.7	8.66	46.9	22.6	4.06		

Notes: **Bold** indicates >25% change in amino acid composition from the overall mean.
Underlined indicates >50% change in amino acid composition from the overall mean.

Amino acid percents are on a per protein basis. The percent amino acid composition of each protein in that location were averaged; they are unweighted for chain length. See Table 3.22A for weighted percents (% of all amino acids) of proteins in nonmembrane locations of *E. coli.*
* We also looked at Extracellular proteins cost removing the SEC signal peptide when present (except FlgD for which a SEC signal is predicted; however, FlgD is secreted by the flagella system(112)). The average changed to 20.433 ATPs per amino acid.

Table 3.8: Percent of amino acids in curli specific genes.

Standard Competition Ranking

	% Gly	Rank Gly	% Ser	Rank Ser	%GS	Rank GS	% Asn	Rank Asn	% Gln	Rank Gln	%NQ	Rank NQ	%GSNQ	Rank GSNQ
Mean	7.12		5.81		12.93		3.87		4.42		8.29		21.22	
Median	7.06		5.61		12.84		3.68		4.24		8.16		21.01	
STDEVP	2.50		2.07		3.09		1.79		2.06		2.77		3.84	
CsgA	**19.21**	6	**9.27***	227	**28.48**	6	**10.60**	**22**	7.28**	354	**17.89**	**18**	**46.36**	**2**
CsgB	9.93	492	8.61	361	**18.54**	165	**8.61**	**87**	**9.93**	**55**	**18.54**	**12**	**37.08**	**8**
CsgC	3.64	4028	**14.55**	12	**18.18**	198	3.64	2218	**10.91**	**33**	**14.55**	**102**	**32.73**	**47**
CsgD	4.17	3884	8.33	420	12.50	2377	**8.80**	**75**	4.17	2214	12.96	223	25.46	438
CsgE	4.65	3693	5.43	2358	10.08	3669	4.65	1160	3.88	2483	8.53	1907	18.61	3415
CsgF	8.70	1077	7.97	536	16.67	419	**11.59**	**13**	**7.97**	**218**	**19.57**	**7**	**36.23**	**13**
CsgG	6.86	2323	7.22	846	14.08	1405	5.78	531	5.78	940	11.55	460	25.63	411
>P_{95}	11.11		9.33		18.04		7.14		7.97		13.03		27.78	
>P_{90}	10.11		8.33		16.60		6.06		7.01		11.64		25.49	

* 94.7th Percentile

** 91.8th Percentile

Note: Bold ≥95th percentile

Fractional Ranking

	Rank G	Rank S	Rank GS	Rank N	Rank Q	Rank NQ	Rank GSNQ
CsgA	6	227	6	22	354	18	2
CsgB	492	361	165	87	55	12	8
CsgC	4029.5(4)	12	201.5(8)	2223(11)	34(3)	102	47
CsgD	3889.5(12)	427.5(16)	2392(31)	75.5(2)	2224(21)	224(3)	438
CsgE	3694.5(4)	2359.5(4)	3669.5(2)	1168.5(18)	2484.5(4)	1909.5(6)	3417.5(6)
CsgF	1080.5(8)	536	423.5(10)	13	218	7	13
CsgG	2323	847(3)	1405.5(2)	531	940	460	411

Note: Numbers in parenthesis are number of proteins with same % amino acid.

Table 3.9: Percentiles of flagella, curli, and Type 1 pili proteins.

Percentile	99.9999	99.999	99.99	99.9	99	95	90	80	70
Avg E	18.017	18.036	18.227	18.916	20.430	21.427	21.894	22.404	22.785
Savings	5.538	5.519	5.328	4.639	3.125	2.128	1.661	1.151	0.770

Flagella

Percentile	Avg Cost	Rank	Name	Length	Location	Role	Savings per aa
99.885	19.1548	6	FliC	498	Cell surface appendage	Structural	4.4001
99.793	19.3949	10	FlgM	97	Secreted&released(68, 77)	Regulatory	4.1601
99.677	19.6541	15	FlgD	231	Cell surface appendage	Assembly	3.9008
99.608	19.7397	18	FliK	375	Secreted&released(95)	Assembly	3.8152
99.539	19.8705	21	FlgF	251	Periplasmic	Structural	3.6844
99.493	19.9291	23	FliD	468	Cell surface appendage	Structural	3.6258
99.308	20.1981	31	FlgI	365	Periplasmic	Structural	3.3568
98.985	20.4642	45	FlgC	134	Periplasmic	Structural	3.0907
98.962	20.4713	46	FlgK	547	Cell surface appendage	Structural	3.0836
98.846	20.5953	51	FlgE	402	Cell surface appendage	Structural	2.9596
98.639	20.6565	60	FlgG	260	Periplasmic	Structural	2.8984
98.5	20.6683	66	FliE	104	Cytoplasmic *	Structural	2.8866
98.015	20.8473	87	FlgL	317	Cell surface appendage	Structural	2.7076
97.831	20.9435	95	FlgB	138	Periplasmic	Structural	2.6114
96.03	21.2669	173	FliO	121	Membrane anchored	Export	2.2880
95.868	21.2995	180	FliF	552	IM Protein	Structural	2.2555
95.476	21.3480	197	FlgA	219	Periplasmic	Chaperone	2.2070
95.407	21.3540	200	FliN	137	Cytoplasmic	Structural	2.2009
95.361	21.3680	202	FliH	228	Cytoplasmic	Export	2.1869
94.368	21.5162	245	FliS	136	Cytoplasmic	Chaperone	2.0387
93.883	21.5544	266	FliG	331	Membrane associated	Structural	2.0005
93.583	21.5961	279	FlgH	232	OM Lipoprotein	Structural	1.9588
87.166	22.0514	557	FliI	457	Cytoplasmic	Export	1.5035
83.796	22.2189	703	FlgJ	313	Periplasmic(102)	Assembly	1.3361
82.018	22.3130	780	MotB	308	Membrane anchored	Structural	1.2419
81.072	22.3541	821	FliY	266	Periplasmic	?	1.2008
79.225	22.4391	901	FlgN	138	Cytoplasmic	Chaperone	1.1158
78.048	22.4958	952	FliA	239	Cytoplasmic	Regulatory	1.0591
70.868	22.7515	1263	FlhE	130	Periplasmic	?	0.8034
66.898	22.8840	1435	FlhA	692	IM Protein	Export	0.6709
66.205	22.9020	1465	FliJ	147	Cytoplasmic	Export	0.6529
59.65	23.0984	1749	FliT	121	Cytoplasmic	Chaperone	0.4566
59.28	23.1126	1765	FliM	334	Cytoplasmic	Structural	0.4423
57.11	23.1824	1859	MotA	295	IM Protein	Structural	0.3725
53.948	23.2740	1996	FlhD	119	Cytoplasmic	Regulatory	0.2810
45.568	23.5490	2359	FlhC	192	Cytoplasmic	Regulatory	0.0060
41.736	23.6766	2525	FliL	154	OM Lipoprotein	?	-0.1217
40.859	23.7039	2563	FlhB	382	IM Protein	Export	-0.1490
17.544	24.7843	3573	FliQ	89	IM Protein	Export	-1.2294
17.314	24.8102	3583	FliP	245	IM Protein	Export	-1.2553
13.805	25.0825	3735	FliZ	183	Cytoplasmic	Regulatory	-1.5276
11.866	25.2406	3819	FliR	261	IM Protein	Export	-1.6857
77.64364	21.90035	969.5		266.8571		**AVERAGE**	1.6546
93.733	21.57525	272.5		242		**MEDIAN**	1.5876
27.17407	1.598852	1177.18		143.2881		**STDEV**	1.5989

Table 3.9: Percentiles of flagella, curli, and Type 1 pili proteins (continued).

Curli

Percentile	Avg Cost	Rank	Name	Length	Location	Role	Savings per aa
99.816	19.3887	9	CsgA	151	Cell surface appendage	Major	4.1662
98.685	20.6517	58	CsgB	151	Cell surface appendage	Minor	2.9033
96.538	21.2036	151	CsgC	110	Periplasmic	?	2.3513
24.146	24.3593	3287	CsgD	216	Cytoplasmic	Regulatory	-0.8044
11.981	25.2326	3814	CsgE	129	Periplasmic	Chaperone	-1.6776
88.874	21.9725	483	CsgF	138	Periplasmic**	Chaperone	1.5824
68.006	22.8498	1387	CsgG	277	OM Lipoprotein	export	0.7051
69.72086	22.23688	1312.714		167.4286		**AVERAGE**	1.3180
88.874	21.97246	483		151		**MEDIAN**	1.5824
37.06872	2.066261	1605.809		58.5174		**STDEV**	2.0663

Type 1 pili

Percentile	Avg Cost	Rank	Name	Length	Location	Role	Savings per aa
99.77	19.5247	11	FimA	182	Cell surface appendage	Major	4.0302
10.942	25.3265	3859	FimB	200	Cytoplasmic	regulatory	-1.7716
70.291	22.7739	1288	FimC	241	Periplasmic	Chaperone	0.7811
72.115	22.7075	1209	FimD	878	OM ß-barrel protein	export	0.8474
30.772	24.0742	3000	FimE	198	Cytoplasmic	Regulatory	-0.5193
84.419	22.1852	676	FimF	176	Cell surface appendage	Adaptor	1.3697
97.554	20.9916	107	FimG	167	Cell surface appendage	Adaptor	2.5633
82.503	22.2920	759	FimH	300	Cell surface appendage	Tip	1.2629
57.941	23.1536	1823	FimI	179	Periplasmic	?	0.4013
29.156	24.1405	3070	FimZ	210	Cytoplasmic	Regulatory	-0.5856
63.5463	22.71698	1580.2		273.1		**AVERAGE**	0.7487
71.203	22.74069	1248.5		199		**MEDIAN**	0.8321
30.61971	1.650011	1326.442		216.1437		**STDEV**	1.6500

*FliE interacts with FlgB and is likely periplasmic (87, 96).

**CsgF has been shown to be cell surface exposed (108).

Table 3.10: Cost savings and amino acid composition or extracellular proteins in *E. coli*.

	Cost vs Average aa		15.66	0.06	12.66	14.66	2.66	12.06	11.06	15.66	-10.9	-4.94	0.06	-2.94	-6.94	-24.6	7.06	15.66	8.66	-46.9	-22.6	4.06
		ATP/aa	**%A**	**%R**	**%N**	**%D**	**%C**	**%E**	**%Q**	**%G**	**%H**	**%I**	**%L**	**%K**	**%M**	**%F**	**%P**	**%S**	**%T**	**%W**	**%Y**	**%V**
Overall	**MEAN**	23.555	9.37	5.6	3.87	5.03	1.29	5.83	4.42	7.12	2.33	6.1	10.59	4.7	2.96	3.91	4.3	5.81	5.36	1.52	2.79	7.09
	MEDIAN	23.398	9.23	5.45	3.68	5.15	1.08	5.9	4.24	7.06	2.17	5.87	10.37	4.27	2.78	3.62	4.24	5.61	5.18	1.32	2.61	6.99
	STDEVP	1.504	2.9	2.43	1.79	2.04	1.24	2.59	2.06	2.5	1.5	2.27	3.2	2.44	1.33	2.06	1.75	2.07	1.95	1.2	1.47	2.28
Extracellular	**MEAN**	20.652	12.11	2.535	6.911	5.805	0.566	3.0305	6.474	8.0422	0.698	4.832	8.45	4.427	1.714	2.558	2.744	9.112	9.857	0.629	2.422	7.087

		ATP/aa	**A**	**R**	**N**	**D**	**C**	**E**	**Q**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
ATP savings per amino Acid	**FliC**	19.155	11.85	2.21	9.64	7.83	0	2.81	5.42	8.84	0	5.62	7.43	5.02	0.8	1	1.2	8.63	13.05	0	2.01	6.63
	CsgA	19.389	10.6	1.32	10.6	5.3	0	1.32	7.28	19.21	1.99	3.31	5.96	2.65	1.32	2.65	1.32	9.27	5.96	0.66	2.65	6.62
	FimA	19.525	19.78	1.65	6.04	4.4	1.1	1.65	5.49	8.79	1.1	3.3	8.24	2.75	0.55	3.85	1.1	7.69	12.09	0	1.1	9.34
	FlgM	19.395	11.34	5.15	5.15	7.22	0	5.15	6.19	3.09	0	5.15	9.28	6.19	3.09	0	5.15	11.34	11.34	0	0	5.15
	FlgD	19.654	7.79	1.3	6.49	4.76	0	3.46	6.49	10.82	0.87	6.06	9.52	1.73	1.3	1.73	3.03	10.39	15.58	0.43	0.87	7.36
	FliK	19.74	12.53	1.87	2.4	6.93	0	5.07	9.6	4.53	1.87	3.73	11.2	3.2	1.07	1.87	8.8	8.27	10.67	0.27	0	6.13
	FliD	19.929	11.54	1.5	6.62	7.26	0	1.5	4.91	6.41	0.21	6.2	9.62	5.56	1.28	1.71	0.64	14.32	14.1	0.43	2.14	4.06
	FlgK	20.471	12.07	2.93	8.59	6.76	0	2.93	8.04	7.86	0.18	4.94	8.59	3.29	1.65	3.29	1.65	8.78	8.78	0.37	2.74	6.58
	FlgE	20.595	10.2	1.49	11.94	5.22	0	1.99	6.22	9.7	0.25	4.23	7.46	2.99	2.24	4.23	2.74	8.71	10.7	0.5	2.49	6.72
	CsgB	20.652	15.23	3.31	8.61	3.31	0	1.99	9.93	9.93	0	7.95	6.62	5.3	3.31	1.99	0.66	8.61	5.3	0	3.97	3.97
	FlgL	20.847	11.67	3.47	5.36	6.94	0	5.05	8.52	5.99	0.32	4.1	8.83	4.73	3.47	2.52	1.58	10.41	8.2	0.63	2.84	5.36
	FimG	20.992	13.77	1.8	4.79	5.39	1.8	1.2	5.99	7.78	1.2	4.19	7.78	3.59	1.2	1.8	1.8	9.58	12.57	1.2	3.59	8.98

		Savings/aa	**A**	**R**	**N**	**D**	**C**	**E**	**Q**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
Percent of total savings for each amino Acid	**FliC**	4.4	8.8	-0.05	**16.59**	9.33	-0.78	-8.26	2.52	6.1	5.79	0.53	-0.04	-0.22	3.4	**16.3**	-4.97	**10.06**	**15.14**	**16.22**	4.02	-0.43
	CsgA	4.17	4.59	-0.06	**20.43**	0.94	-0.83	-13.03	7.61	**45.42**	0.9	3.3	-0.07	1.45	2.72	7.48	-5.04	**13.02**	1.25	9.67	0.76	-0.45
	FimA	4.03	**40.43**	-0.06	6.82	-2.31	-0.13	-12.5	2.95	6.49	3.34	3.43	-0.03	1.42	4.15	0.42	-5.61	7.32	**14.46**	**17.71**	9.49	2.27
	FlgM	4.16	7.4	-0.01	3.9	7.7	-0.83	-1.95	4.7	-15.17	6.13	1.12	-0.02	-1.05	-0.23	**23.18**	1.45	**20.82**	**12.45**	**17.16**	15.18	-1.89
	FlgD	3.9	-6.35	-0.07	8.5	-1.01	-0.88	-7.31	5.88	14.86	4.1	0.05	-0.02	2.24	2.95	**13.79**	-2.3	**18.39**	**22.7**	**13.09**	11.16	0.28
	FliK	3.82	**12.96**	-0.06	-4.89	7.31	-0.9	-2.4	**15.02**	-10.63	1.33	3.06	0.01	1.15	3.44	**13.22**	8.33	**10.09**	12.04	**15.43**	16.55	-1.02
	FliD	3.63	9.34	-0.07	9.6	9.03	-0.95	-14.41	1.51	-3.07	6.38	-0.13	-0.02	-0.7	3.21	**14.98**	-7.13	**36.75**	**20.88**	**14.16**	4.07	-3.39
	FlgK	3.08	**13.67**	-0.05	**19.37**	8.24	-1.12	-11.35	**13**	3.75	7.62	1.86	-0.04	1.34	2.95	4.98	-6.08	**15.07**	9.59	**17.59**	0.34	-0.67
	FlgE	2.96	4.36	-0.08	**34.5**	0.95	-1.16	-15.63	6.73	**13.65**	7.69	3.12	-0.06	1.7	1.69	-2.62	-3.73	**15.34**	15.61	**16.23**	2.3	-0.51
	CsgB	2.9	**31.59**	-0.05	**20.65**	-8.69	-1.19	-15.95	**21.01**	**15.17**	8.78	-3.15	-0.08	-0.61	-0.85	**16.36**	-8.85	**15.11**	-0.19	**24.59**	-9.24	-4.36
	FlgL	2.71	**13.29**	-0.05	6.96	**10.33**	-1.27	-3.47	**16.74**	-6.53	8.14	3.64	-0.04	-0.04	-1.31	**12.65**	-7.1	**26.62**	9.09	**15.43**	-0.42	-2.59
	FimG	2.56	**26.87**	-0.09	4.53	2.04	0.52	-21.78	6.77	4.05	4.83	3.67	-0.07	1.27	4.76	**20.36**	-6.9	**23.05**	**24.37**	5.92	-7.1	3
	Average	>1.5ATPs	**13.91**	-0.06	**12.25**	3.65	-0.79	-10.67	8.7	6.17	5.42	1.71	-0.04	0.66	2.24	**11.76**	-3.99	**17.64**	**13.12**	**15.27**	3.92	-0.81
		ALL	**17.6**	-0.09	**14.79**	6.47	-0.57	-11.87	8.58	-1.23	9.3	2.78	-0.06	-1.79	4.54	**10.88**	-5.89	**19.54**	**14.1**	**15.21**	-3.3	1.1

Bold: Greater than 10% of total savings due to composition of that amino acid

Percent of Total Cost Savings for *E. coli* Major Extracellular Fiber Subunits

	Excluding Aromatics				**Aromatics**		
	Top 2	**Savings**	**>10% Savings**	**Savings**	**FW**	**FYW**	**FYWH**
CsgA	GN	65.85	GNS	78.86	17.16	17.91	18.81
FimA	AT	54.89	AT	54.89	18.13	27.62	30.96
FliC	NT	31.72	NTS	41.78	32.52	36.54	42.33

Table 3.11: Flagella protein location vs average synthetic cost (ASC).

	Student T test (assuming heteroscedastic)						Two tailed		One tailed	
	Num.	Mean	Variance	Degrees Freedom	T crit. value	Pooled Variance	P value	Critical Value*	P value	Critical Value*
vs Cytoplasm	**13**	**22.532**	**1.364**							
IM	10	23.104	2.264	17	0.994	1.750	0.33405	2.110	0.16703	1.740
Periplasm	9	21.201	1.059	19	2.821	1.242	**0.01091**	2.093	**0.00546**	1.729
OM	2	22.636	2.164	1	0.096	1.426	0.93899	12.706	0.46949	6.314
Extracellular	8	19.973	0.366	19	6.590	0.997	**2.624E-06**	2.093	**1.312E-06**	1.729
vs Cyto & IM	**23**	**22.780**	**1.754**							
Cytoplasm	13	22.532	1.364	28	0.584	1.617	0.56359	2.048	0.28179	1.701
IM	10	23.104	2.264	15	0.588	1.902	0.56529	2.131	0.28265	1.753
Periplasm	9	21.201	1.059	19	3.587	1.569	**0.00196**	2.093	**0.00098**	1.729
OM	2	22.636	2.164	1	0.134	1.772	0.91528	12.706	0.45764	6.314
Extracellular	8	19.973	0.366	26	8.036	1.419	**1.626E-08**	2.056	**8.128E-09**	1.706
vs all flagella	**42**	**21.900**	**2.556**							
Cytoplasm	13	22.532	1.364	27	1.550	2.286	0.13273	2.052	0.06636	1.703
IM	10	23.104	2.264	14	2.246	2.504	**0.04139**	2.145	**0.02069**	1.761
Periplasm	9	21.201	1.059	18	1.656	2.312	0.11504	2.101	0.05752	1.734
OM	2	22.636	2.164	1	0.688	2.547	0.61616	12.706	0.30808	6.314
Extracellular	8	19.973	0.366	29	5.902	2.237	**2.085E-06**	2.045	**1.043E-06**	1.699

BOLD = Statistically Significant. P <0.05

* 5% Confidence Interval

Table 3.12: Flagella subunit ASC/economic rank versus stoichiometry & chain length.

With FliC included (26 total)	
Comparing Avg AA Cost & Number of Subunits	
Linear regression	
y = 17322 - 761.2x R= 0.31913	
Spearman R	-0.25497
Rank Difference Squares Sum	3662
t-test value for hypothesis r = 0	-1.29179
p-level	0.208733
Kendall Tau	-0.12521
Inversions Count	177
Z	0.896972
p-level	0.369734
Gamma	-0.12739
Pearson Correlation Coefficient	-0.31913

Without FliC (25 total)	
Comparing Avg AA Cost & Number of Subunits	
Linear regression	
y = 84.851 - 2.8219x R= 0.13516	
Spearman R	-0.16159
Rank Difference Squares Sum	3012
t-test value for hypothesis r = 0	-0.78528
p-level	0.440307
Kendall Tau	-0.05094
Inversions Count	152
Z	0.356929
p-level	0.721145
Gamma	-0.0519
Pearson Correlation Coefficient	-0.13516

Comparing Economic Rank & Number of Subunits	
Linear regression	
y = 1178.2 - 0.4517x R= 0.14222	
Spearman R	-0.25497
Rank Difference Squares Sum	3662
t-test value for hypothesis r = 0	-1.29179
p-level	0.208733
Kendall Tau	-0.12521
Inversions Count	177
Z	0.896972
p-level	0.369734
Gamma	-0.12739
Pearson Correlation Coefficient	-0.14222

Comparing Economic Rank & Number of Subunits	
Linear regression	
y = 28.262 - 0.0055992x R= 0.21034	
Spearman R	-0.16159
Rank Difference Squares Sum	3012
t-test value for hypothesis r = 0	-0.78528
p-level	0.440307
Kendall Tau	-0.05094
Inversions Count	152
Z	0.356929
p-level	0.721145
Gamma	-0.0519
Pearson Correlation Coefficient	-0.21034

Comparing Avg AA cost & Chain Length (all 42 proteins)	
Linear regression	
Pearson Correlation Coefficient	-0.21355
Spearman R	-0.18857
Rank Difference Squares Sum	14667
t-test value for hypothesis r = 0	-1.21443
p-level	0.231703
Kendall Tau	-0.12674
Inversions Count	484
Z	1.18265
p-level	0.236948
Gamma	-0.12689
Pearson Correlation Coefficient	-0.21355

Table 3.13: Comparison of protein abundance and average synthetic cost (ASC) or mean amino acid mass in *E. coli.*

						Number per cell vs. Avg Cost				Number per cell vs. Avg Mass			
						Spearman		Pearson		Spearman		Pearson	
	Location	Proteins	Avg. # per cell	Avg Cost	Avg Mass	R	P value	R	P value	R	P value	R	P value
	Overall	1270	1362.67	22.98	110.58	**-0.1263**	**6.31E-06**	**-0.0717**	**0.0105**	**-0.0750**	**0.0075**	-0.0512	0.0684
Whole Cell Lysates	Overall without Ribosomal Proteins	1217	529.21	22.98	110.55	**-0.1349**	**2.33E-06**	**-0.0691**	**0.0159**	**-0.0995**	**0.0005**	**-0.0578**	**0.0437**
	Cytoplasmic	1022	1625.28	22.95	110.81	**-0.1078**	**0.0006**	**-0.0902**	**0.0039**	**-0.0773**	**0.0135**	**-0.0632**	**0.0435**
	Cytoplasmic without Ribosomal Proteins	969	592.86	22.95	110.78	**-0.1187**	**0.0002**	**-0.0978**	**0.0023**	**-0.1019**	**0.0015**	**-0.0709**	**0.0273**
Masuda et al 2009 (89)	Inner membrane	134	348.65	23.51	109.80	0.0038	0.9656	0.1262	0.1461	-0.1584	0.0676	0.0066	0.9392
	Integral Membrane Protein	81	371.32	24.27	109.84	0.1829	0.1022	**0.2394**	**0.0313**	-0.1156	0.3042	**0.0061**	**0.9570**
	Inner membrane lipoprotein	3	162.75	21.25	107.65	-	-	-0.8752	0.3215	-	-	-0.8093	0.3997
	Membrane anchored	43	337.14	22.44	110.14	-0.0854	0.5860	-0.0880	0.5747	-0.1093	0.4854	0.0015	0.9926
	Membrane associated	3	213.32	22.65	108.07	-	-	0.5438	0.6340	-	-	-0.0805	0.9487
	Membrane Lipoprotein	3	328.33	21.77	107.64	-	-	0.9649	0.1691	-	-	0.4960	0.6696
	Periplasmic w N-term memb anchor	1	32.03	22.16	111.15	-	-	-	-	-	-	-	-
	Periplasmic	63	147.88	22.76	109.65	**-0.3890**	**0.0016**	**-0.3009**	**0.0166**	**-0.4100**	**0.0008**	**-0.2913**	**0.0205**
	Outer membrane	51	265.22	22.63	109.27	-0.2168	0.1264	0.0290	0.8397	-0.2627	0.0626	-0.0437	0.7609
	Outer membrane Lipoprotein	31	140.62	22.11	108.35	**-0.5761**	**0.0007**	**-0.4159**	**0.0200**	**-0.4994**	**0.0042**	**-0.3591**	**0.0472**
	Outer Membrane ß-barrel protein	20	458.36	23.43	110.68	0.1579	0.5061	-0.0841	0.7244	-0.0692	0.7720	-0.2258	0.3385
	Overall	961	1385.17	23.18	110.53	**-0.2868**	**1.08E-19**	**-0.0741**	**0.0216**	-0.0196	0.5440	-0.0506	0.1171
Membrane Enriched Fraction	Cytoplasmic	579	2265.89	22.95	110.87	**-0.1250**	**0.0026**	**-0.1035**	**0.0127**	**-0.1381**	**0.0009**	-0.0793	0.0566
	Inner membrane	294	7.47	23.80	110.22	-0.1115	0.0563	**-0.1212**	**0.0378**	-0.2400	3.20E-05	-0.0904	0.1221
	Integral Membrane Protein	208	1.54	24.30	110.23	**0.1511**	**0.0293**	**0.1558**	**0.0246**	**-0.2283**	**0.0009**	**-0.1823**	**0.0084**
Masuda et al 2009 (89)	Inner membrane lipoprotein	3	162.75	21.25	107.65	-	-	-0.8752	0.3215	-	-	-0.8093	0.3997
	Membrane anchored	74	9.31	22.67	110.44	-0.1596	0.1744	0.0930	0.4305	-0.1601	0.1731	-0.0227	0.8478
	Membrane associated	3	213.32	22.65	108.07	-	-	0.5438	0.6340	-	-	-0.0805	0.9487
	Membrane Lipoprotein	3	17.67	21.77	107.64	-	-	**-0.9998**	**0.0138**	-	-	-0.2296	0.8525
	Periplasmic w N-term memb anchor	3	1.58	23.12	111.58	-	-	0.7273	0.4816	-	-	0.9170	0.2612
	Periplasmic	33	174.87	22.66	109.78	**-0.4170**	**0.0158**	**-0.3849**	**0.0270**	**-0.3652**	**0.0367**	**-0.3454**	**0.0490**
	Outer membrane	54	207.98	22.57	109.14	0.0094	0.9461	0.0632	0.6496	0.0271	0.8457	0.0328	0.8136
	Outer membrane Lipoprotein	33	96.59	22.10	108.42	-0.2361	0.1860	**-0.3545**	**0.0429**	-0.1899	0.2898	-0.2224	0.2135
	Outer Membrane ß-barrel protein	21	383.00	23.29	110.26	0.2519	0.2706	-0.0215	0.9262	0.2286	0.3190	-0.0383	0.8689
	Cell Surface Appendage	1	1.45	19.52	99.41	-	-	-	-	-	-	-	-

BOLD: $P < 0.05$

Notes: Echobase sequence and location data were merged with abundance values of Masuda et al (89) using ECK numbers. GatA had two abundance values in whole cell lysates. InsH and yi81 have multiple Echobase number but only one ECK number; so ASC and Mass values were averaged.

Table 3.13: Comparison of protein abundance and average synthetic cost (ASC) or mean amino acid mass in *E. coli* (continued)

	Location	Proteins	Avg Number per cell	Avg Cost	Avg Mass	Number per cell vs. Avg Cost				Number per cell vs. Avg Mass			
						Spearman		Pearson		Spearman		Pearson	
						R	P value	R	P value	R	P value	R	P value
Cytosolic Fraction	Overall	1091	201883.19	22.95	110.54	**-0.1606**	**9.69E-08**	**0.0653**	**0.0309**	**-0.0696**	**0.0215**	0.0502	0.0973
	Overall without Ribosomal Proteins	1038	3685.60	22.95	110.50	**-0.1747**	**1.45E-08**	**-0.0844**	**0.0065**	**-0.1035**	**0.0008**	-0.0388	0.2113
	Cytoplasmic	874	251724.34	22.91	110.75	**-0.1248**	**0.0002**	**0.0849**	**0.0120**	**-0.0663**	**0.0499**	0.0530	0.1174
	Cytoplasmic without Ribosomal Proteins	821	4353.39	22.91	110.72	**-0.1440**	**3.44E-05**	**-0.0980**	**0.0049**	**-0.0995**	**0.0043**	-0.0494	0.1572
	Inner membrane	99	428.06	23.53	109.91	-0.0805	0.4285	-0.0532	0.6008	**-0.2200**	**0.0287**	**-0.2605**	**0.0092**
Ishihama et al 2008 (70)	Integral Membrane Protein	59	365.79	24.37	110.34	0.2052	0.1190	0.0922	0.4872	-0.0968	0.4657	-0.0782	0.5561
	Inner membrane lipoprotein	4	507.25	21.60	107.92	-	-	-0.3656	0.6344	-	-	-0.5773	0.4227
	Membrane anchored	29	457.27	22.29	109.69	-0.3169	0.0940	-0.1789	0.3532	-0.2479	0.1948	-0.2518	0.1876
	Membrane associated	4	541.25	23.05	109.46	-	-	-0.2223	0.7777	-	-	-0.3074	0.6926
	Membrane Lipoprotein	2	384.00	21.62	108.03	-	-	-	-	-	-	-	-
	Periplasmic w N-term mem anchor	1	147.00	23.44	111.77	-	-	-	-	-	-	-	-
	Periplasmic	77	1322.96	22.84	109.59	0.0642	0.0505	-0.1758	0.1262	-0.2120	0.0642	-0.1067	0.3557
	Outer membrane	41	2616.32	22.61	109.27	-0.2947	0.0615	-0.0260	0.8719	**-0.3497**	**0.0250**	-0.1445	0.3675
	Outer membrane Lipoprotein	21	1432.67	21.95	108.14	**-0.5651**	**0.0076**	**-0.4722**	**0.0307**	**-0.4462**	**0.0426**	-0.2605	0.2542
	Outer Membrane ß-barrel protein	20	3859.16	23.30	110.47	0.1173	0.6224	-0.0602	0.8010	-0.2165	0.3591	-0.3991	0.0813

Note: 12 proteins exluded (plasmid origins: P07676, P06717,Q00191,P22997,P62593,P27190,P02978,P00642,Q52280,Q57154 - Not in K-12: P15484,P11257)

BOLD: P < 0.05

Notes: Echobase sequence and location data were merged with the abundance values of Ishihama et al (70) using Uniprot numbers. We excluded 12 proteins that did not have Echobase information. Ten had plasmid origins: P07676, P06717, Q00191, P22997, P62593, P27190, P02978, P00642, Q52280, Q57154, and two were not found K-12: P15484 and P11257.

Table 3.14: Comparison of protein chain length and average synthetic cost (ASC) or mean amino acid mass in *E. coli.*

					Length vs. Avg Cost				Length vs. Avg Mass			
			Resp.		Pearson		Spearman		Pearson		Spearman	
Location	Number	Avg Length	Avg Cost	Avg Mass	R	P value	R	P value	R	P value	R	P value
Overall - *E. coli*	4333	312.46	23.55	111.01	**-0.0992**	**5.92E-11**	**-0.0503**	**0.0009**	**-0.1225**	**6.00E-16**	**-0.1738**	**0, t=-11.62**
Cytoplasmic	2859	296.50	23.32	111.61	**-0.1329**	**9.86E-13**	**-0.1263**	**1.23E-11**	**-0.1414**	**3.08E-14**	**-0.1949**	**0, t=-10.62**
Inner membrane	972	357.55	24.65	109.99	**-0.3051**	**0, t=-9.98**	**-0.2850**	**1.08E-19**	**-0.1115**	**0.0005**	**-0.1469**	**4.22E-06**
Integral Membrane Protein	777	369.58	25.01	109.78	**-0.4555**	**0, t=-14.25**	**-0.4232**	**0, t=-13.00**	**-0.0961**	**0.0074**	**-0.1275**	**0.0004**
Inner membrane lipoprotein	14	485.50	21.99	109.27	0.1662	0.5701	-0.2489	0.3908	0.0715	0.8080	-0.0419	0.8870
Membrane anchored	158	284.58	23.37	110.99	**-0.2592**	**0.0010**	**-0.2887**	**0.0002**	-0.1158	0.1472	-0.1588	0.0463
Membrane associated	10	390.50	22.71	109.97	0.5009	0.1403	0.4681	0.1725	0.0841	0.8172	0.0790	0.8282
Membrane Lipoprotein	7	439.29	22.97	109.51	0.0601	0.8982	0.2500	0.5887	0.5479	0.2030	0.6071	0.1482
Periplasmic w N-term memb anchor	6	271.33	23.48	111.99	-0.4745	0.3416	-0.2571	0.6228	-0.4790	0.3364	-0.3143	0.5441
Periplasmic	337	309.42	22.85	109.72	0.1030	0.0590	**0.1835**	**0.0007**	0.1189	0.0291	0.1389	0.0107
Outer membrane	149	334.95	22.85	109.62	-0.0217	0.7929	0.1089	0.1862	0.0678	0.4111	0.1858	0.0233
Outer membrane Lipoprotein	96	216.46	22.56	109.01	-0.0435	0.6739	0.0324	0.7543	0.0413	0.6892	0.1490	0.1474
Outer Membrane ß-barrel protein	53	549.57	23.37	110.73	**-0.4286**	**0.0014**	-0.3440	0.0117	-0.2493	0.0718	-0.1236	0.3779
Extracellular	16	280.44	20.65	104.90	-0.1676	0.5351	-0.1118	0.6800	0.0041	0.9880	-0.1472	0.5865
Yeast All	6732	449.90	26.59	113.20	**-0.1853**	**0, t=-15.47**	**-0.2173**	**0, t=-18.26**	**-0.0428**	**0.0004**	**-0.0952**	**4.83E-15**
Yeast Cytoplasmic	2845	498.02	25.86	112.86	0.0164	0.3824	0.0119	0.5261	0.0247	0.1872	-0.0078	0.6783
Yeast Extracellullar	54	479.54	24.88	107.16	-0.1672	0.2268	0.0081	0.9535	-0.0316	0.8203	0.0959	0.4905
Yeast Cell Wall	44	493.25	24.04	105.27	-0.0991	0.5220	0.0394	0.7998	0.0073	0.9627	0.2272	0.1381
***M. tuberculosis* All**	3927	339.23	22.15	107.14	**-0.1116**	**2.36E-12**	-0.0171	0.2848	**-0.1952**	**0, t=-12.47**	**-0.1437**	**1.08E-19**
***M. tuberculosis* Cytoplasmic**	2111	348.93	22.25	108.56	-0.0439	0.0438	-0.0154	0.4793	**-0.1451**	**2.09E-11**	**-0.1841**	**1.53E-17**
***M. tuberculosis* Extracellullar**	140	513.19	19.45	96.00	**-0.4957**	**4.73E-10**	**-0.4958**	**4.72E-10**	**-0.5585**	**7.51E-13**	**-0.5670**	**2.81E-13**
***M. tuberculosis* Cell Wall**	8	552.38	20.97	103.19	-0.4829	0.2255	-0.0952	0.8225	-0.4522	0.2606	-0.0952	0.8225
***P. syringae* All**	5608	329.66	23.05	110.08	**-0.0967**	**3.98E-13**	**-0.0595**	**8.16E-06**	**-0.0867**	**7.67E-11**	**-0.1284**	**0, t=-9.69**
***P. syringae* Cytoplasmic**	1973	317.70	22.96	110.98	**-0.1083**	**1.42E-06**	**-0.0913**	**4.92E-05**	**-0.1407**	**3.41E-10**	**-0.1934**	**4.34E-18**
***P. syringae* Extracellular**	84	433.73	21.58	107.30	-0.0855	0.4394	-0.1262	0.2525	-0.0902	0.4147	-0.1653	0.1330
***P. syringae* HOPs & helpers**	51	437.31	21.48	107.68	-0.0635	0.6577	-0.2346	0.0975	-0.0445	0.7567	-0.2175	0.1253

BOLD: P < 0.01

Table 3.14: Comparison of protein chain length and average synthetic cost (ASC) or mean amino acid mass in *E. coli* (continued).

E. coli proteins					Length vs. Avg Cost				Length vs. Avg Mass			
					Pearson		Spearman		Pearson		Spearman	
Protein Length	Number	Avg Length	Avg Cost	Avg Mass	R	P value	R	P value	R	P value	R	P value
All proteins	4333	312.46	23.55	111.01	**-0.0992**	**5.92E-11**	**-0.0503**	**0.0009**	**-0.1225**	**6.00E-16**	**-0.1738**	**0, t=-11.62**
≤25 amino acids	7	17.57	30.50	121.29	-0.7388	0.0578	-0.8469	0.0162	-0.5140	0.2379	-0.3784	0.4026
≤50 amino acids	65	37.82	25.61	113.55	**-0.4416**	**2.31E-04**	**-0.2536**	**0.0416**	-0.2284	0.0672	-0.1155	0.3595
≤75 amino acids	240	57.76	24.46	113.03	**-0.3644**	**6.01E-09**	**-0.2314**	**0.0003**	-0.1140	0.0781	-0.0710	0.2732
≤100 amino acids	478	72.91	23.95	112.32	**-0.3524**	**2.01E-15**	**-0.2698**	**2.05E-09**	**-0.1645**	**3.04E-04**	**-0.1442**	**0.0016**
≤125 amino acids	718	86.09	23.88	112.20	**-0.2215**	**2.00E-09**	**-0.1504**	**5.20E-05**	**-0.1172**	**0.001658**	**-0.1124**	**0.0026**
≤150 amino acids	954	98.88	23.81	112.07	**-0.1888**	**4.17E-09**	**-0.1285**	**6.90E-05**	**-0.1104**	**6.37E-04**	**-0.1131**	**0.0005**
≤175 amino acids	1187	111.28	23.75	111.97	**-0.1697**	**4.00E-09**	**-0.1150**	**7.15E-05**	**-0.1092**	**1.63E-04**	**-0.1129**	**9.68E-05**
≤200 amino acids	1438	124.62	23.71	111.82	**-0.1574**	**1.95E-09**	**-0.1098**	**2.99E-05**	**-0.1296**	**8.24E-07**	**-0.1361**	**2.22E-07**
≤300 amino acids	2403	175.21	23.64	111.50	**-0.1033**	**3.90E-07**	**-0.0549**	**0.0071**	**-0.1451**	**8.83E-13**	**-0.1504**	**1.26E-13**
≤500 amino acids	3759	251.50	23.59	111.01	**-0.0674**	**3.53E-05**	-0.0318	0.0513	**-0.2223**	**0, t=-13.97**	**-0.2296**	**0, t=-14.46**
≤1000 amino acids	4276	300.17	23.56	111.01	**-0.0794**	**2.03E-07**	**-0.0412**	**0.0070**	**-0.1320**	**4.45E-18**	**-0.1765**	**0, t=-11.72**
≥50 amino acids	4272	316.39	23.53	110.97	**-0.0774**	**4.13E-07**	-0.0331	0.0306	**-0.1129**	**1.35E-13**	**-0.1648**	**0, t=-10.92**
≥100 amino acids	3861	341.79	23.50	110.84	**-0.0706**	**1.13E-05**	-0.0191	0.2360	**-0.0830**	**2.42E-07**	**-0.1303**	**4.37E-16**
≥150 amino acids	3388	372.17	23.49	110.71	**-0.0669**	**9.69E-05**	-0.0157	0.3594	**-0.0465**	**0.0068**	**-0.0937**	**4.65E-08**
≥200 amino acids	2902	405.27	23.48	110.61	**-0.0767**	**3.50E-05**	-0.0222	0.2311	-0.0132	0.4765	**-0.0595**	**0.0013**
≥300 amino acids	1938	482.59	23.45	110.39	**-0.0848**	**1.87E-04**	0.0087	0.7017	**0.0695**	**0.0022**	**0.0699**	**0.0021**
≥500 amino acids	581	709.12	23.35	110.97	**-0.2317**	**1.60E-08**	**-0.1908**	**3.64E-06**	-0.0355	0.3935	0.0479	0.2488
≥1000 amino acids	57	1234.70	22.82	110.50	**-0.3158**	**0.016716**	-0.3325	0.0115	-0.1590	0.2375	-0.0148	0.9131

BOLD: P < 0.01

Table 3.15: Comparison of the ASCs of extracellular vs intracellular fiber and fiber-associated proteins in *E. coli*.

	Extracellular	Intracellular
Fiber proteins:	RecA, MreB, ZapB (YiiU) RodZ (YfgA), FtsZ, MinD	FliC, FlgE, FimA, CsgA PapA* (P04127), FlaK**
Fiber-associated proteins‡:	*FtsA, ZipA, ZapA*, FtsE FtsX, FtsK, FtsQ, FtsL FtsB, FtsW, FtsI, FtsN *MinC, MinE*, DicB, *MreC* MreD	*FliD, FlgK, FlgL, FlgD*, *FliK* *CsgB*, *FimF*, FimG, FimH, PapE(P08407)*, PapF (P08408)* PapG(P13720)*, *PapK*(P62532)*

Note: Direct fiber-associated proteins (contact major fiber protein) are italicized above

	Mean	Student T Value	P Value†	Degrees of Freedom	Cost Diff (ATP)	U-Test Z	U-Test P Value
vs Intracellular Fibers	21.575						
Extracellular Fibers	19.844	5.410	0.000428	9	1.731	2.8823	0.00395
Extracellular Fibers without PapA	19.696	5.912	0.000357	8	1.879	2.7386	0.00617
Extracellular Fibers without FlaK	19.850	4.713	0.002175	7	1.725	2.7386	0.00617
Extracellular Fibers without PapA, FlaK	19.666	5.087	0.003811	5	1.909	2.5584	0.01052
vs Intracellular Fiber-associated	22.944						
Extracelluar Fiber-associated	21.534	2.352	0.026514	26	1.410	2.532	0.01134
Extracellular Fiber-associated without Pap Proteins	20.782	4.265	0.000269	24	2.162	3.4224	0.00062
vs Intracellular Direct Fiber-associated	22.031						
Extracelluar Direct Fiber-associated	20.752	2.843	0.014807	12	1.280	1.8074	0.0707
Extracelluar Direct Fiber-associated without PapK	20.536	3.512	0.004863	11	1.495	2.2857	0.02227
vs Cytoplasmic Proteins	23.555						
Intracellular Fibers	21.575	9.952	0.000175	5	1.980	3.5948	0.00032
Extracellular Fibers	19.844	14.685	2.65E-05	5	3.711	4.1923	2.76E-05
Extracellular Fibers wihout PapA	19.696	15.435	0.000103	4	3.859	3.8367	0.00012
Extracellular Fibers without FlaK	19.850	11.991	0.000277	4	3.705	3.8256	0.00013
Extracellular Fibers wihout PapA, FlaK	19.666	12.157	0.001198	3	3.889	3.4321	0.0006
Intracellular Fiber-associated	22.944	1.523	0.147252	16	0.611	2.4607	0.01387
Extracellular Fiber-associated	21.534	4.523	0.000698	12	2.021	4.366	1.27E-05
Intracellular Direct Fiber-associated	22.031	5.334	0.003104	5	1.524	2.9116	0.0036
Extracelluar Direct Fiber-associated	20.752	8.025	8.93E-05	7	2.803	4.4242	9.68E-06

† Student T test: Assumes Unequal Variance. Confidence Intervals are 2 tailed

* Not in K-12

** Average ASC of Flagellin phase 2 variable forms in Table 3.18D. Not in K-12

‡ Intracellular fiber associated proteins located in the cytoplasm or inner membrane compartments.

Table 3.16: Calculations of cost savings and fitness.

Description	Value calculated/used	Literature values & Notes
Average Doubling Time	30 minutes	30 minutes(2)
Number of H^+ per ATP	4 (143, 158)	4 (143, 158)
Number of H^+ per flagellar rotation	1240 (94)	800 (93, 99), ~1000 (37), 1200 (38, 106), 1240±240 (94), 1303 (162),1170-1270 (162),
Flagella Rotation rates*.**	100Hz (73, 106)	10-20 (134),13.4 (18), 100 (73, 84, 93, 106), 170 (134), 200(134), 250 (37),60-270 (94, 134), 300 (20, 38, 101, 134)
Number of H^+/s (100Hz, 1flag.)	124,000	-
Number of H^+/s (100Hz, 5flag.)	620,000	-
Growth disadvantage flagella	-	2% (106)
Flagellar rotation cost**	-	0.1% total energy growth conditions (106)
Growth of flagella	-	10000 aa/s [49, 50]
Respiratory chain total	20 million H^+/s (106)	20 million H^+/s (106)
Total H^+ per doubling (30 min.)	36 billion	or 9 billion ATPs if all H^+ are converted to ATPs
F_0F_1-ATPases H^+/s	-	3 million (106), 8 million(106)
ATP molecules in cell	-	0.5 – 3 million (2), 0.3-1.4 billion(58)
ATP Cost per cell***	20 billion	6-11 billion (theoretical) (98), 12-20 billion (experimental) (98), 20-60 billion (3, 145), 55 billion(2)
Cost per aa (% each aa is same)	27.36	-
Average Cost per aa (MG1655)	23.555	-
FliC Cost per aa	19.155	-
FimA Cost per aa	19.525	-
Flagella: FliC		
Number of Flagella per cell	5	0-15 (8 typical) (105, 106), 5-10 (106),10 (2, 38), 9.9±2.9 log & 4.5±2.1 in stationary (78), 6.6±1.4 (74), 6±2 (124), 2.5 to 4.4 (69), 4 (73)
Subunits FliC per flagella	20000 (73, 106)	20000 (29, 37, 73, 87, 106), 5340/helix turn (19, 60), 30000 (167), 33000 (3000x11) (2)
Savings per aa vs average	4.4001	-
Length of FliC	498 (97)	-
Savings per FliC	2191.2 ATPs	-
One Gly to Trp substitution in FliC	6,260,000 ATPs	using 5 flagella

Table 3.16: Calculations of cost savings and fitness (continued).

Description	Value calculated/used	Literature values & Notes
Flagella: FliC (continued)		
One Gly to Trp substitution in FliC	0.0313%(10^{-4}) of total cellular cost	using 5 flagella & $20x10^9$ ATPs/cell
Total savings per flagellum	43824894 ATP	20000 FliC subunits
Saving for 5 flagella	219124470 ATP	-
Saving for 5 flagella	876497883 H^+	$4H^+$/ATP
Time Swimming (100Hz ,1 flag.)	1412 seconds	23.5 min: 1 flag. savings, constantly spinning
% Savings Total Respiratory Chain	2.43%	using 30 minute doubling and 5 flagella
% Savings ATP Cost per cell	1.096%	using 5 flagella & $20x10^9$ ATPs/cell
Type 1 Pili: FimA		
Number of Pili per cell	300	200-500(21, 73),100-1000 (106), 100-300(2, 105, 106), 100-200(85), 171-200 (140), 100-500 (157)
Number of FimA per pili	1000	1000 (21, 73), 3000 (73), 500 to 3000 (26, 59, 110), >>100(76), ~2800 (59) (19.31nm*27 & 2um)
Savings per aa vs average	4.0302	-
Length of FimA	182 (97)	-
Savings per FimA	733.5 ATPs	-
Total savings for Type 1 pili	220048920 ATP	-
Total savings for Type 1 pili	880548129 H^+	$4H^+$/ATP
Time Swimming (100Hz ,5 flag.)	1420 seconds	23.7 min., all 5 flagella spinning
% Savings Total Respiratory Chain	2.45%	30 minute doubling
% Savings ATP Cost per cell	1.100%	using 300 fimbriae & $20x10^9$ ATPs/cell
Typical Aerobic Gram Neg. Organism		
Extracellular protein aa cost (50% GC)	21.443 ATPs/aa	from eq. of line fitting Fig. 3.5B
Total protein aa cost (50%GC)	23.492 ATPs/aa	from eq. of line fitting Fig. 3.5B
Savings per Extracellular protein	2.0495 ATPs/aa	8.274% lower cost vs Total proteins
Total proteins/cell	$2.35x10^6$(106)	$3.6x10^6$ (2)
Extracellular proteins/cell	$5.0x10^5$	$1x10^6$(2); 100875 (5 flagella); 300900 (300 Type I Pili); $5x10^5$ S-layer proteins per cell(133)
Extracellular protein savings	1.537%	Assumes all amino acids are recovered. 300aa/protein, 2.0495ATPs/aa, $2x10^{10}$ ATPs/cell, $5x10^5$ extracellular proteins per cell

Table 3.16: Calculations of cost savings and fitness (continued).

Description	Value calculated/used	Literature values & Notes
Typical Aerobic Gram Neg. Organism (continued)		
Extracellular protein synthetic cost	16.08%	Assumes all amino acids are lost. 300aa/protein, 21.443 ATPs/aa, $2x10^{10}$ ATPs/cell, $5x10^{5}$ extracellular proteins per cell
Extracellular protein synthetic cost if aa biases = cellular proteins	17.62%	Assumes all amino acids are lost. if 300aa/protein, 23.492 ATPs/aa, $2x10^{10}$ ATPs/cell, $5x10^{5}$ extracellular proteins per cell

Direct Competition — Assumes savings directly affect doubling times

Generations	Fold Increase	% Cheaper Organism
100	2.922	74.503
200	8.539	89.516
250	14.596	93.588
500	213.043	99.533
1000	45387.3	99.998

Notes: Most values are for *E. coli* K-12 or *S. typhimurium* except in a few instances (*Streptococcus(84, 94))*

*Many values are maximum rates and depend on temperature(20, 94) and other factors such as PMF(49) and pH(93, 99). Values can differ significantly depending on how they are measured (bundle, individual fiber, tethered cell, bead on polyhook, electrorotation of tethered cells)(141) (134). Above 100-175 Hz(28) torque decreases significantly (efficiency plummets) and approaches zero at ~300Hz (effect is greater at lower temperatures and lower pH)(17, 20, 49, 99). For free-swimming cells the motor saturation point is about 100Hz (MacNab in (107)).

** Source of flagella power is PMF. Under aerobic respiration flagella rotation is saturation as it is a small percent of the total protons pumped. Under anaerobic glycolysis, ATP hydrolysis becomes the major source of flagellar power and swimming speed is reduced(106).

*** Values vary considerably and depend on conditions including growth rate, aerobic vs anaerobic conditions, availability and type of carbon source (glucose vs acetate), etc. Generally aerobic growth and richer media conditions require less ATPs/cell. A good discussion is listed in source (98) (references cited therein are (3, 11, 14, 45, 47, 64, 107, 115, 144-147, 156)). The calculation requires multiplying the cells per gram (dry cell weight) times $YATP_{max}$ (Number of ATP per gram of cells). Cell weight used in (98) is 0.28pg per cell (40 minutes growth on glucose(106) = 3.6 trillion cells per gram); however, cell weight varies with growth conditions (106) and changes would affect values listed in table. Theoretical values are smaller than experimental values because the former exclude some cell maintenance cost including component turnover, transport cost, and membrane polarization(98).

Table 3.17A: Comparisons of the mean ASCs of proteins in locations of *E. coli* using different cost values.

		Akashi & Gojobori		Craig & Weber		Molecular Mass	
	Number	Mean	% of Overall	Mean	% of Overall	Mean	% of Overall
Overall	4333	23.5549		20.3139		111.0052	
Cytoplasm	2859	23.3171	98.99	19.8168	97.55	111.6093	100.54
Inner Membrane	972	24.6529	104.66	22.2954	109.75	109.9862	99.08
Integral Membrane Protein	777	25.0102	106.18	22.9046	112.75	109.7845	98.90
Membrane anchored	158	23.3743	99.23	20.0457	98.68	110.9876	99.98
Periplasmic N-term mem anchor	6	23.4800	99.68	19.6510	96.74	111.9889	100.89
Inner Membrane Lipoprotein	14	21.9869	93.34	18.3978	90.57	109.2690	98.44
Membrane associated	10	22.7075	96.40	19.2283	94.66	109.9723	99.07
Membrane Lipoprotein*	7	22.9700	97.52	19.8973	97.95	109.5073	98.65
Periplasm	337	22.8538	97.02	19.3281	95.15	109.7200	98.84
Outer Membrane	149	22.8521	97.02	19.4946	95.97	109.6250	98.76
OM ß-barrel protein	53	23.3737	99.23	20.2872	99.87	110.7320	99.75
OM Lipoprotein	96	22.5642	95.79	19.0570	93.81	109.0138	98.21
Cell surface appendage	12	20.4738	86.92	17.1651	84.50	103.8034	93.51
Secreted and released	4	21.1865	89.95	17.1547	84.45	108.1931	97.47
Extracellular	16	20.6520	87.68	17.1625	84.49	104.9008	94.50

		Cabons per aa		Nitrogens per aa		Sulfur per aa*		Mean ΔG per aa	
	Number	Mean	% Overall	Mean	% Overall	Mean	%Overall	Mean	% Overall
Overall	4333	4.9663		1.3596		0.0425		206.0956	
Cytoplasm	2859	4.9530	99.73	1.3857	101.92	0.0425	100.00	207.2800	100.57
Inner Membrane	972	5.0646	101.98	1.2950	95.25	0.0452	106.30	204.6274	99.29
Periplasm	337	4.8686	98.03	1.3387	98.46	0.0371	87.33	202.4783	98.24
Outer Membrane	149	4.8464	97.59	1.3373	98.36	0.0394	92.57	202.7336	98.37
Extracellular	16	4.5437	91.49	1.2744	93.73	0.0228	53.63	191.1559	92.75

*U-test of S/aa: Extracellular vs Cytoplasmic, P=1.47E-6. For Sec- (Table 3.7) %Overall =42.8%, P=1.33E-9.

	Mean cost	Spearman			Kendall				Pearson		
vs Akashi&Gojobori	23.555	R	t-test	p value	Tau	Z	p value	Gamma	R	t-test	p value
Craig&Weber	20.314	0.905	139.76	→0	190.58	18809.6	→0	0.73907	0.918	151.88	→0
Molecular Mass	111.005	0.571	45.78	→0	89.06	8789.73	→0	0.42936	0.592	48.38	→0
Carbons per Amino Acid	4.966	0.902	137.89	→0	279.17	27553.2	→0	0.91605	0.916	150.31	→0
Nitrogens per Amino Acid	1.36	0.058	3.86	0.00012				0.05176	0.034	2.23	0.02573
Seawater ΔG(kJ /mol) 18°C	206.096	0.622	52.26	→0	95.42	9417.33	→0	0.46002	0.636	54.24	→0
vs Molecular Mass	111.005										
Seawater ΔG(kJ/mol) 18°C	206.096	0.900	135.90	→0				0.733	0.903	138.71	→0
Carbons per Amino Acid	4.966	0.784	83.12	→0	201.12	19850	→0	0.61254	0.802	88.29	→0
Nitrogens per Amino Acid	1.36	0.653	56.67	→0				0.4741	0.645	55.55	→0

Table 3.17A: Comparisons of the mean ASCs of proteins in locations of *E. coli* using different cost values (continued).

Mann Whitney U Test	Akashi & Gojobori		Craig & Weber		Molecular Mass	
	Z	p value	Z	p value	Z	p value
vs Cytoplasm						
Inner Membrane	23.196	4.98E-119	29.497	3.18E-191	13.025	8.85E-39
Integral Membrane Protein	27.638	3.90E-168	34.510	5.65E-261	13.352	1.16E-40
Membrane anchored	0.649	0.5164	0.013	0.9897	2.258	0.0240
Periplasmic w N-term mem anchor	0.367	0.7139	0.245	0.8068	0.277	0.7820
Inner Membrane Lipoprotein	4.153	3.28E-05	3.795	0.0001	2.908	0.0036
Membrane associated	1.777	0.0755	1.286	0.1985	1.831	0.0670
Membrane Lipoprotein*	1.124	0.2609	0.470	0.6384	1.950	0.0512
Periplasm	6.029	1.65E-09	5.698	1.21E-08	9.772	1.48E-22
Outer Membrane	3.713	2.05E-04	2.510	0.0121	5.975	2.30E-09
OM ß-barrel protein	0.305	0.7604	2.374	0.0176	1.843	0.0653
OM Lipoprotein	4.892	9.98E-07	4.906	9.31E-07	6.151	7.72E-10
Cell surface appendage	5.606	2.07E-08	4.883	1.05E-06	5.690	1.27E-08
Secreted and released	2.269	0.0233	2.264	0.0236	1.980	0.0477
Extracellular	5.983	2.19E-09	5.355	8.55E-08	5.912	3.38E-09
vs All						
Cytoplasm	6.223	4.89E-10	8.631	6.09E-18	7.263	3.78E-13
Inner Membrane	19.530	6.17E-85	24.304	1.77E-130	8.729	2.58E-18
Integral Membrane Protein	24.013	2.05E-127	29.427	2.51E-190	9.405	5.20E-21
Membrane anchored	2.102	0.0356	2.156	0.0311	0.204	0.8381
Periplasmic w N-term mem anchor	0.051	0.9594	0.735	0.4621	0.807	0.4198
Inner Membrane Lipoprotein	4.275	1.91E-05	4.034	5.49E-05	2.215	0.0268
Membrane associated	2.103	0.0355	1.777	0.0756	1.181	0.2374
Membrane Lipoprotein*	1.317	0.1878	0.836	0.4030	1.410	0.1587
Periplasm	8.243	1.69E-16	8.648	5.25E-18	6.727	1.74E-11
Outer Membrane	5.149	2.62E-07	4.599	4.25E-06	3.921	8.83E-05
OM ß-barrel protein	0.761	0.4466	0.633	0.5269	0.479	0.6316
OM Lipoprotein	5.890	3.86E-09	6.231	4.62E-10	4.559	5.14E-06
Cell surface appendage	5.567	2.60E-08	4.958	7.12E-07	5.496	3.89E-08
Secreted and released	2.369	0.0179	2.393	0.0167	1.643	0.1004
Extracellular	6.001	1.96E-09	5.487	4.10E-08	5.578	2.44E-08

Table 3.17B: Comparisons of the mean ASCs of proteins in locations of *S. cerevisiae* using different cost values.

		Wagner Respiratory		Wagner Fermentative		Bragg & Wagner Respiratory		Bragg & Wagner Fermentative		Amino Acid Molecular Mass	
	Number	Mean	% of Overall	Mean	% of Overall	Mean	% of Overall	Mean	% of Overall	Mean	% of Overall
Overall	6732	26.59		6.61		25.14		5.81		113.20	
Cytoplasm	2845	25.86	97.25	6.46	97.68	24.32	96.73	5.61	96.51	112.86	99.70
PM	186	27.05	101.70	6.54	98.99	25.94	103.17	5.94	102.11	111.66	98.64
Integral/Endo Memb.	172	28.46	107.04	6.69	101.22	27.42	109.04	6.12	105.19	112.81	99.65
Cell Wall	44	24.04	90.39	5.64	85.31	23.21	92.30	5.18	89.17	105.27	93.00
Extracellular	54	24.88	93.54	5.92	89.56	24.00	95.46	5.44	93.57	107.16	94.66
Extracellular & Cell Wall	98	24.50	92.13	5.79	87.65	23.65	94.04	5.32	91.59	106.31	93.92

	Mean	Spearman			Kendall				Linear (Pearson)		
vs Yeast respiratory (Wagner)	26.5937	R	t-test	p value	Tau	Z	p value	Gamma	R	t-test	p value
Yeast respiratory (Bragg&Wagner)	25.145	0.959	278.2	→0	772.5	95050	→0	0.83524	0.966	308.5	→0
Yeast fermentative (Wagner)	6.61024	0.677	75.37	→0	380.3	46794	→0	0.49749	0.69	78.23	→0
Yeast fermentative (Bragg&Wagner)	5.81345	0.693	78.9	→0				0.51175	0.713	83.42	→0
Molecular Mass	113.198	0.483	45.31	→0				0.34333	0.547	53.6	→0
vs Yeast fermentative (Wagner)											
Yeast respiratory (Bragg&Wagner)		0.586	59.32	→0	237.7	29252	→0	0.42314	0.604	62.11	→0
Yeast fermentative (Bragg&Wagner)		0.802	110.1	→0	520.4	64025	→0	0.62318	0.855	135.1	→0
Molecular Mass		0.59	60.02	→0				0.42631	0.647	69.67	→0

Table 3.17B: Comparisons of the mean ASCs of protein in locations of *S. cerevisiae* using different cost values (continued).

Mann Whitney U Test	Wagner Respiratory		Wagner Fermentative		Bragg & Wagner Respiratory		Bragg & Wagner Fermentative		Amino Acid Molecular Mass	
vs Overall	Z	p value	Z	p value	Z	p value	Z	p value	Z	p value
Cytoplasm	14.8616	5.85E-50	12.7545	2.94E-37	15.4240	1.13E-53	15.0629	2.84E-51	5.6203	1.91E-08
PM	4.3898	1.13E-05	0.2869	0.7741849	6.0996	1.06E-09	4.9427	7.71E-07	7.0659	1.60E-12
Integral/Endo Membranes	12.3372	5.71E-35	3.8335	0.0001263	13.2731	3.32E-40	9.4490	3.42E-21	2.2301	0.0257411
Cell Wall	6.8980	5.27E-12	8.5229	1.56E-17	5.3273	9.97E-08	6.1732	6.69E-10	8.6931	3.53E-18
Extracellular	4.8461	1.26E-06	7.1035	1.22E-12	3.1973	0.001387	12.6762	8.01E-37	7.6434	2.11E-14
Extracellular & Cell Wall	8.1895	2.62E-16	10.9442	7.08E-28	5.9212	3.20E-09	10.1865	2.28E-24	11.4573	2.16E-30
vs Cytoplasm										
PM	9.0114	2.04E-19	3.7753	0.0001598	10.9088	1.05E-27	9.9169	3.52E-23	5.6918	1.26E-08
Integral/Endo Membranes	16.5602	1.35E-61	8.1420	3.89E-16	17.4758	2.19E-68	14.3009	2.16E-46	0.7752	0.4382106
Cell Wall	5.7802	7.46E-09	7.9861	1.39E-15	3.7201	0.0001991	4.8510	1.23E-06	8.5915	8.59E-18
Extracellular	3.0802	0.0020684	6.0747	1.24E-09	1.0085	0.3132149	1.8219	0.0684758	7.2476	4.24E-13
Extracellular & Cell Wall	6.1067	1.02E-09	9.7773	1.41E-22	3.2127	0.0013148	4.5628	5.05E-06	11.0433	2.36E-28

Table 3.17C: Comparisons of the mean ASCs of proteins in locations of *M. tuberculosis* H37Rv using different cost values.

		Akashi & Gojobori		Craig & Weber		Molecular Mass	
	Number	Mean	% of Overall	Mean	% of Overall	Mean	% of Overall
Overall	3927	22.1469		19.3088		107.1443	
Cytoplasm	2111	22.2479	100.46	19.1672	99.27	108.5627	101.32
Cytoplasmic Membrane	598	22.9850	103.78	20.8788	108.13	106.1764	99.10
Unknown	1070	21.8415	98.62	18.9533	98.16	106.3746	99.28
Extracellular	140	19.4465	87.81	17.5376	90.83	95.9992	89.60
Cell Wall	8	20.9739	94.70	17.8469	92.43	103.1890	96.31
Extracellular &Cell Wall	148	19.52908	88.18	17.55428	90.91	96.38786	89.96

	Mean cost	Spearman				Pearson		
vs Akashi & Gojobori	22.1456	**R**	**t-test**	**p value**	**Gamma**	**R**	**t-test**	**p value**
Craig & Weber	19.3088	0.8744062	112.9059985	→0	0.697734461	0.8728684	112.0696964	→0
Molecular Mass	107.1443	0.7032694	61.97531318	→0	0.532250842	0.7803322	78.17459575	→0

Mann Whitney U Test	Akashi & Gojobori		Craig & Weber		Molecular Mass	
	Z	**p value**	**Z**	**p value**	**Z**	**p value**
vs Cytoplasm						
Cytoplasmic Membrane	10.7167	8.495E-27	19.0031	1.61E-80	14.4926	1.35E-47
Unknown	8.9557	3.375E-19	5.1720	2.316E-07	15.0537	3.27E-51
Extracellular	12.3710	3.752E-35	10.6110	2.65E-26	15.3676	2.70E-53
Cell Wall	2.4495	0.0143057	2.0390	0.0414474	3.4030	0.0006665
Extracellular and Cell Wall	12.5622	3.405E-36	10.7611	5.26E-27	15.6863	1.88E-55
vs All						
Cytoplasm	2.0838	0.0371784	1.7299	0.0836427	11.3553	6.98E-30
Cytoplasmic Membrane	11.7610	6.203E-32	17.8934	1.33E-71	7.8344	4.66E-15
Unknown	7.5395	4.718E-14	6.4176	1.385E-10	7.3012	2.85E-13
Extracellular	11.8675	1.747E-32	10.6696	1.41E-26	13.6224	2.95E-42
Cell Wall	2.2392	0.025142	2.0735	0.0381271	2.7638	0.005713
Extracellular and Cell Wall	12.0425	2.123E-33	10.8408	2.21E-27	13.8675	9.96E-44

Table 3.18A: ASC of *P. syringae* HOPs and HOP Helpers

Mean Mass	ASC	Percentile	Total aa	Description	Note
101.564	20.107	98.841	422	gi|28868125|PSPTO0905 glycosyl hydrolase family 5	HOP
107.648	20.432	98.056	488	gi|28871892|PSPTO4776 type III effector HopPmaI(Pto)	HOP
105.695	20.432	98.038	416	gi|28870457|PSPTO3293 hypothetical protein	HOP
104.244	20.476	97.895	404	gi|28870456|PSPTO3292 hypothetical protein	HOP
107.477	20.486	97.860	553	gi|28870262|PSPTO3087 avirulence protein AvrPtoB	HOP
107.706	20.560	97.628	279	gi|28867704|PSPTO0474 hypothetical protein	HOP
104.303	20.566	97.592	486	gi|28868580|PSPTO1372 type III effector HopPtoA1	HOP
103.549	20.625	97.432	404	gi|28872174|PSPTO5061 candidate type III effector Hop protein	HOP
108.166	20.629	97.396	287	gi|28867301|PSPTO0061 candidate type III helper HolPtoY	HOP
105.613	20.746	96.736	712	gi|28868583|PSPTO1375 type III effector HopPtoM	HOP
105.278	20.749	96.701	118	gi|28871729|PSPTO4597 candidate type III effector HolPtoZ	HOP
107.772	*20.868*	95.809	*384*	*gi|29171488|type III effector HopX1|pDC3000A*	*HOP*
104.950	21.013	94.650	487	gi|28871842|PSPTO4718 type III effector HopPtoA2	HOP
108.109	21.103	93.990	338	gi|28867285|PSPTO0044 type III effector HopPtoK	HOP
106.294	21.176	93.508	705	gi|28868096|PSPTO0876 type III effector HopPtoD1	HOP
107.385	21.217	93.223	83	gi|28871829|PSPTO4703 candidate type III effector HolPtoAA	HOP
105.429	21.302	92.367	177	gi|28871720|PSPTO4588 candidate type III effector Hop protein	HOP
107.500	21.337	91.849	1957	gi|28868103|PSPTO0883 candidate type III effector HolPtoR	HOP
108.709	21.614	87.837	1795	gi|28868585|PSPTO1377 avirulence protein AvrE(Pto)	HOP
109.748	21.700	86.374	468	gi|28871846|PSPTO4722 type III effector HopPtoD2	HOP
108.056	21.739	85.536	168	gi|28871724|PSPTO4592 HopPtoO-related protein	HOP
108.020	21.811	84.038	466	gi|28868614|PSPTO1406 type III effector HopPtoB1	HOP
111.655	21.947	80.917	164	gi|28871145|PSPTO4001 avirulence protein AvrPto(DC3000)	HOP
110.451	21.973	80.257	380	gi|28872465|PSPTO5354 type III effector HopPsyA(Pto)	HOP
109.687	22.003	79.597	204	gi|28867732|PSPTO0502 type III effector HopPtoF	HOP
108.953	22.115	76.743	284	gi|28868774|PSPTO1568 type III effector HopPtoJ	HOP
110.450	22.153	75.745	403	gi|28868127|PSPTO0907 hypothetical protein	HOP
109.548	22.190	74.835	632	gi|28871821|PSPTO4691 type III effector HopPtoI	HOP
109.462	22.216	74.050	298	gi|28871726|PSPTO4594 type III effector HopPtoS3	HOP
108.876	22.248	72.998	447	gi|28868097|PSPTO0877 candidate type III effector HolPtoQ	HOP
108.806	*22.291*	71.874	*378*	*gi|29171494|type III effector HopT1-1|pDC3000A*	*HOP*
108.140	22.306	71.518	389	gi|28871725|PSPTO4593 candidate type III effector HolPtoU2	HOP
111.456	22.382	69.146	218	gi|28867816|PSPTO0588 type III effector HopPtoH	HOP
107.665	22.388	68.914	269	gi|28867817|PSPTO0589 type III effector HopPtoC	HOP
110.507	22.493	65.525	261	gi|28868126|PSPTO0906 virulence protein putative	HOP
110.799	*22.545*	63.831	*283*	*gi|29171493|type III effector HopO1-1|pDC3000A*	*HOP*
110.456	22.742	57.410	899	gi|28870054|PSPTO2872 type III effector HopPtoL	HOP
110.845	22.761	56.626	493	gi|28871849|PSPTO4727 type III effector HopPtoG	HOP
109.895	22.888	52.042	152	gi|28868121|PSPTO0901 candidate type III effector Hop protein	HOP
110.105	22.898	51.685	350	gi|28868578|PSPTO1370 type III effector HopPtoN	HOP
113.754	23.257	39.736	211	gi|28871468|PSPTO4331 type III effector HopPtoE	HOP
113.518	23.372	35.919	264	gi|28867731|PSPTO0501 type III effector HoptoS2	HOP
110.529	23.382	35.581	390	gi|28871844|PSPTO4720 hypothetical protein |HopVI	HOP
113.349	24.055	19.547	276	gi|28868239|PSPTO1022 avirulence protein AvrPpiB1(Pto)	HOP
113.349	*24.055*	19.529	*276*	*gi|29171482|type III effector HopAM1-1|pDC3000A*	*HOP*
98.635	18.191	100.000	370	gi|28868590|PSPTO1382 type III helper protein HrpZ(Pto)	helper
99.849	18.886	99.839	324	gi|28869866|PSPTO2678 type III helper protein HopPtoP	helper
101.162	19.213	99.661	424	gi|28868581|PSPTO1373 type III helper protein HrpW(Pto)	helper
100.609	19.300	99.590	113	gi|28868589|PSPTO1381 type III helper protein HrpA(Pto) - HrpA	helper
103.797	20.130	98.752	780	gi|28868613|PSPTO1405 type III helper protein HrpK(Pto)	helper
106.065	20.681	97.075	555	gi|28871243|PSPTO4101 type III helper protein HopPmaH(Pto)	helper
109.359	21.812	84.002	495	gi|28868586|PSPTO1378 memb-bound lytic murein transglycosylase D putative	helper

Table 3.18B: ASC of plant and animal Type III effectors

ASC Akashi	Craig & Weber	Avg. Mass	Total aa	Set	SwissProt Number	Organism	Protein Name	Description
20.903	18.068	104.882	219	A	A6M3N5	*Yersinia pestis CA88-4125*	YopE	Translocated host-GTPase-activating protein
20.816	18.203	104.699	219	A	Q7BRY7	*Yersinia enterocolitica*	YopE	Yop effector YopE
21.066	17.544	108.657	468	A	A9R9K8	*Yersinia pestis bv. Antiqua (strain Angola)*	YopH	Protein-tyrosine-phosphatase YopH
21.116	17.572	108.807	468	A	Q7BRY8	*Yersinia enterocolitica*	YopH	Yop effector YopH
24.338	20.640	115.285	182	A	B0A3S4	*Yersinia pestis biovar Orientalis str. F1991016*	YopK	YopK protein
24.349	20.657	115.279	182	A	Q56935	*Yersinia pseudotuberculosis*	YopK	Yop targeting protein yopK, yopQ
24.186	20.330	114.345	182	A	Q7BS06	*Yersinia enterocolitica*	YopQ	YopQ
23.655	19.168	112.741	288	A	B0HNN9	*Yersinia pestis biovar Antiqua str. B42003004*	YopJ	Effector protein YopJ
23.583	19.066	112.500	288	A	Q93KQ5	*Yersinia enterocolitica*	YopP	Yop effector YopP
22.126	19.043	112.922	409	A	A6M3U5	*Yersinia pestis CA88-4125*	YopM	Leucine-rich 15-repeat translocated effector protein
21.896	18.906	112.760	550	A	Q663L9	*Yersinia pseudotuberculosis*	YopM	YopM; putative targeted effector protein
21.880	18.725	112.590	367	A	Q93KU8	*Yersinia enterocolitica*	YopM	Yop effector YopM
24.494	21.363	112.702	322	A	B0A3S3	*Yersinia pestis biovar Orientalis str. F1991016*	YopT	Cysteine protease YopT
24.498	21.321	112.883	322	A	Q93RN4	*Yersinia pseudotuberculosis*	YopT	Cysteine protease yopT
24.616	21.494	113.109	322	A	P27475	*Yersinia enterocolitica*	YopT	Cysteine protease yopT
24.266	21.144	112.369	322	A	P0C2N1	*Yersinia enterocolitica*	YopT1	Cysteine protease yopT1
22.839	19.117	111.671	732	A	A9ZFE7	*Yersinia pestis biovar Orientalis str. IP275*	YopO	Protein kinase YopO
22.830	19.126	111.630	732	A	Q05608	*Yersinia pseudotuberculosis*	YpkA	Protein kinase ypkA
22.810	19.130	111.381	728	A	Q56921	*Yersinia enterocolitica*	YpkA	Protein kinase A
23.583	19.066	112.500	288	A	Q93KQ5	*Yersinia enterocolitica*	YopP	Yop effector YopP
22.781	19.129	111.438	729	A	O85239	*Yersinia enterocolitica*	YopO	Protein kinase YopO
22.304	19.321	111.091	165	A	A9ZER0	*Yersinia pestis biovar Orientalis str. IP275*	YscH	Type III secretion protein YopR
22.398	19.397	111.273	165	A	Q663I2	*Yersinia pseudotuberculosis*	YscH	Yop proteins translocation protein H
22.392	19.339	111.486	165	A	Q7BRZ4	*Yersinia enterocolitica*	YscH	Secreted protein YopR
21.351	17.877	108.502	397	A	O34020	*Chlamydophila caviae*	CopN	CopN protein
21.249	17.635	108.741	399	A	Q9Z8L4	*Chlamydophila pneumoniae*	LcrE	CopN
19.550	15.618	102.481	880	A	Q824H6	*Chlamydophila caviae*	TARP	Putative uncharacterized protein
19.198	15.473	101.606	1005	A	O84462	*Chlamydia trachomatis*	TARP	Translocated actin-recruiting phosphoprotein
19.763	15.717	102.729	755	A	Q9Z7Y1	*Chlamydophila pneumoniae*	TARP	Uncharacterized protein CPn_0572/CP_0177/CPj0572/CpB0594

Table 3.18B: ASC of plant and animal Type III effectors (continued)

ASC Akashi	Craig & Weber	Avg. Mass	Total aa	Set	SwissProt Number	Organism	Protein Name	Description
21.423	17.632	109.251	355	A	Q46210	*Chlamydophila caviae*	IncA	Inclusion membrane localised protein
22.769	19.764	111.022	273	A	Q3KMQ0	*Chlamydia trachomatis (strain A/HAR-13 / ATCC VR-571B)*	IncA	Inclusion membrane protein A
20.893	16.872	108.947	390	A	Q9Z8Z8	*Chlamydophila pneumoniae*	IncA	Inclusion membrane protein A
21.423	17.632	109.251	355	A	Q46210	*Chlamydophila caviae*	IncA	Inclusion membrane localised protein
23.328	20.796	106.398	115	A	O84235	*Chlamydia trachomatis*	IncB	Inclusion Membrane Protein B
21.250	18.310	102.826	176	A	Q9Z8P7	*Chlamydophila pneumoniae*	IncB	Inclusion Membrane Protein B
22.060	18.675	105.129	186	A	O30783	*Chlamydophila caviae*	IncC	Inclusion membrane protein C
22.179	20.320	103.458	178	A	O84236	*Chlamydia trachomatis*	IncC	Inclusion Membrane Protein C
21.858	18.096	105.469	203	A	Q9Z8P6	*Chlamydophila pneumoniae*	IncC	Inclusion Membrane Protein C
21.689	19.667	105.749	141	A	Q9RPQ1	*Chlamydia trachomatis*	IncD	Inclusion membrane protein D
21.961	20.227	102.555	132	A	O84118	*Chlamydia trachomatis*	IncE	Inclusion membrane protein E
21.540	20.361	100.179	104	A	O84119	*Chlamydia trachomatis*	IncF	Inclusion membrane protein F
20.819	19.213	104.114	167	A	Q3KMQ1	*Chlamydia trachomatis (strain A/HAR-13 / ATCC VR-571B)*	IncG	Inclusion membrane protein G
23.027	18.720	110.704	288	A	Q9Z9F5	*Chlamydophila pneumoniae*		Putative uncharacterized protein
22.585	18.683	116.207	651	A	Q9Z7W9	*Chlamydophila pneumoniae*		CPj0585 protein
21.076	17.135	108.391	336	A	P40613	*Salmonella typhimurium*	SpaN	Surface presentation of antigens protein spaN
21.335	16.859	107.918	685	A	Q56027	*Salmonella typhimurium*	SipA	Cell invasion protein sipA
22.637	19.744	110.667	788	A	Q9RPH0	*Salmonella typhimurium*	SspH2	Leucine-rich repeat protein
24.193	19.820	114.525	336	A	Q56061	*Salmonella typhimurium*	SifA	acc:Q56061 Protein sifA
23.209	19.708	110.953	782	A	Q8ZNR3	*Salmonella typhimurium*	SopA	Secreted effector protein of Salmonella
22.881	19.104	110.369	561	A	Q57QR2	*Salmonella choleraesuis*	SopB	Outer protein
24.126	20.126	113.954	317	A	P40722	*Salmonella typhimurium*	SopD	Sop effector protein sopD
22.878	18.523	110.906	240	A	O52623	*Salmonella typhimurium*	SopE	Guanine nucleotide exchange factor sopE
22.997	18.873	110.126	240	A	Q7CQD4	*Salmonella typhimurium*	SopE2	Guanine nucleotide exchange factor sopE2
22.271	18.242	110.552	543	A	P74873	*Salmonella typhimurium*	SptP	Effector protein sptP
19.334	15.437	103.303	574	A	Q58I88	*Escherichia coli*	Tir	Translocated intimin receptor
20.532	16.174	106.515	192	A	Q47184	*Escherichia coli*	EspA	EspA
19.658	15.963	104.528	312	A	Q8XC86	*Escherichia coli O157:H7*	EspB	EspB
20.383	17.533	104.453	374	A	Q7DB81	*Escherichia coli O157:H7*	EspD	EspD

Table 3.18B: ASC of plant and animal Type III effectors (continued)

ASC Akashi	Craig & Weber	Avg. Mass	Total aa	Set	SwissProt Number	Organism	Protein Name	Description
20.291	17.663	101.753	248	A	Q7DB85	*Escherichia coli O157:H7*	EspF	EspF
22.146	18.472	110.271	398	A	O85646	*Escherichia coli*	EspG	EspG
22.035	17.562	110.309	337	A	Q8X2D5	*Escherichia coli O157:H7*	TccP	EspF-like protein
21.972	17.320	110.165	250	A	A2A0X3	*Escherichia coli O157:H-*	TccP2	Type III secreted effector protein
24.522	20.954	114.361	329	A	B2NN32	*Escherichia coli O157:H7 str. EC4196*	NleB	NleB
24.774	20.313	116.249	224	A	Q8XBX7	*Escherichia coli*	St47	Putative uncharacterized protein
22.746	19.514	112.119	321	P	P13835	*Pseudomonas syringae pv. glycinea*	AvrB	Avirulence protein B
20.746	17.572	105.613	712	P	Q887D0	*Pseudomonas syringae pv. tomato*	HopM1	Effector protein hopM1
22.248	18.733	108.876	447	P	Q888Y7	*Pseudomonas syringae pv. tomato*	HopQ1-1	Type III effector HopQ1-1
21.570	17.695	111.071	220	P	Q7BE94	*Pseudomonas syringae pv. maculicola*	AvrRpm1	Avr protein
22.115	18.995	108.953	284	P	Q886L1	*Pseudomonas syringae pv. tomato*	HopAF1	Type III effector HopAF1
20.629	16.300	108.166	287	P	Q88BF6	*Pseudomonas syringae pv. tomato*	HopY1	Type III effector HopY1
22.748	19.508	109.567	413	P	Q889A9	*Pseudomonas syringae pv. tomato*	HopAJ1	Type III effector HopAJ1
20.625	18.323	103.549	404	P	Q87V79	*Pseudomonas syringae pv. tomato*	HopAN1	Type III effector HopAN1
18.886	15.792	99.849	324	P	Q882F0	*Pseudomonas syringae pv. tomato*	HopP1	Type III helper protein HopP1
20.936	17.863	104.344	484	P	Q8RP03	*Pseudomonas syringae pv. maculicola*	HopPtoA1Pma	Type III effector HopPtoA1Pma
21.337	18.449	107.500	1957	P	Q888Y1	*Pseudomonas syringae pv. tomato*	HopR1	Type III effector HopR1
20.432	15.798	107.648	488	P	Q87W07	*Pseudomonas syringae pv. tomato*	HopI1	Type III effector HopI1
22.221	18.616	110.507	375	P	Q08370	*Pseudomonas syringae pv. syringae*	HrmA	Protein hrmA
22.306	19.222	108.140	389	P	Q87WF7	*Pseudomonas syringae pv. tomato*	HopT1-2	Type III effector HopT1-2
23.257	18.988	113.754	211	P	Q87X57	*Pseudomonas syringae pv. tomato*	HopE1	HopPtoE
22.761	19.051	110.845	493	P	Q87W42	*Pseudomonas syringae pv. tomato*	HopG1	HopPtoG
22.382	18.289	111.456	218	P	Q88A09	*Pseudomonas syringae pv. tomato*	HopH1	Type III effector HopH1
22.742	19.811	110.456	899	P	Q881L7	*Pseudomonas syringae pv. tomato*	HopL1	Type III effector HopL1
21.276	18.272	107.266	204	P	Q9K2L5	*Pseudomonas syringae pv. phaseolicola*		ORF2
23.382	19.896	110.529	390	P	Q87W46	*Pseudomonas syringae pv. tomato*	HopV1	Type III effector HopV1
20.560	16.496	107.706	279	P	Q88AB8	*Pseudomonas syringae pv. tomato*	HopAS1	Type III effector HopAS1
21.412	18.878	106.338	914	P	Q7PC62	*Pseudomonas syringae pv. syringae (strain B728a)*	HopAE1	Effector protein hopAE1
22.760	19.457	111.585	2042	P	Q7PC42	*Pseudomonas syringae pv. syringae (strain B728a)*	HopAC1	Putative type III effector HolPtoACPsy

Table 3.18B: ASC of plant and animal Type III effectors (continued)

ASC Akashi	Craig & Weber	Avg. Mass	Total aa	Set	SwissProt Number	Organism	Protein Name	Description
23.385	19.566	111.770	311	P	Q52530	*Pseudomonas syringae pv. phaseolicola*	AvrD	Avirulence gene D (Fragment)
21.811	18.605	108.020	466	P	Q9L6W4	*Pseudomonas syringae pv. tomato*	HopB1	Putative uncharacterized protein
22.396	19.459	107.661	269	P	Q9F3T4	*Pseudomonas syringae pv. pisi*	AvrPpiC2	Probable cysteine protease avirulence protein avrPpiC2
20.763	17.388	107.871	380	P	Q52394	*Pseudomonas syringae pv. phaseolicola*	AvrPphE	AvrPphE protein
20.924	17.284	110.495	539	P	Q48B61	*Pseudomonas syringae pv. phaseolicola (strain 1448A / Race 6)*	HopAB1	Effector protein hopAB1
22.493	18.929	110.507	261	P	Q888W0	*Pseudomonas syringae pv. tomato*	HopAI1	Type III effector HopAI1
22.891	19.421	110.412	716	P	Q7PC45	*Pseudomonas syringae pv. syringae (strain B728a)*	HopAG1	Type III effector HopAG1
22.837	19.386	111.171	907	P	P11437	*Pseudomonas syringae pv. glycinea*	AvrA	Avirulence protein A
20.248	16.680	108.453	220	P	Q52432	*Pseudomonas syringae*	AvrRps4	Avirulence protein (Fragment)
21.209	18.155	106.180	710	P	Q48BE0	*Pseudomonas syringae pv. phaseolicola (strain 1448A / Race 6)*	HopD1	Type III effector HopD1
21.183	17.132	108.895	494	P	Q52389	*Pseudomonas syringae*		Putative uncharacterized protein
22.898	20.353	110.105	350	P	Q9JP32	*Pseudomonas syringae pv. tomato*	HopN1	Type III effector HopN1
22.190	18.750	109.548	632	P	Q87W65	*Pseudomonas syringae pv. tomato*	HopAD1	Effector protein hopAD1
20.681	16.777	106.065	555	P	Q87XS5	*Pseudomonas syringae pv. tomato*	HopAK1	Type III effector HopAK1
20.130	16.451	103.797	780	P	Q9L6W3	*Pseudomonas syringae pv. tomato*	HrpK	HrpK
22.398	19.397	111.273	165	T	Q663I2	*Yersinia pseudotuberculosis*	YscH	Yop proteins translocation protein H
21.853	18.000	111.435	221	T	A6M3R1	*Yersinia pestis CA88-4125*	YscL	Regulator of YscN ATPase
21.300	16.348	114.370	210	T/C	P94517	*Bacillus subtilis*	YscB	Uncharacterized protein yscB
25.537	23.396	112.485	217	T	B0HZP5	*Yersinia pestis biovar Antiqua str. E1979001*	YscR	Type III secretion apparatus protein YscR
24.656	22.194	114.704	209	T	A6M3R2	*Yersinia pestis CA88-4125*	YscK	Putative type III secretion protein
22.285	19.206	108.845	439	T	B0HZP9	*Yersinia pestis biovar Antiqua str. E1979001*	YscN	Type III secretion apparatus H+-transporting two-sector ATPase YscN
26.185	22.101	124.197	74	T	P94518	*Bacillus subtilis*	YscA	Uncharacterized protein yscA
23.313	20.360	114.903	114	T	A6M3T7	*Yersinia pestis CA88-4125*	YscY	Putative type III secretion protein
24.047	21.816	109.519	226	T	A9K514	*Burkholderia mallei ATCC 10399*	BsaW	Type III secretion system protein BsaW
22.304	19.321	111.091	165	T	A9R9K1	*Yersinia pestis bv. Antiqua (strain Angola)*	YscH	Type III secretion protein YopR
25.009	23.730	108.941	261	T	B0HZP3	*Yersinia pestis biovar Antiqua str. E1979001*	YscT	Type III secretion apparatus protein YscT
24.292	21.738	109.688	216	T	A9K4S2	*Burkholderia mallei ATCC 10399*	SctR	Type III secretion inner membrane protein SctR
25.582	23.305	111.982	215	T	B4TH61	*Salmonella enterica subsp. enterica serovar Heidelberg str. SL476*	EpaP	Type III secretion apparatus protein, YscR/HrcR family
20.598	16.591	107.774	115	T	A9R9K6	*Yersinia pestis bv. Antiqua (strain Angola)*	YscM	Type III secretion apparatus protein YscM

Table 3.18B: ASC of plant and animal Type III effectors (continued)

ASC Akashi	Craig & Weber	Avg. Mass	Total aa	Set	SwissProt Number	Organism	Protein Name	Description
24.737	20.926	112.202	360	T	Q1MQX2	*Lawsonia intracellularis (strain PHE/MN1-00)*	YscU	Translocation protein in type III secretion
22.895	19.674	110.756	244	T	B0HZN2	*Yersinia pestis biovar Antiqua str. E1979001*	YscJ	Type III secretion apparatus lipoprotein YscJ
26.951	25.885	112.116	262	T	Q2SC30	*Hahella chejuensis (strain KCTC 2396)*	YscT	Type III secretory pathway, component EscT
24.800	21.288	114.022	354	T	B0HNK1	*Yersinia pestis biovar Antiqua str. B42003004*	YscU	Type III secretion appartus protein YscU
21.199	17.197	114.500	157	T	Q2SC35	*Hahella chejuensis (strain KCTC 2396)*	YscO	Putative type III secretion protein YscO
23.524	21.406	112.054	307	T	B0HZP6	*Yersinia pestis biovar Antiqua str. E1979001*	YscQ	Type III secretion apparatus protein YscQ
21.164	18.476	106.859	318	T	Q2SC33	*Hahella chejuensis (strain KCTC 2396)*	YscQ	Putative type III secretion protein YscQ
21.567	17.859	110.771	455	T	B0HZP7	*Yersinia pestis biovar Antiqua str. E1979001*	YscP	Type III secretion system needle length determinant YscP
24.256	21.250	109.410	90	T	Q2SC31	*Hahella chejuensis (strain KCTC 2396)*	YscS	Type III secretory pathway, component EscS
23.313	20.360	114.903	114	T	A6M3T7	*Yersinia pestis CA88-4125*	YscY	Putative type III secretion protein
21.853	18.000	111.435	221	T	A6M3R1	*Yersinia pestis CA88-4125*	YscL	Regulator of YscN ATPase
24.186	21.875	109.465	216	T	A3NKX4	*Burkholderia pseudomallei (strain 668)*	YscR	Type III secretion apparatus protein, YscR/HrcR family
23.176	20.201	112.344	137	T/C	B0HZP0	*Yersinia pestis biovar Antiqua str. E1979001*	YscB	YopN-specific chaperone YscB
21.825	16.696	110.013	115	T	B0HNJ2	*Yersinia pestis biovar Antiqua str. B42003004*	YscI	Type III secretion apparatus protein YscI
26.638	24.934	112.527	271	T	Q1MQX3	*Lawsonia intracellularis (strain PHE/MN1-00)*	YscT	Type III secretion component protein SctT
24.255	21.801	108.506	88	T	B0HZP4	*Yersinia pestis biovar Antiqua str. E1979001*	YscS	Type III secretion apparatus protein YscS
26.114	24.110	112.648	86	T	Q1MQX4	*Lawsonia intracellularis (strain PHE/MN1-00)*	YscS	Probable translocation protein in type III secretion
24.866	21.158	112.733	360	T	Q254G9	*Chlamydophila felis (strain Fe/C-56)*	YscU	Type III secretory flagellar biosynthesis Yop translocation protein U
24.255	21.801	108.506	88	T	B0HZP4	*Yersinia pestis biovar Antiqua str. E1979001*	YscS	Type III secretion apparatus protein YscS
22.252	18.163	111.146	89	T	Q2T727	*Burkholderia thailandensis*	YscF	Type III secretion system needle protein
23.422	19.651	110.529	700	T	Q1MQX5	*Lawsonia intracellularis (strain PHE/MN1-00)*	YscV	Putative type III secretion pore protein
21.825	16.696	110.013	115	T	B0HNJ2	*Yersinia pestis biovar Antiqua str. B42003004*	YscI	Type III secretion apparatus protein YscI
24.308	20.499	113.148	351	T	Q2SC29	*Hahella chejuensis (strain KCTC 2396)*	YscU	Type III secretory pathway, component EscU
20.598	16.591	107.774	115	T	A9R9K6	*Yersinia pestis bv. Antiqua (strain Angola)*	YscM	Type III secretion apparatus protein YscM
21.567	17.859	110.771	455	T	B0HZP7	*Yersinia pestis biovar Antiqua str. E1979001*	YscP	Type III secretion system needle length determinant YscP
22.294	18.393	109.614	908	T	Q252Q1	*Chlamydophila felis (strain Fe/C-56)*	YscC	Type III secretion general secretion pathway protein D
23.524	21.406	112.054	307	T	B0HZP6	*Yersinia pestis biovar Antiqua str. E1979001*	YscQ	Type III secretion apparatus protein YscQ
25.009	23.730	108.941	261	T	B0HZP3	*Yersinia pestis biovar Antiqua str. E1979001*	YscT	Type III secretion apparatus protein YscT
23.176	20.201	112.344	137	T/C	B0HZP0	*Yersinia pestis biovar Antiqua str. E1979001*	YscB	YopN-specific chaperone YscB

Table 3.18B: ASC of plant and animal Type III effectors (continued)

ASC Akashi	Craig & Weber	Avg. Mass	Total aa	Set	SwissProt Number	Organism	Protein Name	Description
24.800	21.288	114.022	354	T	B0HNK1	*Yersinia pestis biovar Antiqua str. B42003004*	YscU	Type III secretion appartus protein YscU
24.800	21.288	114.022	354	T	B0HNK1	*Yersinia pestis biovar Antiqua str. B42003004*	YscU	Type III secretion appartus protein YscU
22.895	19.674	110.756	244	T	B0HZN2	*Yersinia pestis biovar Antiqua str. E1979001*	YscJ	Type III secretion apparatus lipoprotein YscJ
23.655	18.971	123.230	154	T	B0HNK7	*Yersinia pestis biovar Antiqua str. B42003004*	YscO	Type III secretion protein YscO
25.537	23.396	112.485	217	T	B0HZP5	*Yersinia pestis biovar Antiqua str. E1979001*	YscR	Type III secretion apparatus protein YscR
20.598	16.591	107.774	115	T	A9R9K6	*Yersinia pestis bv. Antiqua (strain Angola)*	YscM	Type III secretion apparatus protein YscM
22.895	19.674	110.756	244	T	B0HZN2	*Yersinia pestis biovar Antiqua str. E1979001*	YscJ	Type III secretion apparatus lipoprotein YscJ
24.890	22.733	110.423	227	T	B0HZP5	*Yersinia pestis biovar Antiqua str. E1979001*	YscR	Type III secretion apparatus protein YscR
21.634	18.026	110.591	435	T	P68587	*Yersinia pestis. P68587 was used instead of P40295(obsolete)*	YscP*	Yop proteins translocation protein P
23.054	20.119	112.604	122	T	B0HZQ3	*Yersinia pestis biovar Antiqua str. E1979001*	YscX	Type III secretion protein YscX
24.850	22.316	110.742	215	T	Q2SC32	*Hahella chejuensis (strain KCTC 2396)*	YscR	Type III secretory pathway, component EscR
21.853	18.000	111.435	221	T	A6M3R1	*Yersinia pestis CA88-4125*	YscL	Regulator of YscN ATPase
24.656	22.194	114.704	209	T	A6M3R2	*Yersinia pestis CA88-4125*	YscK	Putative type III secretion protein
22.304	19.321	111.091	165	T	A9R9K1	*Yersinia pestis bv. Antiqua (strain Angola)*	YscH	Type III secretion protein YopR
25.009	23.730	108.941	261	T	B0HZP3	*Yersinia pestis biovar Antiqua str. E1979001*	YscT	Type III secretion apparatus protein YscT
24.255	21.801	108.506	88	T	B0HZP4	*Yersinia pestis biovar Antiqua str. E1979001*	YscS	Type III secretion apparatus protein YscS
23.524	21.406	112.054	307	T	B0HZP6	*Yersinia pestis biovar Antiqua str. E1979001*	YscQ	Type III secretion apparatus protein YscQ
21.825	16.696	110.013	115	T	B0HNJ2	*Yersinia pestis biovar Antiqua str. B42003004*	YscI	Type III secretion apparatus protein YscI
21.856	18.549	107.370	439	T	Q2SC36	*Hahella chejuensis (strain KCTC 2396)*	YscN	Putative Yops secretion ATP synthase YscN
23.280	21.455	110.495	123	C	P61380	*Yersinia pestis*	SycN	Chaperone protein sycN
25.028	21.492	116.655	132	C	Q9ZGW6	*Yersinia pestis*	SycT	Chaperone protein sycT
23.323	20.688	110.675	146	C	O34021	*Chlamydophila caviae*	SycE	Scc1 protein

Code: A = Animal Pathogens Set
P = Plant Symbiont Set
T = Type III Secretion System
C = TTSS Related Chaperone

Table 3.18C: ASC of various Autotransporters

Total aa	ASC	Mass/aa	Uniprot (updated)	Name	Uniprot (65)	Uniprot (Domain) Description	Amino acids	Description & Notes	Organism	% Savings α vs β ASC	% Savings α vs β Mass/aa
1045	21.898	107.490	P09489	Ssp	P09489	All	1-1045	Extracellular serine protease	*Serratia marcescens*		
618	21.956	106.792		Ssp		Extracellular serine protease	28-645			2.12	1.84
277	22.432	108.798		Ssp		Autotransporter	769-1045				
1036	21.377	103.842	Q54483	SSP-h1	BAA33455	All	1-1036	SSP-h1	*Serratia marcescens*		
1034	21.385	104.205	Q54484	SSP-h2	BAA11383	All	1-1034	SSP-h2	*Serratia marcescens*		
985	21.069	104.230	Q9ZNI6	PspA	BAA36466	All	1-985	Serine protease homologue	*Pseudomonas fluorescens*		
1036	21.242	103.061	Q9ZNI5	PspB	BAA36467	All	1-1036	Serine protease homologue	*Pseudomonas fluorescens*		
934	22.900	110.886	Q51846	Ssa1	AAA80490	All	1-934	Serotype-1-specific antigen	*Pasteurella haemolytica*		
1039	21.479	105.663	Q93L97	SphB1	CAC44081	All	1-1039	Autotransporter subtilisin-like protease	*Bordetella pertussis*		
1083	21.429	104.915	Q8GKS5	NaiP	AAN71715	All	1-1083	NalP	*Neisseria meningitidis*		
1296	22.143	107.826	Q48247	VacA	Q48247	All	1-1296	Vacuolating cytotoxin autotransporter	*Helicobacter pylori*		
989	21.927	107.621		VacA		Vacuolating cytotoxin	34-1022			4.39	0.67
274	22.935	108.346		VacA		Autotransporter	1023-1296				
1286	20.626	102.841	Q03155	AIDA-1	Q03155	All	1-1286	AIDA-I autotransporter	*Escherichia coli*		
797	19.441	99.755		AIDA-1		Adhesin AIDA-I	50-846			12.88	7.51
440	22.316	107.850		AIDA-1		AIDA-I translocator	847-1286				
1102	21.757	105.469	Q7BCK4	IcsA	AAA26547	All	1-1102	OM protein icsA autotransporter	*Shigella flexneri*		
706	20.648	102.633		IcsA		Outer membrane protein icsA	53-758			13.13	7.77
344	23.769	111.278		IcsA		Outer membrane protein icsA translocator	759-1102				
955	21.523	105.967	Q9Z625	MisL	AAD16954	All	1-955	MisL			
989	20.535	102.218	Q9XD84	TibA	AAD41751	All	1-989	Adhesin/invasin tibA	*Escherichia coli*		
623	19.363	98.760		TibA		Adhesin/invasin tibA	55-677			13.30	8.45
312	22.334	107.872		TibA		Adhesin/invasin tibA translocator	677-989				
1039	20.340	102.798	P39180	Ag43	P39180	All	1-1039	Antigen 43	*Escherichia coli (strain K12*		
499	19.235	99.738		Ag43		Antigen 43 alpha chain	53-551			9.13	5.48
488	21.167	105.516		Ag43		Antigen 43 beta chain	552-1039				
2035	19.832	101.727	Q9XCJ4	ShdA	AAD25110	All	1-2035	ShdA	*Salmonella typhimurium*		
619	22.999	110.449	Q9JPL5	AutA	CAB89117	All	1-619	Autotransporter A	*Neisseria meningitidis*		
1377	22.183	107.632	Q47692	Tsh	I54632	All	1-1377	Temp-sens hemagglutinin tsh autotransporter	*Escherichia coli*		
1048	21.713	106.705		Tsh		Temperature-sensitive hemagglutinin tsh	53-1100			7.09	3.32
277	23.370	110.368		Tsh		Temperature-sensitive hemagglutinin tsh translocator	1101-1377				
1364	22.067	107.038	Q8VSL2	SepA	CAC05786	All	1-1364	Serine protease sepA autotransporter	*Shigella flexneri*		
1033	21.575	105.858		SepA		Serine protease sepA	57-1089			7.57	3.93
275	23.343	110.191		SepA		Serine protease sepA translocator	1090-1386				
1305	22.281	107.865	Q9EZE7	EspC	AAC44731	All	1-1305	Serine protease espC	*Escherichia coli O127:H6*		
965	21.890	107.230		EspC		Secreted autotransporter protein espC	54-1018			6.53	2.87
287	23.420	110.397		EspC		Autotransporter protein espC translocator	1019-1305				

Table 3.18C: ASC of various Autotransporters (continued)

										% Savings α vs β	
Total aa	**ASC**	**Mass/aa**	**Uniprot (updated)**	**Name**	**Uniprot (65)**	**Uniprot (Domain) Description**	**Amino acids**	**Description & Notes**	**Organism**	**ASC**	**Mass/aa**
1300	22.726	109.031	Q7BSW5	EspP	CAA66144	All	1-1300	Serine protease espP	*Escherichia coli O157:H7*		
968	22.449	108.906		EspP		Secreted autotransporter protein espP	56-1023			3.91	0.46
277	23.362	109.410		EspP		Autotransporter protein espP translocator	1024-1300				
1295	22.258	107.916	O68900	Pet	AAC26634	All	1-1295	Serine protease pet autotransporter	*Escherichia coli O44:H18*		
966	21.868	107.589		Pet		Serine protease pet	53-1018			6.29	1.51
277	23.335	109.240		Pet		Serine protease pet translocator	1019-1295				
1372	22.209	106.732	Q7BS42	Pic	AAD23953	All	1-1372	Serine protease pic autotransporter	*Escherichia coli O44:H18*		
1040	21.710	105.553		Pic		Serine protease pic	56-1095			8.37	4.49
277	23.693	110.519		Pic		Serine protease pic translocator	1096-1372				
1285	22.634	108.683	Q9L8L1	SigA	AAF67320	All	1-1285	Exported serine protease SigA	*Shigella flexneri*		
1295	22.495	108.127	Q8FDW4	SatA	AAG30168	All	1-1295	Serine protease sat autotransporter	*Escherichia coli*		
969	22.304	108.052		SatA		Serine protease sat	50-1018			4.45	1.13
277	23.342	109.283		SatA		Serine protease sat translocator	1019-1295				
1377	22.298	107.678	Q6YGS9	Vat	AAO21903	All	1-1377	Vacuolating autotransporter toxin	*Escherichia coli*		
1359	22.521	108.404	Q79JW1	EpeA	AAL18821	All	1-1359	Autotransporter protease	*Escherichia coli*		
1364	22.514	108.269	Q84GK0	EatA	AAO17297	All	1-1364	Serine protease eatA	*Escherichia coli*		
1042	22.119	107.684		EatA		Secreted autotransporter protein eatA	57-1098			6.95	2.97
266	23.771	110.976		EatA		Autotransporter protein eatA translocator	1099-1364				
1363	22.239	107.199	Q93SE3	EspI	CAC39286	All	1-1363	Putative uncharacterized protein	*Escherichia coli*		
1335	21.984	106.112	Q9LA58	EaaA	AAF63237	All	1-1335	EaaA	*Enterobacterial phage P-EibA*		
1335	22.005	106.109	Q9LA54	EaaC	AAF63038	All	1-1335	EaaC	*Enterobacterial phage P-EibA*		
910	20.844	102.676	P14283	Pertactin	P14283	All	1-910	Pertactin autotransporter	*Bordetella pertussis*		
677	20.196	101.366		Pertactin		Outer membrane protein P.69	35-711			12.50	5.72
199	23.080	107.512		Pertactin		Pertactin translocator	712-910				
1010	20.537	102.336	Q45340	BrkA	AAA51646	All	1-1010	BrkA	*Bordetella pertussis*		
642	20.547	102.566	Q6U948	Tef	AAQ82668	All	1-644	Tracheal colonization factor protein	*Bordetella pertussis*		
915	20.942	103.707	O66044	Vag8	AAC31247	All	1-915	Vag8	*Bordetella pertussis*		
1531	21.467	104.984	O84818	PmpD	O84818	All	1-1531	Probable outer membrane protein pmpD	*Chlamydia trachomatis*		
1223	20.881	103.954		PmpD		Probable outer membrane protein pmpD	21-1243			12.75	4.85
288	23.932	109.255		PmpD		Autotransporter	1244-1531				
1723	21.299	104.223	Q9Z812	Pmp20	Q9Z812	All	1-1723	Probable outer membrane protein pmp20	*Chlamydia pneumoniae*		
1412	20.532	102.401		Pmp20		Probable outer membrane protein pmp20	22-1433			17.03	9.55
290	24.747	113.206		Pmp20		Autotransporter	1434-1723				
1609	21.775	106.183	Q9Z6U5	Pmp21	Q9Z6U5	All	1-1609	Probable outer membrane protein pmp21	*Chlamydia pneumoniae*		
1297	21.005	104.880		Pmp21		Probable outer membrane protein pmp21	31-1327			16.22	6.35
282	25.072	111.992		Pmp21		Autotransporter	1328-1609				

Table 3.18C: ASC of various Autotransporters (continued)

										% Savings α vs β	
Total aa	**ASC**	**Mass/aa**	**Uniprot (updated)**	**Name**	**Uniprot (65)**	**Uniprot (Domain) Description**	**Amino acids**	**Description & Notes**	**Organism**	**ASC**	**Mass/aa**
1552	21.875	110.301	Q9AIU6	Iga1	NP_283693	All	1-1552	IgA1 protease	*Neisseria meningitidis*		
1449	22.479	109.768	Q9F3Z5	App	CAC14670	All	1-1449	Adhesion penetration protein	*Neisseria meningitidis*		
1849	21.342	109.756	P45386	Iga1	P45386	All	1-1849	Immunoglobulin A1 protease autotransporter	*Haemophilus influenzae*		
996	22.522	110.587		Iga1		Immunoglobulin A1 protease	26-1021			-13.93	-1.79
828	19.768	108.637		Iga1		Immunoglobulin A1 protease translocator	1022-1849				
1394	22.758	111.494	P45387	Hap	P45387	All	1-1394	Adhesion & penetration protein autotransporter	*Haemophilus influenzae*		
1114	22.382	110.904		Hap		Adhesion and penetration protein	26-1139			6.96	2.80
255	24.057	114.101		Hap		Autotransporter	1140-1394				
2249	19.884	99.740	P15921	rOmpA	P15921	All	1-2249	Outer membrane protein A	*Rickettsia rickettsii*		
1932	19.304	98.059		rOmpA		Outer Membrane Protein A	29-1961			17.72	11.25
288	23.462	110.486		rOmpA		Autotransporter	1962-2249				
1654	20.530	101.673	Q53047	rOmpB	Q53047	All	1-1654	Outer membrane protein B	*Rickettsia rickettsii*		
1333	20.068	100.370		rOmpB		120 kDa surface-exposed protein	1-1333			12.19	7.36
293	22.855	108.344		rOmpB		32 kDa beta peptide	1362-1654				
656	21.781	106.469	O52756	ApeE	AAC38796	All	1-656	OM N-acetyl Phe β-naphthyl ester-cleaving esterase	*Salmonella typhimurium*		
646	22.076	107.726	O33407	EstA	AAB61674	All	1-646	Esterase estA - Not cleaved	*Pseudomonas aeruginosa*		
373	21.392	105.286		EstA		Extracellular	25-397				
281	22.725	110.640		EstA		Autotransporter	366-646				
645	23.098	109.610	P40601	Lip-1	P40601	All	1-645	Lipase 1	*Photorhabdus luminescens*		
358	21.682	106.840		Lip-1		Lipase I	25-382			13.13	6.13
263	24.958	113.817		Lip-1		Autotransporter	383-645				
650	22.532	109.947	Q7WVT7	McaP	AAP97134	All	1-650	McaP	*Moraxella catarrhalis*		
741	22.395	108.403	O52269	BabA	AAC38081	All	1-741	Adhesin binding fucosylated histo-blood group antigen	*Helicobacter pylori*		
651	22.998	110.831	Q9ZLB8	SabA	AAD06240	All	1-651		*Helicobacter pylori J99*		
518	22.573	108.314	O32641	AlpA	CAB05386	All	1-518	Adhesin	*Helicobacter pylori*		
886	21.435	113.534	Q83YP8	Aae	AAP21063	All	1-886	Autotransporter/adhesin Aae	*Aggregatibacter actinomycetemcomitans HK1651*		
1070	23.659	111.937	Q9EZV7	NanB	AAG35309	All	1-1070	Sialidase NanB			
									Average	8.72	4.55

Table 3.18D: ASC of various extracellular proteins of interest

Total aa	Avg Cost of Amino Acid	Economic Rank / Total	Ranking Organism	SwissProt Number	Organism	Protein Name	Description
185	20.58595	51 / 4333	MG1655	P04127	*Escherichia coli*	PapA	Pap fimbrial major pilin protein
182	19.52473	11 / 4333	same	P04128	*Escherichia coli* K-12	FimA	Type I pili major pilin protein
498	19.15482	6 / 4333	same	P04949	*Escherichia coli* K-12	FliC	Flagellum subunit
368	19.5663	13 / 4333	MG1655	B3SGP3	*Escherichia coli*	FlkA	Flagellum subunit - phase variable form
502	19.91474	23 / 4333	MG1655	B3SGN1	*Escherichia coli*	FlkA	Flagellum subunit - phase variable form
556	19.71115	16 / 4333	MG1655	B3SGN7	*Escherichia coli*	FlkA	Flagellum subunit - phase variable form
423	19.91773	23 / 4333	MG1655	Q75SX7	*Escherichia coli*	FlkA	Flagellum subunit - phase variable form
529	19.97108	25 / 4333	MG1655	Q75SY3	*Escherichia coli*	FlkA	Flagellum subunit - phase variable form
495	19.55818	13 / 4333	MG1655	P06179	*Salmonella typhimurium*	FliC	Flagellum subunit
507	19.46193	11 / 4333	MG1655	P52616	*Salmonella typhimurium*	FljB	Flagellum subunit - phase variable form
638	22.27445	1764 / 5567	*P. aeruginosa* PAO1	P11439	*Pseudomonas aeruginosa*	ToxA	Exotoxin A
453	20.61148	158 / 5567	*P. aeruginosa* PAO1	Q51451	*Pseudomonas aeruginosa*	ExoS	Exoenzyme S - Ranked 155 in PSORTDB
378	22.45503	2111 / 5567	*P. aeruginosa* PAO1	Q9I1S4	*Pseudomonas aeruginosa*	ExoY	Adenylate cyclase
687	21.24309	463 / 5567	*P. aeruginosa* PAO1	O34208	*Pseudomonas aeruginosa*	ExoU	*not in PSORTDB*
457	20.49256	135 / 5567	*P. aeruginosa* PAO1	Q9I788	*Pseudomonas aeruginosa*	ExoT	Exoenzyme T
479	21.67724	854 / 5567	*P. aeruginosa* PAO1	Q03023	*Pseudomonas aeruginosa*	AprA	Alkaline metalloproteinase
429	21.30256	501 / 5567	*P. aeruginosa* PAO1	Q9HXZ5	*Pseudomonas aeruginosa*	Eno	Enolase
311	21.71383	891 / 5567	*P. aeruginosa* PAO1	P26876	*Pseudomonas aeruginosa*	Lip	Lactonizing lipase
462	21.67619	852 / 5567	*P. aeruginosa* PAO1	Q9HWK6	*Pseudomonas aeruginosa*	PrpL	Lysyl endopeptidase
85	21.39882	588 / 5567	*P. aeruginosa* PAO1	P95434	*Pseudomonas aeruginosa*	PscF	Major component Type III needle structure
1072	21.17603	135 / 3885	*Yersinia pestis* CO92	Q9F288	*Yersinia pestis*	YapE	Type V secretion
170	21.20765	140 / 3885	*Yersinia pestis* CO92	P26948	*Yersinia pestis*	Caf1	F1 capsule antigen
147	19.37347	23 / 3737	same	B8H0U8	*Caulobacter crescentus* CB15N	HfaA	Holdfast attachment protein HfaA
415	18.4106	3 / 3737	same	B8H0V0	*Caulobacter crescentus* CB15N	HfaD	Holdfast attachment protein HfaD-not in PSORTDB
1073	19.15443	14 / 3737	same	B8H2X0	*Caulobacter crescentus* CB15N	RsaA	S-layer protein
1026	18.81969	24 / 3737	CB15N	P35828	*Caulobacter crescentus (Caulobacter vibrioides)*	RsaA	S-layer protein
862	21.58898	147 / 4112	*B. Subtilus* 168	P94217	*Bacillus anthracis*	Eag	S-layer protein EA1
939	20.2361	1 / 1838	*C. jejuni* RM1221	P35827	*Campylobacter fetus*	SapA	S-layer protein

Table 3.19: Comparison of the ASC of Type III secretion effectors with cytoplasmic proteins.

			E. coli K-12* Cytoplasm		*E. coli* K-12* Overall			*P. syringae pv. tomato*** Cytoplasm		*P. syringae pv. tomato*** Overall		
	ASC	**Num.**	**Z**	**P value**	**Z**	**P value**	**Savings**	**Z**	**P value**	**Z**	**P value**	**Savings**
Animal Type III Effectors†	22.233	64	5.835	5.39E-09	6.456	1.08E-10	1.32	4.064	4.81E-05	4.225	2.39E-05	0.81
Plant Type III Effectors†	21.74	38	7.181	6.90E-13	7.375	1.64E-13	1.81	5.878	4.16E-09	5.766	8.11E-09	1.31
Animal and Plant Type III Effectors†	22.049	102	8.928	4.35E-19	9.561	1.17E-21	1.51	6.722	1.79E-11	6.835	8.19E-12	1
Type III Secretion System, Chaperones†	23.468	65	***0.957***	***0.3385***	***0.163***	***0.8703***	0.09	2.764	5.71E-03	2.257	0.024037	-0.42
***E. coli* Type III Effectors†**	21.564	10	3.1	1.94E-03	3.254	1.14E-03	1.99	2.77	5.60E-03	2.743	6.09E-03	1.48
***E. coli&Salmonella* Type III Effectors†**	22.162	20	3.171	1.52E-03	3.539	4.02E-04	1.39	2.197	0.028041415	2.318	0.02046	0.88
***P. syringae sp.* Type III Effectors†**	21.74	38	7.181	6.90E-13	7.375	1.64E-13	1.81	5.878	4.16E-09	5.766	8.11E-09	1.31
***P. syringae pv tomato* HOPs**	21.757	44	7.966	1.67E-15	8.123	4.54E-16	1.8	6.679	2.40E-11	6.489	8.61E-11	1.29
***P. syringae pv tomato* HOP Helpers**	19.745	7	4.45	8.59E-06	4.424	9.67E-06	3.81	4.387	1.15E-05	4.303	1.69E-05	3.3
***P. syringae pv tomato* HOP w Helpers**	21.48	51	9.027	1.77E-19	9.17	4.73E-20	2.07	7.801	6.22E-15	7.612	2.70E-14	1.57

Table 3.19: Comparison of the ASC of Type III secretion effectors with cytoplasmic proteins (continued).

			E. coli K-12* Cytoplasm		*E. coli* K-12* Overall		*P. syringae pv. tomato*** Cytoplasm		*P. syringae pv. tomato*** Overall	
	Mean Mass	**Num.**	**Z**	**P value**	**Z**	**P value**	**Z**	**P value**	**Z**	**P value**
Animal Type III Effectors†	109.29	64	4.062	4.87E-05	2.817	4.84E-03	2.661	7.80E-03	***0.789***	***0.429864***
Plant Type III Effectors†	108.6	38	5.529	3.21E-08	4.43	9.43E-06	4.509	6.50E-06	2.695	7.03E-03
Animal and Plant Type III Effectors†	109.033	102	6.535	6.36E-11	4.906	9.28E-07	4.798	1.60E-06	2.259	0.0239
Type III Secretion System, Chaperone†	111.579	65	***0.787***	***0.431***	***0.827***	***0.408218***	***1.088***	***0.2765355***	3.525	4.23E-04
***E. coli* Type III Effectors†**	108.191	10	2.366	0.018	2	0.0455	2.041	0.0412	***1.447***	***0.148***
***E. coli&Salmonella* Type III Effectors†**	109.513	20	2.339	0.0193	***1.609***	***0.1076***	***1.596***	***0.1106***	***0.434***	***0.664***
***Pseudomonas* Type III Effectors†**	108.6	38	5.529	3.21E-08	4.43	9.43E-06	4.509	6.50E-06	2.695	7.03E-03
***P. syringae pv tomato* HOPs**	108.457	44	6.281	3.36E-10	5.165	2.40E-07	5.261	1.43E-07	3.423	6.20E-04
***P. syringae pv tomato*HOP Helpers**	102.782	7	4.192	2.77E-05	4.043	5.28E-05	4.108	3.98E-05	3.863	1.12E-04
***P. syringae pv tomato* HOP w Helpers**	107.678	51	7.368	1.73E-13	6.284	3.29E-10	6.384	1.73E-10	4.603	4.17E-06

		Num.	**ASC**	**Mean Mass**
E. coli* K-12	**Cytoplasm**	2859	23.317	111.609
E. coli* K-12	**Overall**	4333	23.555	111.005
P. syringae pv. tomato*	**Cytoplasm**	1973	22.961	110.982
P. syringae pv. tomato*	**Overall**	5608	23.047	110.075

***BOLD ITALIC* =** Not significant, P > 0.05

Sequences and locations: Echobase (67, 97)

** Sequences and locations: PSORTDB (122)

† From Arnold et al PLOS Pathology 5 (2009)

Table 3.20: Amino acid composition and cost savings in different compartments of *P. syringae*.

	Mean ATPs per aa	%A	%R	%N	%D	%C	%E	%Q	%G	%H	%I	%L	%K	%M	%F	%P	%S	%T	%W	%Y	%V	Total Amino Acids	Number
Overall Mean	23.05	10.55	6.45	3.09	5.36	1.13	5.66	4.48	7.47	2.33	5.03	11.20	4.00	2.58	3.67	4.75	6.14	5.07	1.41	2.49	7.13	329.67	5607
Overall Median	22.94	10.37	6.36	2.96	5.44	0.91	5.67	4.36	7.39	2.23	4.92	10.97	3.54	2.37	3.51	4.63	5.93	4.95	1.23	2.41	7.06	282.00	
Overall STDEVP	1.37	3.01	2.48	1.48	1.98	1.09	2.30	1.99	2.43	1.38	1.98	3.28	2.40	1.25	1.74	1.96	2.09	1.89	1.11	1.33	2.19	276.48	
Extracellular all	21.58	10.90	4.93	**4.89**	6.06	**0.60**	4.52	4.86	8.26	2.40	4.13	8.59	4.76	2.22	3.04	4.95	**9.12**	**6.61**	**0.85**	2.30	6.00	433.73	84
Extracellular - HOP	21.74	11.08	**4.00**	**4.85**	**6.76**	**0.69**	**4.03**	4.56	8.82	**1.71**	4.52	8.44	4.16	**1.88**	3.28	4.44	**8.26**	**7.75**	**1.22**	2.77	6.80	428.18	33
HOPs w Helpers	21.48	10.79	5.54	**4.91**	5.60	**0.55**	4.84	5.06	7.90	2.85	3.87	8.69	**5.16**	2.44	2.89	5.29	**9.67**	5.87	**0.61**	2.00	5.48	437.31	51
Outer Membrane	22.25	10.09	5.86	**4.96**	6.43	**0.34**	4.69	5.04	8.99	**1.35**	**3.72**	9.56	3.67	**1.55**	3.47	4.22	**7.84**	**6.45**	1.42	**3.84**	6.52	649.23	115
Periplasm	22.65	11.41	**4.18**	3.85	6.27	**0.63**	5.08	4.55	7.74	**1.50**	4.56	9.37	**6.54**	2.40	3.71	5.19	6.17	5.43	1.38	2.77	7.26	414.95	110
Inner Membrane	23.83	11.36	5.38	2.66	**3.63**	**0.81**	**3.90**	3.72	8.18	1.76	6.18	13.93	**2.80**	3.05	**4.59**	4.43	6.17	5.09	**1.90**	2.41	8.05	412.61	961
Unknown	22.91	10.73	6.46	3.21	5.42	1.20	5.43	4.85	7.28	2.32	4.55	10.56	4.28	2.59	3.50	5.03	6.51	5.31	1.46	2.56	6.74	282.73	2364
Cytoplasm	22.96	9.91	7.19	2.92	5.99	1.29	6.93	4.36	7.20	2.72	5.19	10.96	4.10	2.43	3.48	4.56	5.44	4.63	1.13	2.35	7.21	317.70	1973
Cost Savings (ATP)		15.66	0.06	12.66	14.66	2.66	12.06	11.06	15.66	-10.9	-4.94	0.06	-2.94	-6.94	-24.6	7.06	15.66	8.66	-46.9	-22.6	4.06		

Notes: **Bold** indicates >25% change in amino acid composition from the overall mean.

Underlined indicates >50% change in amino acid composition from the overall mean.

Amino acid percents are on a per protein basis. The percent amino acid composition of each protein in that location were averaged; they are unweighted for chain length. See Table 3.22B for weighted percents (% of all amino acids) of proteins in nonmembrane locations of *P. syringae*.

	Savings	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	Count
Extracellular-HOPs	**>1.5 ATPs**	**13.34**	-0.08	**13.86**	8.64	-0.95	-9.75	0.79	8.47	6.45	1.01	-0.08	-0.49	2.28	5.39	-5.29	**25.09**	**12.51**	**16.96**	2.07	-0.23	14
Extracellular-HOPs	**>0 ATPs**	**17.00**	-0.10	**13.25**	**12.36**	-0.48	-11.11	-3.07	**18.47**	2.37	1.13	-0.06	-0.35	3.52	7.54	-4.09	**21.40**	9.43	9.25	5.36	-1.82	23
HOPS and Helpers	**>1.5 ATPs**	**11.61**	-0.03	**10.16**	0.00	-0.93	-6.04	3.22	7.12	1.22	3.18	-0.06	-1.13	1.17	**11.20**	2.36	**25.68**	4.94	**18.70**	**10.16**	-2.53	24
HOPS and Helpers	**>0 ATPs**	-3.64	-0.05	**17.69**	1.88	-1.29	-5.84	6.24	0.45	-12.90	5.33	-0.12	-4.00	1.75	4.86	2.20	**53.36**	1.16	**38.54**	2.77	-8.39	47
Extraceullar All	**>1.5 ATPs**	**12.25**	-0.05	**11.53**	3.19	-0.94	-7.41	2.33	7.62	3.15	2.38	-0.07	-0.90	1.58	9.06	-0.46	**25.46**	7.73	**18.06**	7.18	-1.68	38
Extraceullar All	**>0 ATPs**	3.14	-0.06	**16.23**	5.32	-1.02	-7.57	3.18	6.37	-7.88	3.95	-0.10	-2.80	2.33	5.74	0.13	**42.86**	3.88	**28.92**	3.62	-6.23	70

Table 3.21: Comparison of the ASC of cellular and extracellular serine proteases.

					U-Test		Student T-test			Savings (ATP)	
	ASC	Mass	Number	Length	Z	P value	T	Deg. Free	P value	per aa	per protein
vs Cellular Serine proteases	22.547	109.500	5592	494.01							
All Extracelluar Serine proteases	22.157	107.408	191	737.93	4.700	2.60E-06	5.872	206	1.71E-08	0.390	287.97
Extracelluar Serine proteases†	21.827	106.596	82	897.34	6.474	9.52E-11	6.657	84	2.70E-09	0.720	646.09
vs *Bacillus sp.* Cellular Serine proteases	22.454	108.502	447	502.75							
***Bacillus sp.* Extracellular Serine proteases**	21.210	103.899	23	455.61	5.185	2.16E-07	6.558	25	7.19E-07	1.245	567.08
***Bacillus sp.* Extracellular Serine proteases†**	21.273	104.409	18	485.87	4.305	1.67E-05	5.244	19	4.62E-05	1.181	573.85
vs *E. coli* Cellular Serine proteases	23.126	110.089	192	401.68							
***E. coli* Extracellular Serine proteases**	22.381	107.912	10	1335.56	3.856	1.15E-04	6.581	64	1.01E-08	0.745	995.41
***E. coli* Extracellular Serine proteases †**	22.319	107.914	8	1335.60	3.822	1.33E-04	6.729	49	1.74E-08	0.807	1077.88
E. coli Extracellular Serine proteases ‡	21.994	107.306	10	1005.30	4.639	3.51E-06	8.443	30	2.02E-09	1.133	1138.73
E. coli Extracellular Serine proteases †,‡	21.973	107.316	8	1005.69	4.302	1.69E-05	8.461	23	1.62E-08	1.154	1160.14
vs *E. coli* K-12 Cytoplasm	23.555	111.005	4333	312.46							
***E. coli* Cellular Serine proteases**	23.126	110.089	192	401.68	6.214	5.16E-10	4.546	216	9.09E-06	0.429	172.14
***E. coli* Cellular Serine proteases †**	23.142	110.393	26	387.65	2.151	3.15E-02	1.478	25	0.152	0.412	159.88
***E. coli* Extracellular Serine proteases**	22.381	107.912	10	1335.56	3.299	9.70E-04	16.626	11	3.84E-09	1.174	1567.75
***E. coli* Extracellular Serine proteases†**	22.319	107.914	8	1335.60	3.005	2.66E-03	16.631	9	4.59E-08	1.236	1650.24
E. coli Extracellular Serine proteases ‡	21.994	107.306	10	1005.30	4.108	3.99E-05	15.490	10	2.57E-08	1.561	1569.54
E. coli Extracellular Serine proteases †,‡	21.973	107.316	8	1005.69	3.720	1.99E-04	15.259	8	3.37E-07	1.582	1591.13

Note: Cellular serine proteases are serine proteases from Uniprot that are not secreted.

† Homologs (Genes with same name and similar ASC, lengths, sequences) were averaged.

‡ All the *E. coli* secreted serine proteases were autotransporters. Just the secreted domain was analyzed here.

Table 3.22A: Effects of protein structure on ASC in *E. coli*.

Mean values based on per protein amounts (unweighted for length)

	Mean				Filtered Database			Neural Networks only			Unfiltered Database†		
	Count	Mass/aa	ATPs/aa	Length	%H	%E	%C	%H	%E	%C	%H	%E	%C
Overall	3212	111.38	23.26	297.75	39.82	17.54	42.64	41.41	17.80	40.79	33.91	21.14	44.95
Cytoplasmic	2859	111.61	23.32	296.47	40.58	17.08	42.33	42.40	17.27	40.33	ND	ND	ND
Periplasmic	337	109.72	22.85	309.42	33.85	20.96	45.18	33.47	22.00	44.53	34.13	20.88	45.00
Extracellular	16	104.90	20.65	280.44	29.29	26.64	44.07	31.14	24.82	44.04	29.38	26.72	43.90

Linear Correlation	Filtered Database						Neural Netwoks only						Unfiltered Database†					
vs ASC (ATPs/aa)	%H		%E		%C		%H		%E		%C		%H		%E		%C	
	R	P value	R	P value	R	P value	R	P value	R	P value	R	P value	R	P value	R	P value	R	P value
Overall	0.0408	0.0208	-0.0291	0.0996	-0.0360	0.0412	-0.0010	0.9527	0.0164	0.3514	-0.0174	0.3249	0.0167	0.3440	-0.0172	0.3290	-0.0072	0.6815
Cytoplasmic	0.0219	0.2408	0.0011	0.9552	-0.0379	0.0427	-0.0311	0.0964	0.0501	**0.0073**	-0.0023	0.9003	ND	ND	ND	ND	ND	ND
Periplasmic	-0.0102	0.8518	-0.0622	0.2547	0.0939	0.0852	-0.0300	0.5826	-0.0184	0.7359	0.0798	0.1439	-0.0121	0.8255	-0.0619	0.2572	0.0963	0.0776
Extracellular	0.2177	0.4179	0.0324	0.9053	-0.4340	0.0931	0.3288	0.2138	-0.0063	0.9816	-0.5458	0.0287	0.2022	0.4528	0.0599	0.8255	-0.4339	0.0931

Notes: ND - Not Done **BOLD:** P < 0.01 **† Overall contains only data from Periplasmic and Extracellular proteins**

Comparing Protein Secondary Structure between Locations

	Filtered Database				Neural Networks Only			
	Student T*		Mann Whitney		Student T*		Mann Whitney	
vs Cytoplasmic	T value	P value	Z	P value	T value	P value	Z	P value
%H Periplasmic	6.622	1.16E-10	7.199	6.05E-13	9.639	1.08E-19	10.347	4.32E-25
%E Periplasmic	5.317	1.79E-07	4.551	5.33E-06	7.341	1.21E-12	7.734	1.04E-14
%C Periplasmic	4.768	2.62E-06	5.507	3.65E-08	8.383	8.15E-16	9.792	1.22E-22
%H Extracellular	2.023	**0.061**	2.487	0.013	2.695	0.017	3.058	0.002
%E Extracellular	2.212	0.043	2.108	0.035	2.305	0.036	2.628	0.009
%C Extracellular	0.557	**0.586**	0.656	**0.512**	1.498	**0.155**	1.348	**0.178**

Notes: Comparisons are versus same % secondary structure in cytoplasmic proteins

* Assumes unequal variance, two tailed. **BOLD –** P > 0.01. H-Helix, E- Strand, C-Coil.

Table 3.22A: Effects of protein structure on ASC in *E. coli* (continued).

Filtered Database (Percent of all amino acids in the Secondary Structure of each location)

		%A	%C	%D	%E	%F	%G	%H	%I	%K	%L	%M	%N	%P	%Q	%R	%S	%T	%V	%W	%Y
Overall	**Helix**	12.77	1.06	4.77	8.46	3.47	3.06	2.23	5.71	13.09	5.21	2.92	3.02	2.24	6.03	7.11	4.63	4.14	5.79	1.51	2.76
	Coil	7.37	1.32	7.75	5.68	2.74	12.23	2.91	3.24	6.23	4.86	2.40	5.73	7.80	4.07	5.20	6.97	5.91	4.29	1.01	2.29
	Strand	6.72	1.70	2.57	4.31	5.42	3.58	2.32	11.40	12.37	3.31	2.60	2.01	1.38	2.96	5.06	4.26	6.51	15.38	1.82	4.31
Cytoplasm	**Helix**	12.63	1.09	4.76	8.74	3.45	3.02	2.33	5.78	13.13	5.09	2.92	2.95	2.23	5.97	7.33	4.52	4.06	5.79	1.50	2.74
	Coil	7.25	1.38	7.69	5.86	2.79	12.33	3.04	3.29	6.38	4.68	2.43	5.51	7.79	3.99	5.40	6.86	5.79	4.30	0.98	2.26
	Strand	6.68	1.86	2.53	4.46	5.43	3.53	2.43	11.71	12.58	3.20	2.63	1.84	1.37	2.81	5.18	4.07	6.27	15.54	1.76	4.13
Extracellular	**Helix**	12.79	0.33	5.53	3.73	3.00	3.26	0.40	5.93	13.32	5.53	1.73	6.53	0.60	9.13	2.93	9.46	7.79	4.86	0.67	2.47
	Coil	12.01	0.46	7.90	2.98	1.74	13.44	0.46	2.46	4.41	4.05	1.64	9.65	4.82	3.85	1.64	10.83	11.29	4.82	0.31	1.23
	Strand	9.74	0.29	3.47	2.22	3.66	4.05	1.06	8.20	9.74	3.09	1.25	3.66	1.54	7.62	2.60	6.17	12.54	13.60	0.77	4.73
Periplasm	**Helix**	14.22	0.86	4.75	5.94	3.74	3.41	1.24	5.04	12.76	6.48	3.01	3.49	2.45	6.52	5.18	5.53	4.83	5.82	1.72	3.01
	Coil	8.08	0.90	8.20	4.50	2.37	11.47	2.05	2.85	5.17	6.28	2.21	7.17	8.03	4.64	3.82	7.67	6.61	4.16	1.26	2.56
	Strand	6.83	0.70	2.79	3.44	5.41	3.85	1.65	9.48	11.13	4.07	2.48	3.09	1.49	3.75	4.38	5.44	7.83	14.42	2.25	5.51

Table 3.22A: Effects of protein structure on ASC in *E. coli* (continued).

Average Cost		Filtered Database		Neural Networks		Unfiltered Database†	
		ATPs/aa	% total	ATPs/aa	% total	ATPs/aa	% total
Overall	**Helix**	23.904	39.37	24.017	41.21	23.608	32.90
	Coil	21.420	43.41	20.871	41.63	20.902	46.59
	Strand	26.792	17.23	27.607	17.16	26.463	20.50
Cytoplasm	**Helix**	23.936	40.22	24.055	42.30	ND	ND
	Coil	21.490	42.98	20.949	41.08	ND	ND
	Strand	26.844	16.80	27.627	16.63	ND	ND
Extracellular	**Helix**	21.797	33.45	21.746	32.81	21.831	33.41
	Coil	18.255	43.44	17.875	43.37	18.188	43.10
	Strand	23.623	23.11	24.320	23.82	23.616	23.49
Periplasm	**Helix**	23.679	32.67	23.708	32.68	23.685	32.68
	Coil	21.019	46.88	20.426	46.09	21.009	46.09
	Strand	26.599	20.45	27.638	21.22	26.604	21.22

Percent Cost of		vs Cytoplasmic		vs Overall	
		Filtered	NN	Filtered	NN
Extracellular	**Helix**	91.06	90.40	91.19	90.55
	Coil	84.94	85.33	85.22	85.64
	Strand	88.00	88.03	88.17	88.09
Periplasm	**Helix**	98.93	98.56	99.06	98.71
	Coil	97.81	97.50	98.13	97.87
	Strand	99.09	100.04	99.28	100.11

Notes: Values on this page are based on the total amino acids in all of that secondary structure of every protein in that location. Values in the table on the page above are from averaging % secondary structure in each protein.
ND – not done. NN – Neural Networks.

† Overall contains only data from Periplasmic and Extracellular proteins

Table 3.22B: Effects of protein structure on ASC in *P. syringae*.

Mean values based on per protein amounts (unweighted for length)

	Mean				Filtered Database			Neural Networks only			Unfiltered Database†		
	Count	Mass/aa	ATPs/aa	Length	%H	%E	%C	%H	%E	%C	%H	%E	%C
Overall	2167	110.72	22.89	327.13	40.80	16.50	42.70	43.41	15.98	40.61	36.02	15.88	48.10
Cytoplasmic	1973	110.98	22.96	317.70	41.28	16.56	42.16	44.13	15.90	39.97	ND	ND	ND
Periplasmic	110	108.72	22.65	414.95	36.86	18.37	44.78	37.22	17.86	44.92	36.88	18.36	44.75
Extracellular	84	107.30	21.58	433.73	34.68	12.64	52.67	34.60	15.37	50.04	34.88	12.62	52.49
HOPS&Helpers	51	107.68	21.48	437.31	41.31	8.03	50.65	38.23	10.89	50.88	41.57	8.03	50.40
Extracelluar-HOPS	33	106.70	21.74	428.18	24.43	19.77	55.80	28.98	22.28	48.74	24.54	19.73	55.73

Linear Correlation	Filtered Database						Neural Netwoks only						Unfiltered Database†					
vs. ASC (ATPs/aa)	%H		%E		%C		%H		%E		%C		%H		%E		%C	
	R	P value	R	P value	R	P value	R	P value	R	P value	R	P value	R	P value	R	P value	R	P value
Overall	-0.1329	**5.21E-10**	0.1340	**3.84E-10**	0.0825	**0.0001**	-0.1055	**8.57E-07**	0.1363	**1.90E-10**	0.0376	**7.98E-02**	0.0196	0.3616	0.0071	0.7421	-0.0331	0.1239
Cytoplasmic	-0.1717	**1.60E-14**	0.1259	**2.04E-08**	0.1605	**7.33E-13**	-0.1484	**3.45E-11**	0.1432	**1.64E-10**	0.1069	**1.96E-06**	ND	ND	ND	ND	ND	ND
Periplasmic	0.0346	0.7196	-0.1271	0.1858	0.0619	0.5205	-0.0466	0.6290	-0.0256	0.7909	0.1126	0.2416	0.0275	0.7758	-0.1280	0.1828	0.0707	0.4628
Extracellular	-0.0908	0.4116	0.1980	0.0710	-0.0434	0.6949	-0.1792	0.1029	0.1871	0.0884	0.0705	0.5238	-0.1008	0.3615	0.2057	0.0605	-0.0362	0.7435
HOPS&Helpers	0.1766	0.2152	0.2297	0.1049	-0.3477	0.0124	0.0621	0.6650	0.1996	0.1603	-0.2331	0.0998	0.1668	0.2420	0.2245	0.1132	-0.3377	0.0154
Extracelluar-HOPS	-0.2626	0.1398	0.1538	0.3929	0.2212	0.2161	-0.3455	0.0489	0.1440	0.4241	0.4587	**0.0073**	-0.2704	0.1280	0.1713	0.3405	0.2178	0.2235

Notes: ND - Not Done **BOLD:** P < 0.01 **† Overall contains only data from Periplasmic and Extracellular proteins**

Comparing Protein Secondary Structure between Locations

	Filtered Database				Neural Networks Only			
	Student T*		Mann Whitney		Student T*		Mann Whitney	
vs Cytoplasmic	T value	P value	Z	P value	T value	P value	Z	P value
H Periplasmic	3.510	6.21E-04	2.538	1.12E-02	6.302	4.22E-09	5.249	1.53E-07
E Periplasmic	2.356	0.020	2.126	0.034	2.682	8.29E-03	2.705	0.007
C Periplasmic	3.046	0.003	2.243	0.025	7.983	6.56E-13	6.670	2.55E-11
H Extracellular	3.410	0.001	2.721	0.007	6.154	2.00E-08	5.509	3.60E-08
E Extracellular	3.328	0.001	4.377	1.20E-05	0.478	**0.634**	1.436	**0.151**
C Extracellular	8.029	4.32E-12	8.232	2.22E-16	10.689	1.23E-17	9.724	2.37E-22
H HOPs	0.017	**0.987**	0.547	**0.584**	3.629	6.31E-04	3.075	0.002
E HOPs	6.334	2.93E-10	6.580	4.70E-11	5.447	1.23E-06	4.499	6.83E-06
C HOPs	5.469	1.31E-06	5.495	3.92E-08	9.032	3.10E-12	8.055	7.93E-16
H Extracellular - HOPs	5.417	5.38E-06	5.079	3.80E-07	5.528	3.88E-06	5.061	4.18E-07
E Extracellular - HOPs	1.508	**0.141**	1.146	**0.252**	3.337	0.002	5.652	1.59E-08
C Extracellular - HOPs	8.648	1.05E-17	6.438	1.21E-10	5.988	1.00E-06	3.297	0.001

Notes: Comparisons are versus same % secondary structure in cytoplasmic proteins * Assumes unequal variance, two tailed. **BOLD –** P > 0.01. H-Helix, E- Strand, C-Coil.

Table 3.22B: Effects of protein structure on ASC in *P. syringae* (continued).

		Filtered Database (Percent of all amino acids in the Secondary Structure of each location)																			
		%A	%C	%D	%E	%F	%G	%H	%I	%K	%L	%M	%N	%P	%Q	%R	%S	%T	%V	%W	%Y
Overall	**Helix**	13.61	0.95	4.98	8.44	3.35	3.57	2.34	5.04	13.96	4.51	2.45	2.31	2.33	5.65	8.14	4.68	3.83	6.17	1.26	2.45
	Coil	8.36	1.13	8.12	5.61	2.72	12.52	2.87	2.96	6.79	4.11	2.22	4.40	8.32	3.92	5.91	7.25	5.46	4.57	0.84	1.95
	Strand	6.76	1.67	2.75	4.47	5.59	3.61	2.54	10.66	13.58	3.02	2.25	1.64	1.47	3.09	5.83	4.17	5.41	16.19	1.56	3.75
Cytoplasm	**Helix**	13.46	1.00	4.96	8.67	3.35	3.54	2.41	5.10	14.11	4.29	2.43	2.23	2.31	5.60	8.42	4.51	3.73	6.19	1.25	2.45
	Coil	8.11	1.23	8.10	5.82	2.76	12.57	3.01	3.02	6.96	3.95	2.28	4.17	8.30	3.85	6.19	7.02	5.33	4.57	0.83	1.94
	Strand	6.70	1.78	2.76	4.58	5.46	3.57	2.61	10.82	13.78	2.90	2.21	1.54	1.46	3.04	6.00	4.07	5.17	16.36	1.52	3.67
Periplasm	**Helix**	15.45	0.50	5.28	6.46	3.52	3.89	1.33	4.58	12.51	7.11	2.53	2.78	2.85	5.73	5.35	5.41	4.38	5.97	1.66	2.72
	Coil	9.84	0.51	8.53	4.59	2.79	11.77	1.80	2.57	5.69	6.08	2.10	5.25	8.97	4.45	3.90	7.46	5.70	4.62	1.03	2.36
	Strand	7.52	0.83	2.39	3.93	6.49	4.25	1.57	9.12	11.48	4.38	2.69	2.28	1.86	3.37	4.40	4.79	6.78	15.60	2.15	4.12
Extracellular	**Helix**	14.37	0.51	4.89	6.24	3.03	3.83	2.30	4.38	12.64	5.71	2.76	3.30	1.96	6.49	6.03	7.31	5.20	5.99	0.94	2.09
	Coil	10.06	0.42	7.90	3.87	2.10	12.70	2.18	2.54	5.69	4.17	1.57	6.47	7.86	4.16	4.31	10.18	6.95	4.45	0.70	1.72
	Strand	6.75	0.78	3.09	3.05	6.84	3.38	2.51	9.74	12.70	3.40	2.40	2.73	0.93	3.59	4.59	5.37	8.31	13.46	1.43	4.96
HOPS	**Helix**	13.97	0.49	4.97	6.52	2.98	3.65	2.62	4.07	12.63	6.12	3.02	2.98	2.13	6.50	6.49	7.27	4.77	6.00	0.80	2.04
	Coil	9.60	0.35	7.01	4.08	1.98	12.14	2.63	2.72	5.70	4.48	1.66	6.44	8.84	4.36	4.78	10.92	6.31	4.14	0.53	1.33
	Strand	7.20	0.93	2.24	3.27	7.14	3.22	3.65	9.05	13.30	3.65	2.34	2.45	0.93	3.71	6.27	4.85	6.54	14.50	0.98	3.76
Extracellular - HOPs	**Helix**	15.60	0.56	4.66	5.42	3.17	4.40	1.35	5.32	12.69	4.49	1.98	4.23	1.45	6.48	4.66	7.44	6.48	5.98	1.39	2.25
	Coil	10.69	0.52	9.13	3.58	2.27	13.48	1.56	2.28	5.66	3.75	1.44	6.52	6.52	3.88	3.66	9.17	7.82	4.87	0.94	2.26
	Strand	6.46	0.68	3.66	2.91	6.64	3.48	1.76	10.19	12.31	3.23	2.44	2.91	0.93	3.52	3.48	5.71	9.47	12.77	1.72	5.74

Table 3.22B: Effects of protein structure on ASC in *P. syringae* (continued).

Average Cost		Filtered Database		Neural Networks		Unfiltered Database†	
		ATPs/aa	% total	ATPs/aa	% total	ATPs/aa	% total
Overall	**Helix**	23.549	40.58	23.691	43.08	22.993	34.90
	Coil	21.138	43.49	20.640	41.58	20.359	50.03
	Strand	26.629	15.93	27.397	15.34	26.658	15.07
Cytoplasm	**Helix**	23.610	41.35	23.752	44.02	ND	ND
	Coil	21.258	42.62	20.751	40.72	ND	ND
	Strand	26.625	16.03	27.429	15.26	ND	ND
Periplasm	**Helix**	23.282	35.95	23.500	36.97	23.287	35.98
	Coil	20.852	46.93	20.373	46.46	20.859	46.96
	Strand	26.590	17.12	27.487	16.57	26.577	17.06
Extracellular	**Helix**	22.601	33.15	22.626	34.584	22.597	33.55
	Coil	19.821	54.17	19.409	50.295	19.813	53.87
	Strand	26.785	12.68	26.713	15.121	26.796	12.58
HOPS	**Helix**	22.664	40.58	22.583	39.21	22.662	41.20
	Coil	19.819	51.19	19.657	50.25	19.809	50.68
	Strand	26.748	8.22	26.667	10.54	26.691	8.12
Extracellular - HOPs	**Helix**	22.413	21.42	22.724	27.28	22.402	21.48
	Coil	19.824	58.86	19.018	50.37	19.819	58.90
	Strand	26.810	19.72	26.746	22.36	26.864	19.62

Percent Cost of		vs Cytoplasmic		vs Overall	
		Filtered	NN	Filtered	NN
Periplasm	**Helix**	98.61	98.94	98.87	99.19
	Coil	98.09	98.18	98.65	98.71
	Strand	99.87	100.21	99.85	100.33
Extracellular	**Helix**	95.73	95.26	95.97	95.50
	Coil	93.24	93.53	93.77	94.03
	Strand	100.60	97.39	100.59	97.50
HOPS	**Helix**	95.99	95.08	96.24	95.32
	Coil	93.23	94.73	93.76	95.24
	Strand	100.46	97.22	100.45	97.34
Extracellular - HOPs	**Helix**	94.93	95.67	95.18	95.92
	Coil	93.25	91.65	93.78	92.14
	Strand	100.69	97.51	100.68	97.62

Notes: Values on this page are based on the total amino acids in all of that secondary structure of every protein in that location. Values in the table on the page above are from averaging % secondary structure in each protein. ND – not done. NN – Neural Networks.

† Overall contains only data from Periplasmic and Extracellular proteins

Table 3.23: Disorder prediction and ASC in nonmembrane proteins of *E. coli* and *P. syringae*.

E. coli **Mean values vased on per protein amounts (unweighted for length)**

	Count	Mass/aa	ATPs/aa	Length	% Ordered	% Disordered
Overall	3212	111.38	23.26	297.75	94.25	5.75
Cytoplasmic	2859	111.61	23.32	296.47	94.50	5.50
Periplasmic	337	109.72	22.85	309.42	92.39	7.61
Extracellular	16	104.90	20.65	280.44	88.81	11.19

E. coli **Linear Correlation of ASC (ATPs/aa) and…**

	% Ordered		% Disordered	
	R	P value	R	P value
Overall	0.1122	1.81E-10	-0.1122	1.81E-10
Cytoplasmic	0.0719	1.20E-04	-0.0719	1.20E-04
Periplasmic	0.2597	1.35E-06	-0.2597	1.35E-06
Extracellular	0.3342	**0.2058**	-0.3342	**0.2058**

P. syringae **Mean values vased on per protein amounts (unweighted for length)**

	Count	Mass/aa	ATPs/aa	Length	% Ordered	% Disordered
Overall	2167	110.72	22.89	327.13	93.70	6.30
Cytoplasmic	1973	110.98	22.96	317.70	94.07	5.93
Periplasmic	110	108.72	22.65	414.95	95.62	4.38
Extracellular	84	107.30	21.58	433.73	82.63	17.37
HOPS&Helpers†	51	107.68	21.48	437.31	77.13	22.87
Extracellular-HOPS	33	106.70	21.74	428.18	91.14	8.86

P. syringae **Linear Correlation of ASC (ATPs/aa) and…**

	% Ordered		% Disordered	
	R	P value	R	P value
Overall	0.1762	1.43E-16	-0.1762	1.43E-16
Cytoplasmic	0.0648	0.0040	-0.0648	0.0040
Periplasmic	0.4273	3.23E-06	-0.4273	3.23E-06
Extracellular	0.4987	1.38E-06	-0.4987	1.38E-06
HOPS&Helpers†	0.5896	5.29E-06	-0.5896	5.29E-06
Extracellular-HOPS	0.4271	**0.0132**	-0.4271	**0.0132**

Table 3.23: Disorder prediction and ASC in nonmembrane proteins of *E. coli* and *P. syringae* (continued).

E. coli	Comparing composition (Filtered Database)			
	Student T*		Mann Whitney	
vs Cytoplasmic	**T value**	**P value**	**Z**	**P value**
Periplasmic				
% Ordered	3.518	4.87E-04	4.337	1.44E-05
% Disordered	3.518	4.87E-04	4.337	1.45E-05
Extracellular				
% Ordered	1.020	**0.324**	0.099	**0.921**
% Disordered	1.020	**0.324**	0.099	**0.921**

P. syringae	Comparing composition (Filtered database)			
	Student T*		Mann Whitney	
vs Cytoplasmic	**T value**	**P value**	**Z**	**P value**
Periplasmic				
% Ordered	1.865	**0.0647**	3.989	6.62E-05
% Disordered	1.865	**0.0647**	3.990	6.62E-05
Extracellular				
% Ordered	5.814	1.07E-07	6.836	8.15E-12
% Disordered	5.814	1.07E-07	6.836	8.16E-12
HOPS†				
% Ordered	8.029	1.31E-10	9.615	6.93E-22
% Disordered	8.029	1.31E-10	9.615	6.93E-22
Extracellular - HOPS				
% Ordered	0.884	**0.3832**	0.962	**0.3360**
% Disordered	0.884	**0.3832**	0.962	**0.3358**

Notes: **BOLD –** P > 0.01 * Assumes unequal variance, two tailed.

† The N terminus (non cleaved signal sequence) of Type III effectors is typically disordered(50);
HOPS have higher % disorder in the rest of their proteins.
Removing the first 30 aa from the HOP proteins changed the % of ordered and disordered regions to 81.36% and 18.64% respectively.

Table 3.23: Disorder prediction and ASC in nonmembrane proteins of *E. coli* and *P. syringae* (continued).

Filtered Database (Percent of all amino acids in the (dis)ordered region of each location)

E. coli		%A	%C	%D	%E	%F	%G	%H	%I	%K	%L	%M	%N	%P	%Q	%R	%S	%T	%V	%W	%Y	ASC	% total
Overall	**Disordered**	9.82	0.44	4.96	7.93	1.42	5.54	2.43	2.71	5.29	8.16	6.98	4.77	5.78	6.87	7.14	8.17	5.95	4.17	0.35	1.12	21.326	4.29
	Ordered	9.36	1.32	5.71	6.48	3.58	7.20	2.55	5.75	10.20	4.58	2.45	3.99	4.45	4.55	5.87	5.47	5.29	6.91	1.39	2.90	23.413	95.71
Cytoplasm	**Disordered**	9.72	0.47	4.97	8.35	1.37	5.46	2.49	2.75	5.23	7.93	6.95	4.65	5.77	6.77	7.53	8.08	5.81	4.20	0.35	1.14	21.344	4.07
	Ordered	9.30	1.38	5.67	6.71	3.59	7.18	2.66	5.83	10.34	4.45	2.48	3.83	4.42	4.50	6.08	5.34	5.15	6.90	1.36	2.83	23.459	95.93
Periplasm	**Disordered**	10.23	0.32	4.82	5.75	1.74	6.11	2.21	2.44	5.41	9.71	7.39	5.48	5.70	7.34	5.30	8.42	6.23	3.95	0.35	1.11	21.338	5.78
	Ordered	9.81	0.88	6.03	4.69	3.54	7.35	1.68	5.07	9.08	5.66	2.22	5.11	4.82	4.93	4.32	6.40	6.28	6.97	1.69	3.45	23.133	94.22
Extracellular	**Disordered**	11.31	0.23	6.11	4.52	1.36	3.62	0.45	2.49	8.14	5.20	4.07	4.30	7.92	7.47	2.26	11.76	12.90	5.20	0.45	0.23	19.763	9.85
	Ordered	11.79	0.40	6.08	2.89	2.74	8.33	0.59	5.22	8.68	4.23	1.31	7.54	2.08	6.38	2.30	9.02	10.14	7.05	0.54	2.69	20.780	90.15
P. syringae																							
Overall	**Disordered**	11.95	0.40	5.91	7.22	1.39	6.17	2.21	2.42	5.93	5.39	5.23	3.58	8.27	5.83	7.63	9.08	5.47	4.84	0.26	0.82	20.540	5.73
	Ordered	10.13	1.18	5.99	6.54	3.56	7.55	2.63	5.19	11.07	4.02	2.14	3.08	4.58	4.41	6.75	5.51	4.75	7.20	1.18	2.54	23.140	94.27
Cytoplasm	**Disordered**	11.47	0.48	6.00	7.78	1.40	5.96	2.30	2.39	6.11	5.31	5.92	3.39	7.91	5.51	8.25	8.49	5.37	4.92	0.27	0.78	20.811	5.00
	Ordered	10.02	1.26	5.94	6.75	3.55	7.47	2.72	5.27	11.27	3.85	2.14	2.93	4.56	4.39	7.02	5.35	4.60	7.25	1.16	2.52	23.211	95.00
Periplasm	**Disordered**	15.14	0.12	6.11	6.35	1.72	6.17	1.17	2.56	5.39	6.98	3.94	2.83	10.90	7.34	4.85	7.28	4.73	5.18	0.21	1.02	20.090	7.28
	Ordered	11.17	0.60	6.33	5.06	3.84	7.77	1.62	4.56	9.43	6.09	2.23	3.93	5.13	4.52	4.48	6.19	5.46	7.12	1.55	2.93	22.914	92.72
Extracellular	**Disordered**	12.71	0.13	5.33	4.78	1.14	7.30	2.30	2.52	5.28	4.91	2.32	5.05	8.69	6.66	5.89	13.17	6.44	4.24	0.24	0.89	19.364	16.32
	Ordered	10.75	0.56	6.48	4.51	3.38	8.83	2.25	4.36	9.58	4.52	2.02	4.92	4.31	4.51	4.72	7.73	6.56	6.47	1.00	2.52	22.067	83.68
HOPS	**Disordered**	12.03	0.14	4.85	4.94	1.12	7.76	2.52	2.62	5.24	4.90	2.03	4.96	9.04	6.82	6.24	13.35	6.40	3.90	0.28	0.85	19.417	22.76
	Ordered	10.92	0.55	6.07	5.02	3.31	8.02	2.76	4.13	10.29	5.13	2.33	4.63	4.41	4.69	5.40	7.64	5.50	6.29	0.80	2.11	22.170	77.24
Extracellular - HOPs	**Disordered**	16.69	0.12	8.17	3.80	1.27	4.60	1.04	1.96	5.52	4.95	4.03	5.52	6.67	5.75	3.80	12.08	6.67	6.21	0.00	1.15	19.057	6.15
	Ordered	10.53	0.59	7.02	3.85	3.46	9.89	1.59	4.66	8.67	3.73	1.61	5.30	4.18	4.28	3.84	7.86	7.93	6.70	1.27	3.06	21.933	93.85

Percent Cost of *E. coli*

		vs Cytoplasmic	vs Overall
Periplasm	**Disordered**	99.98	100.06
	Ordered	98.61	98.81
Extracellular	**Disordered**	92.59	92.67
	Ordered	88.58	88.76

Percent Cost of *P. syringae*

		vs Cytoplasmic	vs Overall
Periplasm	**Disordered**	96.54	97.81
	Ordered	98.72	99.02
Extracellular	**Disordered**	93.05	94.27
	Ordered	95.07	95.36
HOPS	**Disordered**	93.30	94.53
	Ordered	95.52	95.81
Extracellular - HOPs	**Disordered**	91.57	92.78
	Ordered	94.49	94.78

Notes: Values on this page are based on the total amino acids in all the (dis)ordered region of every protein in that location. Values in the table on the pages above are from averaging the % (dis)order in each protein.

Table 3.24: Amino acid composition of extracellular flagellar subunits in endosymbionts & other *Gammaproteobacteria*.

Symbiont	Protein Sequence	Insect Host Name	Nutrition	Presence of Amino Acid Pathway – Source (153, 169) A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
Wigglesworthia glossinidia (5)	PSORTDB	*Glossina brevipalpis*	blood				P		Y	Y	Y				P					P			
Sodalis glossinidius str. 'morsitans' (153)	NCBI	*Glossina morsitans morsitans*	blood		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Buchnera aphidicola str. APS (137)	PSORTDB	*Acyrthosiphon pisum*	phloem		Y			Y			Y	Y	Y	Y	Y		Y			Y	Y		Y
Buchnera aphidicola str. Sg (150)	NCBI	*Schizaphis graminum*	phloem		Y			Y			Y	Y	Y	Y	Y		Y			Y	Y		Y

P=Some genes in pathway are present

	Mean Protein Values Proteins	%GC	Length	Mass/aa	ATPs/aa	%A	%R	%N	%D	%C	%E	%Q	%G	%H	%I	%L	%K	%M	%F	%P	%S	%T	%W	%Y	%V
W. glossinidia	611	22.5	330.1	114.979	25.585	3.6	3.3	8.4	3.8	1.3	5	2.3	5	1.6	13.3	9.5	12.1	2.1	5.5	2.7	7.6	3.7	0.8	3.9	4.2
S. glossinidius	2515	54.51	287.1	110.743	23.261	10.1	6.6	3.4	5.3	1.3	5.4	4.7	7.2	2.4	5.6	10.9	4.2	2.8	3.6	4.5	5.8	5.3	1.3	2.8	6.9
B. aphidicola APS	574	26.35	328	114.303	25.027	4.5	4	7.1	4.1	1.2	5.4	3.2	5.4	2.1	11.7	9.8	10.3	2.3	5	2.9	7.2	4.6	0.9	3.5	4.9
B. aphidicola SG	545	25	326.2	114.531	25.192	4.4	3.9	7.2	4	1.2	5.4	3	5.3	2	11.8	9.8	11	2.2	5.3	2.9	7	4.4	0.8	3.5	4.8
E. coli K-12	4333	50.8	312.5	111.005	23.555	9.4	5.6	3.9	5	1.3	5.8	4.4	7.1	2.3	6.1	10.6	4.7	3	3.9	4.3	5.8	5.4	1.5	2.8	7.1

Notes: Amino acid percents are on a per protein basis. The percent amino acid composition of each protein were averaged; they are unweighted for length.

Table 3.24: Amino acid composition of extracellular flagellar subunits in endosymbionts & other *Gammaproteobacteria* (continued).

Amino Acids	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	
%aa in highly expressed genes (128)	↑	↑	↓		↓	↑		↑		↓	↓	↑		↓		↓†	↑†	↓	↓	↑	
Source of insect host nutrition																					**BOLD**
Vicia faba phloem (molar %)(165)	**3**	0.6	**68.7**	**3.4**		**3.7**	**6.7**	0.8	**3.3**	0.5	0.6	1	0.2	0.4		**4.4**	1	0.2	0.5	1	>3%
Triticum aestivum (>5mM)(127)	Y			Y		Y	Y		Y*	Y	Y	Y		Y		Y	Y			Y	
Triticum aestivum (>10mM)(127)				Y		Y	Y			Y	Y			Y		Y				Y	
Human whole blood (µM)(117)	**325**	nd	48	129	nd	**208**	**528**	**234**	58	51	114	79	19	47	184	118	124	nd	63	**213**	>200µM
Human plasma (µM)(117)	**208**	nd	15	7	nd	35	**344**	**105**	40	37	80	69	15	34	**130**	64	82	nd	41	**146**	>100µM
Human serum (µM)(118)	**351**	91	59	11	10	99	**748**	**325**	88	89	**135**	**187**	8	63	**254**	**155**	**186**	73	70	**274**	>100µM
Mature human serum albumin (%aa)	**10.6**	4.1	2.9	6.2	6.0	**10.6**	3.4	2.1	2.7	1.4	**10.4**	**10.1**	1.0	5.3	4.1	4.1	4.8	0.2	3.1	7.0	>10%
%aa in ext. flagellar prot. **																					
Buchnera aphidicola str. APS	↓		↑		↑			↓	↑	↑		↑					↓			↓	
Buchnera aphidicola str. Sg	↓		↑		↑			↓	↑	↑		↑			↓		↓			↓	
Wigglesworthia glossinidia	↓		↑	↓		↑	↓	↓	↑	↑		↑	↓	↑	↓		↓		↑	↓	

Notes: Selected papers related to amino acid composition in phloem sap – (4, 40, 126, 127, 165), in human blood – (103, 104, 117, 118).

Legend: nd – not done

† only in *W. glossinidia*;

* Histidine in stylus from *B. aphidicola* str. Sg feeding on *T. aestivium*is is ~4.9mM (Adobe Photoshop used to measure graph in (127))

** Trend in amino acid composition in at least 3 of 4 (*Buchnera)* or 4 of 7 (*Wigglesworthia*) in flagellar proteins compared to other *Gammaproteobacteria* (see continued table below).

Table 3.24: Amino acid composition of extracellular flagellar subunits in endosymbionts & other *Gammaproteobacteria* (continued).

		Homologs		Mean Values		Economic Rank **																			
		Total	Size*	Mass/aa	ATPs/aa	Organism	Homologs	%A	%R	%N	%D	%C	%E	%Q	%G	%H	%I	%L	%K	%M	%F	%P	%S	%T	%W	%Y	%V
FliC	γ Proteobacteria	544	516	104.682	19.662			13.1	3.8	8.2	6.5	0.0	4.5	6.5	8.0	0.4	5.8	8.5	4.1	2.5	2.0	0.7	10.0	8.7	0.1	1.1	5.5
	W. glossinidia			109.134	21.100	4/611	514/516	8.8	4.0	8.8	6.7	0.0	6.2	5.5	5.2	0.5	9.5	8.3	7.1	1.4	2.1	1.0	11.9	4.8	0.0	**2.6**	5.7
	S. glossinidius			106.962	20.633	48/2515	489/516	10.7	3.3	6.9	6.7	0.0	6.9	5.7	8.1	**0.0**	9.1	7.2	7.4	**1.0**	1.9	1.4	7.9	8.1	0.0	2.6	5.0
FliD	γ Proteobacteria	284	277	104.682	19.662			10.2	2.9	5.8	7.6	0.0	4.9	5.3	7.5	0.4	5.4	8.9	6.4	2.1	2.5	1.8	10.3	10.2	0.2	1.9	5.8
	W. glossinidia			116.578	24.725	185/611	277/277	**1.6**	2.6	**13.5**	4.2	**0.4**	7.9	**2.0**	4.0	**1.0**	**12.1**	9.7	**15.8**	1.6	**5.3**	1.6	8.7	**2.2**	**0.4**	3.2	**2.4**
	S. glossinidius			108.020	20.970	92/2515	224/277	9.1	3.1	5.6	6.0	0.0	7.6	5.1	7.2	**1.0**	6.2	9.9	8.6	2.5	1.4	2.3	9.7	8.6	**0.4**	1.9	3.9
FliK	γ Proteobacteria	208	160	104.442	19.734			14.1	2.7	3.8	6.2	0.0	5.3	8.2	6.4	2.0	3.2	10.6	4.1	1.9	1.8	6.4	9.6	7.7	0.4	0.4	4.9
	W. glossinidia			120.408	25.627	328/611	155/160	**0.5**	**0.2**	**17.0**	9.0	**0.5**	7.8	**1.9**	**1.0**	**0.5**	**11.2**	8.8	**16.5**	**0.5**	**8.0**	**0.2**	6.6	**1.7**	**0.0**	**5.8**	**2.2**
	S. glossinidius			105.888	20.726	57/2515	151/160	14.2	5.4	2.4	5.6	**0.5**	3.8	8.6	7.5	**4.3**	3.5	11.8	**1.9**	1.1	1.3	7.3	5.9	7.8	0.5	**1.3**	5.1
	B. aphidicola APS			118.654	26.538	479/574	160/160	**0.9**	**0.4**	**11.9**	4.3	**0.4**	4.7	**1.3**	**0.9**	3.0	**14.0**	7.7	**20.4**	2.1	**4.7**	**2.1**	9.8	**3.4**	**0.9**	**4.7**	2.6
	B. aphidicola SG			119.464	26.081	408/545	158/160	**1.3**	2.4	**15.5**	4.2	**1.0**	4.7	**3.1**	**1.3**	2.4	**10.8**	7.1	**17.6**	1.8	**8.4**	**1.3**	7.3	3.9	0.3	**3.1**	**2.4**
FlgD	γ Proteobacteria	232	232	104.719	20.362			8.2	1.7	6.6	5.7	0.0	4.6	7.0	9.6	0.6	4.7	9.3	4.3	2.8	2.5	3.5	10.1	8.9	0.4	1.4	8.0
	W. glossinidia			115.399	23.758	81/611	232/232	**1.4**	1.8	**16.1**	4.7	0.0	7.2	4.3	**3.9**	**1.4**	**11.8**	7.5	**11.5**	2.2	**5.7**	2.5	6.8	6.5	0.4	1.8	**2.5**
	S. glossinidius			107.147	21.125	15/2515	213/232	11.5	2.6	7.8	6.3	**0.4**	4.8	5.6	7.1	0.7	3.0	9.3	6.3	4.1	2.6	4.1	7.8	5.6	0.4	2.6	7.4
	B. aphidicola APS			110.889	22.048	8/574	230/232	**3.4**	1.3	**14.0**	5.9	0.0	3.4	7.2	**4.2**	**1.3**	**12.3**	8.5	8.5	2.1	2.5	4.7	8.9	5.9	0.4	0.8	4.7
	B. aphidicola SG			109.811	22.002	8/545	228/232	**3.5**	1.3	12.4	4.4	0.0	4.0	6.2	4.9	**2.2**	**14.2**	8.0	8.4	2.2	**0.9**	2.2	9.7	8.0	0.4	0.9	6.2
FlgE	γ Proteobacteria	260	253	105.108	20.884			9.5	2.2	9.1	5.9	0.0	3.1	5.8	10.1	0.6	4.3	7.0	3.1	1.9	4.3	3.5	9.5	10.0	0.6	2.8	6.6
	W. glossinidia			108.765	22.121	13/611	250/253	**3.9**	1.2	11.2	4.9	**0.2**	4.4	6.1	9.0	**1.2**	**9.5**	6.3	**6.3**	2.7	4.4	2.7	12.9	**4.6**	0.5	2.9	4.9
	S. glossinidius			107.360	22.062	446/2515	249/253	7.9	1.5	7.7	5.4	**0.5**	4.5	6.2	11.1	**0.2**	5.2	6.7	**6.4**	3.2	4.5	4.2	7.7	7.4	0.7	4.0	5.0
	B. aphidicola APS			112.277	22.797	24/574	252/253	**2.9**	1.7	12.5	6.4	**0.5**	6.1	4.4	6.1	0.7	**9.1**	7.6	**8.1**	2.5	4.4	**1.7**	10.5	7.1	0.7	3.4	3.4
	B. aphidicola SG			112.295	23.113	43/545	253/253	**2.9**	1.5	12.2	6.6	**0.2**	5.6	5.1	6.1	**1.7**	**12.0**	9.3	**7.6**	1.2	5.1	2.0	10.3	**4.4**	0.5	2.7	**2.9**
FlgK	γ Proteobacteria	292	275	106.450	20.722			10.3	3.4	8.0	6.5	0.0	4.2	6.5	8.3	0.7	4.9	9.1	3.8	2.3	3.1	2.5	9.3	7.6	0.2	2.5	6.6
	W. glossinidia			115.250	24.132	118/611	275/275	**3.9**	1.8	12.8	5.3	0.0	**9.0**	**3.1**	4.4	**1.7**	**11.9**	8.3	**10.3**	**0.6**	**6.6**	**0.9**	9.7	**2.6**	0.2	4.4	**2.6**
	S. glossinidius			108.524	20.891	76/2515	202/275	8.5	5.0	9.5	6.8	0.0	5.6	6.1	9.0	1.1	4.7	10.4	4.1	**0.9**	2.5	2.0	7.9	6.7	0.2	3.4	5.6
	B. aphidicola APS			115.122	23.590	88/574	272/275	**4.2**	2.9	11.4	7.2	**0.9**	7.9	3.5	**2.8**	**1.7**	**11.0**	8.8	**9.8**	1.8	3.7	2.0	6.6	4.2	0.2	3.7	5.5
	B. aphidicola SG			116.360	24.132	140/545	274/275	**2.8**	2.6	11.0	8.3	**0.9**	7.7	3.5	**2.8**	**1.7**	**11.6**	8.6	**10.1**	1.7	4.2	1.5	6.8	5.0	0.4	4.4	4.6
FlgL	γ Proteobacteria	264	238	108.528	21.057			9.6	3.7	6.9	7.3	0.0	5.2	7.4	6.6	0.8	4.9	8.9	4.3	2.9	3.3	2.3	10.0	7.0	0.3	2.6	5.9
	W. glossinidia			115.438	23.440	58/611	238/238	**1.0**	3.2	12.8	**1.9**	**1.0**	9.3	4.8	**2.9**	**1.6**	**10.2**	12.1	**8.9**	2.2	3.5	1.3	13.1	**3.5**	0.3	3.2	3.2
	S. glossinidius			107.083	21.309	146/2515	176/238	11.2	3.8	4.8	8.0	0.0	3.5	7.3	9.6	1.3	5.1	9.3	4.8	4.2	2.6	1.9	7.3	5.4	0.3	2.9	6.7

BOLD = Amino acid % is 2X or 1/2 that of same protein average in γ Proteobacteria

Notes: * Homologs whose lengths were 50% (+/-) their respective *E. coli* counterpart were removed. Comparisons were done on the remaining homologs.

** Economic Rank in ATPs/aa (least expensive/total) amongst homologs and within each organism.

Amino acid % in *γproteobacteria* are on a per protein basis. The amino acid compositions of each protein are averaged and unweighted for length.

Buchnera aphidicola strains lack FliA, FliC, FliD, FliL, FliO, FliS, FliT, FlgM and FlgL (88).

Table 3.25: Average synthetic cost, GC-content, descriptions, and auxotrophy of microbes used.

Gram Negative (PSORTDB)	Group	Proteins	Amino Acid Cost Mean (ASC)	STDEV	Mean Amino Acid Mass	Craig & Weber	Mean Chain Length	% GC** content	Genome Size (MB)**	Plasmids	% Coding	CDS** Coverage
Solibacter usitatus Ellin6076	Acidobacteria	7826	23.016	1.232	109.493	20.055	383.0	61.9	9.966	0	90	90
Aquifex aeolicus VF5†	Aquificae	1529	24.763	1.306	114.079	21.361	316.0	43.5	1.551	1	92	94
Aquifex aeolicus VF5 plasmid ece1	-	32	25.284	1.344	116.665	21.698	226.1	36.4	0.039	-	53	54
Cytophaga hutchinsonii ATCC 33406	Bacteroidetes/Chlorobi	3785	24.530	1.555	112.817	20.964	350.8	38.8	4.433	0	89	90
Flavobacterium johnsoniae UW101	Bacteroidetes/Chlorobi	5017	24.443	1.431	113.337	20.685	352.7	34.11	6.097	0	87	87
Flavobacterium psychrophilum JIP02/86	Bacteroidetes/Chlorobi	2412	24.568	1.527	113.394	20.666	333.6	32.5	2.862	0	84	85
Salinibacter ruber DSM 13855†	Bacteroidetes/Chlorobi	2801	22.310	1.293	109.508	19.229	357.8	66.2	3.552	1	84	85
Sulcia muelleri GWSS*	Bacteroidetes/Chlorobi	227	26.031	1.822	115.710	21.522	331.1	22.4	0.246	0	91	92
Chlamydia trachomatis D/UW-3/CX	Chlamydiae	895	23.782	1.365	111.637	20.790	348.9	41.3	1.043	0	89	90
Chlamydophila pneumoniae AR39	Chlamydiae	1112	24.133	1.591	112.374	21.033	327.4	40.6	1.230	0	80	89
Roseiflexus sp. RS-1	Chloroflexi	4517	23.080	1.284	110.071	20.033	363.4	60.4	5.802	0	84	85
Roseiflexus castenholzii DSM 13941	Chloroflexi	4330	23.072	1.285	109.936	20.060	361.4	60.7	5.723	0	82	82
Prochlorococcus marinus AS9601	Cyanobacteria	1921	24.483	1.711	113.434	20.730	262.4	31.3	1.670	0	90	91
Synechocystis sp. PCC 6803†	Cyanobacteria	3167	23.552	1.444	111.340	20.405	326.3	47.7	3.573	4	86	87
Synechocystis sp. PCC 6803 plasmids‡	-	397	23.832	1.386	112.835	20.450	260.6	44.0	0.374	-	82.7	83.3
Thermosynechococcus elongatus BP-1	Cyanobacteria	2475	23.497	1.479	111.154	20.557	312.8	53.9	2.594	0	89	90
Trichodesmium erythraeum IMS101	Cyanobacteria	4451	24.086	1.543	112.993	20.519	346.8	34.14	7.750	0	59	60
Thermus thermophilus HB27†	Deinococcus-Thermus	1982	23.235	1.263	110.119	21.162	303.6	69.4	1.895	1	94	96
Mycoplasma genitalium G37	Firmicutes	484	24.767	1.347	114.243	21.019	363.5	31.7	0.580	0	90	91
Mycoplasma pneumoniae M129	Firmicutes	690	24.348	1.482	113.420	20.788	348.0	40.0	0.816	0	87	88
Fusobacterium nucleatum ATCC 25586	Fusobacteria	2067	24.807	1.826	114.196	20.757	311.5	27.2	2.175	0	88	89
Rhodopirellula baltica SH 1	Planctomycetes	7325	23.011	1.366	110.750	19.720	314.9	55.4	7.146	0	94	97
Caulobacter cresentus CB15	Alphaproteobacteria	3737	22.423	1.311	107.884	19.544	323.6	67.2	4.017	0	90	91
Ehrlichia chaffeensis str. Arkansas	Alphaproteobacteria	1105	24.899	2.069	113.623	21.198	281.5	30.1	1.176	0	79	80
Hodgkinia cicadicola Dsem*	Alphaproteobacteria	169	22.759	1.387	107.651	20.859	257.9	58.4*	0.144	0	90	90*
Pelagibacter ubique HTCC1062	Alphaproteobacteria	1354	24.632	1.619	113.151	20.578	307.6	29.7	1.309	0	95	96
Rhodobacter sphaeroides 2.4.1†‡	Alphaproteobacteria	3857	22.479	1.375	108.138	19.660	314.9	69.0	4.132	5	87.5	88.3
Rhodobacter sphaeroides 2.4.1 plasmids‡	-	385	22.498	1.307	108.263	19.766	313.2	66.8	0.471	-	76.2	76.6
Rickettsia prowazekii Madrid E	Alphaproteobacteria	835	24.691	1.484	113.404	20.684	334.4	29.0	1.112	0	75	75
Silicibacter sp. TM1040†	Alphaproteobacteria	3030	22.720	1.370	109.057	19.642	312.5	60.4	3.201	2	88	89
Silicibacter sp. TM1040 plasmids‡	-	835	22.875	1.315	109.040	19.902	326.9	58.9	0.953	-	85.8	85.9
Wolbachia pipientis wMel	Alphaproteobacteria	1195	24.270	1.508	113.080	20.504	282.7	35.2	1.268	0	80	80
Bordetella pertussis Tohama I	Betaproteobacteria	3436	22.684	1.296	108.604	19.907	326.6	67.7	4.086	0	82	92
Burkholderia xenovorans strain LB400‡	Betaproteobacteria	8702	22.762	1.347	108.813	19.900	320.1	62.6	9.731	0	85.3	86.2
Neisseria gonorrhoeae FA 1090	Betaproteobacteria	2002	23.472	1.530	110.805	20.203	280.8	52.7	2.154	0	78	79
Ralstonia eutropha JMP134†‡	Betaproteobacteria	5846	22.641	1.305	108.409	19.767	329.2	64.8	6.533	2	88	88.6
Ralstonia eutropha JMP134 plasmids‡	-	600	22.725	1.191	108.703	19.786	315.6	64.1	0.723	-	77.8	78.8
Anaeromyxobacter dehalogenans 2CP-C	Deltaproteobacteria	4346	21.913	1.324	106.920	19.421	349.1	74.9	5.013	0	90	91
Geobacter metallireducens GS-15†	Deltaproteobacteria	3519	23.231	1.289	110.169	20.028	343.2	59.5	3.997	1	90	91
Myxococcus xanthus DK 1622	Deltaproteobacteria	7331	22.396	1.253	108.633	22.396	378.2	68.9	9.140	0	90	92
Desulfotalea psychrophila LSv54†	Deltaproteobacteria	3118	23.737	1.438	111.143	20.499	321.5	46.8	3.523	2	85	86

Table 3.25: Average synthetic cost, GC-content, descriptions, and auxotrophy of microbes used (continued).

Gram Negative (PSORTDB)	Group	Proteins	Amino Acid Cost Mean (ASC)	STDEV	Mean Amino Acid Mass	Craig & Weber	Mean Chain Length	% GC** content	Genome Size (MB)**	Plasmids	% Coding	CDS** Coverage
Arcobacter butzleri RM4018	Epsilonproteobacteria	2259	24.608	1.597	113.851	20.601	320.9	27.0	2.341	0	92	93
Campylobacter jejuni RM1221	Epsilonproteobacteria	1838	24.738	1.820	113.963	21.105	293.1	30.3	1.778	0	90	91
Helicobacter pylori J99	Epsilonproteobacteria	1491	24.380	1.408	113.285	20.977	331.4	39.2	1.644	0	90	90
Wolinella succinogenes DSM1740	Epsilonproteobacteria	2044	24.042	1.384	112.206	20.968	322.6	48.5	2.110	0	93	94
Acinetobacter sp. ADP1	Gammaproteobacteria	3325	23.947	1.592	112.141	20.561	316.2	40.4	3.599	0	87	88
Baumannia cicadellinicola Hc	Gammaproteobacteria	595	24.344	1.312	112.730	20.530	327.7	33.2	0.686	0	85	86
Blochmannia floridanus	Gammaproteobacteria	583	25.084	1.480	114.095	21.083	334.6	27.4	0.706	0	83	84
Buchnera aphidicola APS Tokyo†	Gammaproteobacteria	564	25.033	1.485	114.324	20.793	328.3	26.3	0.641	2	86	87
Buchnera aphidicola APS Tokyo plasmids‡	-	10	24.719	1.091	113.104	20.789	311.3	28.6	0.015	-	58.2	62.3
Carsonella ruddii PV	Gammaproteobacteria	182	28.037	2.246	118.846	23.710	274.3	16.6	0.160	0	93	94
Colwellia psychrerythraea 34H	Gammaproteobacteria	4910	23.864	1.625	111.778	20.442	308.0	38.0	5.373	0	84	85
Coxiella burnetii RSA 493†	Gammaproteobacteria	2009	24.286	1.654	112.896	20.857	279.3	42.7	1.995	1	77	85
Escherichia coli K-12 MG1655	Gammaproteobacteria	4333	23.555	1.504	111.005	20.314	312.5	50.8	4.640	0	85	88
Klebsiella pneumoniae MGH 78578†	Gammaproteobacteria	4776	23.360	1.410	110.348	20.338	318.3	57.5	5.315	5	85	86
Legionella pneumophila Paris†	Gammaproteobacteria	3027	24.213	1.451	112.633	20.756	335.4	38.4	3.504	1	87	88
Marinobacter aquaeolei VT8†	Gammaproteobacteria	3858	23.075	1.310	110.691	19.939	336.3	57.3	4.780	2	90	90
Marinobacter aquaeolei VT8 plasmids‡	-	414	23.096	1.242	111.568	19.638	305.1	53.7	0.453	-	83.2	84.2
Photorhabdus luminescens laumondii TTO1	Gammaproteobacteria	4683	23.863	1.462	112.161	20.397	327.4	42.8	5.689	0	80	84
Pseudomonas aeruginosa PAO1	Gammaproteobacteria	5567	22.280	1.303	109.787	20.314	334.1	66.6	6.264	0	89	89
Pseudomonas syringae pv. tomato DC3000†	Gammaproteobacteria	5471	23.050	1.362	110.067	20.080	331.1	58.4	6.397	2	85	85
Pseudomonas syringae pv. tomato DC3000 plasmids‡	-	137	22.901	1.473	110.400	19.734	273.3	55.6	0.141	-	80.6	81.6
Psychromonas ingrahamii 37	Gammaproteobacteria	3545	23.783	1.392	111.555	20.382	332.0	40.1	4.560	0	77	78
Psychrobacter arcticum 273-4	Gammaproteobacteria	2120	23.277	1.446	110.842	19.676	335.1	42.8	2.651	0	80	81
Psychrobacter cryohalolentis K5†	Gammaproteobacteria	2467	23.303	1.449	110.759	19.745	342.9	42.3	3.060	1	83	83
Salmonella typhimurium LT2†	Gammaproteobacteria	4425	23.533	1.502	110.880	20.325	315.3	52.2	4.857	1	86	87
Salmonella typhimurium LT2 plasmid pSLT	-	102	23.496	1.449	112.144	20.146	243.7	53.1	0.094	-	78	79
Shewanella denitrificans OS217	Gammaproteobacteria	3754	23.406	1.361	110.823	20.210	340.1	45.1	4.546	0	84	85
Shewanella oneidensis MR-1†	Gammaproteobacteria	4323	23.526	1.391	111.100	20.328	318.6	46	4.970	1	83	83
Shewanella oneidensis MR-1 plasmid pMR-1	-	148	23.950	1.558	112.853	20.393	259.1	43.7	0.162	-	69	73
Vibrio cholerae O1 biovar eltor N16961‡	Gammaproteobacteria	3835	23.759	1.532	111.782	20.572	304.4	47.5	4.033	0	86.2	86.5
Wigglesworthia glossinidia brevipalpis†	Gammaproteobacteria	611	25.585	1.654	114.979	21.203	330.1	22.5	0.698	1	86	87
Yersinia pestis CO92†	Gammaproteobacteria	3885	23.605	1.523	111.211	20.382	319.4	47.6	4.654	3	80	84
Yersinia pestis CO92 plasmids‡	-	182	23.466	1.419	112.518	19.834	245.3	47.8	0.176	-	75	83
Magnetococcus MC-1	Proteobacteria:Other	3716	23.277	1.354	110.848	20.093	365.2	54.2	4.720	0	86	87
Borrelia burgdorferi B31†	Spirochaetes	851	25.282	1.765	114.786	21.518	333.3	28.6	0.911	21	93	94
Treponema pallidum Nichols	Spirochaetes	1036	23.484	1.277	110.842	20.774	339.3	52.8	1.138	0	92	93
Fervidobacterium nodosum Rt17-B1	Thermotogae	1750	24.703	1.362	113.590	20.985	334.5	35.0	1.949	0	90	90
Thermosipho melanesiensis BI429	Thermotogae	1879	25.129	1.545	114.336	21.347	309.9	31.4	1.915	0	91	92
*Thermotoga petrophila*RKU-1	Thermotogae	1785	24.521	1.318	113.555	21.131	321.6	46.1	1.824	0	94	95
Thermotoga maritima MSB8	Thermotogae	1858	24.522	1.365	113.543	21.151	314.9	46.2	1.861	0	93	94

Table 3.25: Average synthetic cost, GC-content, descriptions, and auxotrophy of microbes used (continued).

Gram Positive (PSORTDB)	Group	Proteins	Amino Acid Cost Mean (ASC)	STDEV	Mean Amino Acid Mass	Craig & Weber	Mean Chain Length	% GC** content	Genome Size (MB)**	Plasmids	% Coding	CDS** Coverage
Arthrobacter aurescens TC1†	Actinobacteria	4041	22.312	1.323	107.435	19.480	333.6	62.3	4.598	2	88	88
Corynebacterium diphtheriae NCTC 13129	Actinobacteria	2272	22.766	1.428	109.039	19.550	320.2	53.5	2.489	0	87	90
Frankia sp. EAN1pec	Actinobacteria	7191	21.871	1.266	106.890	19.081	347.2	71.2	8.982	0	83	84
Kineococcus radiotolerans SRS30216†	Actinobacteria	4480	21.587	1.285	106.114	19.103	320.5	74.4	4.761	2	90	91
Kineococcus radiotolerans SRS30216 plasmids‡	-	201	21.517	1.311	106.690	18.606	268.9	69.6	0.195	-	82.5	83.4
Mycobacterium leprae TN	Actinobacteria	1605	22.472	1.276	108.125	19.503	335.684	57.8	3.268	0	49	77
Mycobacterium tuberculosis H37Rv	Actinobacteria	3927	22.147	1.454	107.144	19.309	339.2	65.6	4.412	0	90	91
Nocardia farcinica IFM 10152†	Actinobacteria	5683	22.098	1.276	107.375	19.249	318.4	70.8	6.021	2	90	90
Nocardioides sp. JS614†	Actinobacteria	4645	22.017	1.229	107.248	19.298	325.6	71.7	4.986	2	91	91
Rubrobacter xylanophilus DSM 9941	Actinobacteria	3140	22.418	1.275	108.883	19.987	313.8	70.5	3.226	0	91	92
Saccharopolyspora erythraea NRRL 2338	Actinobacteria	7197	22.070	1.330	107.635	19.384	322.1	71.1	8.213	0	84	85
Salinispora tropica CNB-440	Actinobacteria	4536	22.045	1.260	107.328	19.311	335.2	69.5	5.183	0	88	88
Streptomyces coelicolor A3(2)†	Actinobacteria	7769	21.925	1.272	107.096	19.187	329.6	72.1	8.668	2	88	89
Thermobifida fusca YX	Actinobacteria	3110	22.291	1.242	108.327	19.432	331.4	67.5	3.642	0	85	85
Tropheryma whipplei Twist	Actinobacteria	808	23.488	1.457	110.222	20.536	328.6	46.3	0.927	0	85	86
Dehalococcoides ethenogenes 195	Chloroflexi	1580	23.676	1.549	110.863	20.329	277.7	48.9	1.470	0	89	90
Deinococcus radiodurans R1†‡	Deinococcus-Thermus	2997	22.360	1.303	108.243	19.709	307.6	67.0	3.061	2	89.4	88.1
Bacillus anthracis Ames	Firmicutes	5311	24.518	1.745	113.244	20.842	263.9	35.4	5.227	0	80	81
Bacillus cereus ATCC 14579†	Firmicutes	5234	24.461	1.682	113.206	20.769	277.3	35.3	5.412	1	80	81
Bacillus subtilis 168	Firmicutes	4112	24.052	1.533	112.397	20.453	296.4	43.5	4.216	0	87	88
Clostridium botulinum A ATCC 19397	Firmicutes	3553	24.558	1.584	113.599	20.294	293.4	28.2	3.863	0	81	81
Clostridium difficile 630†	Firmicutes	3742	24.306	1.469	113.174	20.166	313.0	29.1	4.290	1	82	83
Lactobacillus gasseri ATCC 33323	Firmicutes	1755	24.193	1.584	112.942	20.315	317.4	35.3	1.894	0	88	88
Lactococcus lactis subsp. lactis Il1403	Firmicutes	2321	24.138	1.643	112.694	20.490	288.6	35.3	2.366	0	85	88
Listeria monocytogenes EGD-e	Firmicutes	2846	23.931	1.532	112.117	20.183	306.0	38.0	2.945	0	88	89
Oceanobacillus iheyensis strain HTE831	Firmicutes	3500	24.135	1.689	112.892	20.338	290.9	35.7	3.631	0	84	85
Staphylococcus aureus Newman	Firmicutes	2614	24.320	1.674	113.427	20.321	305.2	32.9	2.879	0	83	83
Staphylococcus aureus RF122	Firmicutes	2515	24.381	1.751	113.398	20.441	292.6	32.8	2.743	0	80	84
Streptococcus mutans UA159	Firmicutes	1960	24.238	1.673	112.788	20.682	295.8	36.8	2.031	0	85	86
Streptococcus pneumoniae TIGR4	Firmicutes	2094	24.240	1.836	112.990	20.792	283.5	39.7	2.161	0	83	84
Caldicellulosiruptor saccharolyticus DSM 8903	Other bacteria	2679	24.743	1.373	114.034	20.923	318.2	35.3	2.970	0	86	86
Yeast												
Saccharomyces cerevisiae S288C	Ascomycetes	6732	26.594	2.145	113.198	N/A		38.3	12.16	0		

† Chromosomal only

‡ Weighted averages were used for multiple plasmids or chromosomes

Bacterial Mean Amino Acid Cost calculated from Ahashi and Gojobori; sequences from PSORTDB except *Escherichia coli* from Echobase and *from NCBI

Yeast Mean Amino Acid Cost calculated from Wagner's respiratory cost; sequences from the MIPS Comprehensive Yeast Genome Database (CYGD)

Table 3.25: Average synthetic cost, GC-content, descriptions, and auxotrophy of microbes used (continued).

Gram Negative (PSORTDB)	Shape	Arrangement	Endospore	Motile	Salinity	Oxygen Requirement	Habitat	Temperature (optimal)
Solibacter usitatus Ellin6076	Rod	Clusters	-	No	Non-halophilic	Aerobic	Terrestrial	Mesophilic(30C)
Aquifex aeolicus VF5†	Coccus	-	-	Yes[L]	-	Aerobic	Specialized	Hyperthermophile (96C)
Aquifex aeolicus VF5 plasmid ece1	-	-	-	-	-	-	-	-
Cytophaga hutchinsonii ATCC 33406	Rod	Singles	-	Yes	-	Aerobic	Multiple	Mesophilic(30C)
Flavobacterium johnsoniae UW101	Rod	-	No	Yes	Non-halophilic	Aerobic	Multiple	Mesophilic(20-30C)
Flavobacterium psychrophilum JIP02/86	Rod	-	No	Yes	-	Aerobic	Aquatic	Psychophilic
Salinibacter ruber DSM 13855†	Curved Rod	-	-	Yes	Extreme halophilic	Aerobic	Specialized	Mesophilic (37-47C)
Sulcia muelleri GWSS*	tubular[L]	-	-	-	-	-	Host associated[L]	Mesophilic [L]
Chlamydia trachomatis D/UW-3/CX	Rod	-	-	-	-	-	Host-associated	Mesophilic (37C)
Chlamydophila pneumoniae AR39	Rod	-	-	-	-	-	Host-associated	Mesophilic (37C)
Roseiflexus sp. RS-1	Rod	Filaments	No	Yes	-	Facultative	Specialized	Thermophilic
Roseiflexus castenholzii DSM 13941	-	Filament	-	Yes	-	Facultative,O_2 resp[L]	Aquatic	Thermophilic (50C)
Prochlorococcus marinus AS9601	-	-	-	-	-	-	Aquatic	Mesophilic
Synechocystis sp. PCC 6803†	Coccus	Aggregates	-	-	-	-	Aquatic	Mesophilic
Synechocystis sp. PCC 6803 plasmids‡	-	-	-	-	-	-	-	-
Thermosynechococcus elongatus BP-1	Rod	Singles	-	-	-	-	Specialized	Thermophilic (55C)
Trichodesmium erythraeum IMS101	Filament	Filaments,Aggregates	-	Yes	-	Aerobic	Aquatic	Mesophilic [L]
Thermus thermophilus HB27†	-	-	-	-	halotolerant**	Aerobic	Specialized	Themophilic (68C)
Mycoplasma genitalium G37	Sphere	Singles	No	Yes	Non-halophilic	Facultative	Host-associated	Mesophilic (37C)
Mycoplasma pneumoniae M129	Sphere	Singles	No	Yes	Non-halophilic	Facultative	Host-associated	Mesophilic (37C)
Fusobacterium nucleatum ATCC 25586	Rod	Singles	No	No	-	Anaerobic	Host-associated	Mesophilic (37C)
Rhodopirellula baltica SH 1	Sphere	Singles,Rosettes	No	Yes	-	Aerobic	Aquatic	Mesophilic (28C)
Caulobacter cresentus CB15	Rod	Singles	-	Yes	-	Aerobic	Aquatic	Mesophillic (35C)
Ehrlichia chaffeensis str. Arkansas	Rod	Singles	-	-	-	-	Host-associated	Mesophilic **
Hodgkinia cicadicola Dsem*	tubular[L]	-	-	-	-	-	Host associated[L]	Mesophilic [L]
Pelagibacter ubique HTCC1062	Rod	Singles	-	No	-	Aerobic	Aquatic	Mesophilic
Rhodobacter sphaeroides 2.4.1†‡	Rod	Chains	-	Yes	-	Facultative	Multiple	Mesophilic(25-35C)
Rhodobacter sphaeroides 2.4.1 plasmids‡	-	-	-	-	-	-	-	-
Rickettsia prowazekii Madrid E	Rod	-	-	-	-	Aerobic	Host-associated	Mesophilic
Silicibacter sp. TM1040†	Rod	-	-	Yes	-	Aerobic[L]	Multiple	Mesophilic [L]
Silicibacter sp. TM1040 plasmids‡	-	-	-	-	-	-	-	-
Wolbachia pipientis wMel	Rod, Coccus	-	-	-	-	-	Host-associated	Mesophilic
Bordetella pertussis Tohama I	Rod	Singles,Pairs,Chains	-	-	-	Aerobic	Host-associated	Mesophilic (35-37C)
Burkholderia xenovorans strain LB400‡	Rod	Singles	No	Yes	-	Aerobic	Multiple	Mesophilic (35-37C)
Neisseria gonorrhoeae FA 1090	Coccus	Singles,Pairs	-	-	-	Aerobic	Host-associated	Mesophilic (35-37C)
Ralstonia eutropha JMP134†‡	Rod, Coccus	-	-	Yes	Non-halophilic	Facultative	Multiple	Mesophilic(30C)
Ralstonia eutropha JMP134 plasmids‡	-	-	-	-	-	-	-	-
Anaeromyxobacter dehalogenans 2CP-C	Rod	-	Yes	Yes	-	Facultative	Terrestrial	Mesophilic(30C)
Geobacter metallireducens GS-15†	Rod	-	-	Yes	-	Aerobic	Aquatic	Mesophilic(30C)
Myxococcus xanthus DK 1622	Rod	Singles,Aggregates	Yes	Yes	-	Aerobic	Terrestrial	Mesophilic(20-30C)
Desulfotalea psychrophila LSv54†	Rod	-	-	-	-	Anaerobic	Speciaized	Psychrophilic (7C)

Table 3.25: Average synthetic cost, GC-content, descriptions, and auxotrophy of microbes used (continued).

Gram Negative (PSORTDB)	Shape	Arrangement	Endospore	Motile	Salinity	Oxygen Requirement	Habitat	Temperature (optimal)
Arcobacter butzleri RM4018	Spiral	Singles, Chains	No	Yes	-	Aerobic	Multiple	Mesophilic
Campylobacter jejuni RM1221	Spiral	Singles, Chains	No	Yes	-	Microaerophilic	Multiple	Mesophilc
Helicobacter pylori J99	Spiral	Singles	-	Yes	-	Aerobic	Host-associated	Mesophilic (37C)
Wolinella succinogenes DSM1740	Spiral, Curved	-	-	Yes	-	Microaerophilic	Host-associated	Mesophilic
Acinetobacter sp. ADP1	Coccobacillus	Pairs	No	No	Non-halophilic	Aerobic	Multiple	Mesophilic (37C)
Baumannia cicadellinicola Hc	Irregular Sphere	-	-	-	-	-	Host associated	Mesophilic [L]
Blochmannia floridanus	Rod	-	-	-	-	-	Specialized, Host-associated[L]	Mesophilic
Buchnera aphidicola APS Tokyo†	-	Singles	-	-	-	-	Host associated	Mesophilic
Buchnera aphidicola APS Tokyo plasmids‡	-	-	-	-	-	-	-	-
Carsonella ruddii PV	-	-	-	-	-	-	Specialized, Host-associated[L]	Mesophilic [L]
Colwellia psychrerythraea 34H	Rod	Singles	-	Yes	-	Facultative	Specialized	Psychophilic (8C,-1C**)
Coxiella burnetii RSA 493†	Coccobacillus	Singles	-	-	-	Facultative	Multiple, Host-associated[L]	Mesophilic (37C)
Escherichia coli K-12 MG1655	Rod	Singles,Pairs	-	Yes	-	Facultative	Host-associated	Mesophilic (37C)
Klebsiella pneumoniae MGH 78578†	Rod	Singles,Pairs,Chains	-	Yes	-	Facultative	Multiple	Mesophilic (37C)
Legionella pneumophila Paris†	Rod	Singles	-	Yes	-	Aerobic	Host-associated	Mesophilic
Marinobacter aquaeolei VT8†	Rod	-	-	Yes	Moderate halophilic	Facultative	Aquatic	Mesophilic (30C)
Marinobacter aquaeolei VT8 plasmids‡	-	-	-	-	-	-	-	-
Photorhabdus luminescens laumondii TTO1	Rod	-	-	Yes	-	Facultative	Host-associated	Mesophilic
Pseudomonas aeruginosa PAO1	Rod	Singles	-	Yes	-	Aerobic	Multiple	Mesophilic (25-30C)
Pseudomonas syringae pv. tomato DC3000†	Rod	Singels	-	Yes	-	Aerobic	Multiple	Mesophilic
Pseudomonas syringae pv. tomato DC3000 plasmids‡	-	-	-	-	-	-	-	-
Psychromonas ingrahamii 37	Rod	-	No	No	-	Anaerobic	Aquatic	Psychrophilic
Psychrobacter arcticum 273-4	Rod	Pairs	-	No	-	-	Specialized	Psychrophilic (22C)
Psychrobacter cryohalolentis K5†	Rod	Pairs	-	No	Moderate halophilic	-	Multiple	Psychrophilic
Salmonella typhimurium LT2†	Rod	Singles,Pairs	-	Yes	-	Facultative	Host-associated	Mesophilic (37C)
Salmonella typhimurium LT2 plasmid pSLT	-	-	-	-	-	-	-	-
Shewanella denitrificans OS217	Rod	Singles,Pairs	-	Yes	-	Facultative	Aquatic	Mesophilic(20-25C)
Shewanella oneidensis MR-1†	Rod	-	-	Yes	-	Facultative	Multiple	Mesophilic
Shewanella oneidensis MR-1 plasmid pMR-1	-	-	-	-	-	-	-	-
Vibrio cholerae O1 biovar eltor N16961‡	Curved	Singles	-	Yes	-	Facultative	Aquatic	Mesophilic(20-30C)
Wigglesworthia glossinidia brevipalpis†	-	-	-	-	-	-	Host-associated	Mesophilic
Yersinia pestis CO92†	Rod	Singles	-	No**	-	Facultative	Multiple	Mesophilic (28-30C)
Yersinia pestis CO92 plasmids‡	-	-	-	-	-	-	-	-
Magnetococcus MC-1	Coccus	-	-	Yes	-	Facultative	Aquatic	Mesophilic [L]
Borrelia burgdorferi B31†	Spiral	-	No	Yes	-	Microaerophilic	Host-associated	Mesophilic
Treponema pallidum Nichols	Spiral	Singles	No	Yes	-	Anaerobic	Host-associated	Mesophilic
Fervidobacterium nodosum Rt17-B1	Rod	Singles,Pairs,Chains	No	Yes	-	Anaerobic	Specialized	Thermophilic (70C)
Thermosipho melanesiensis BI429	Rod	Singles,Pairs,Chains	-	-	-	Anaerobic	Host-associated**	Thermophilic (70C)
*Thermotoga petrophila*RKU-1	Rod	-	No	Yes	-	Anaerobic	Aquatic	Hyperthermophile (80C)
Thermotoga maritima MSB8	Rod	Singles	No	Yes	-	Anaerobic	Specialized	Hyperthermophile (80C)

Table 3.25: Average synthetic cost, GC-content, descriptions, and auxotrophy of microbes used (continued).

Gram Positive (PSORTDB)	Shape	Arrangement	Endospore	Motile	Salinity	Oxygen Requirement	Habitat	Temperature (optimal)
Arthrobacter aurescens TC1†	Rod,Coccus	-	No	Yes	Non-halophilc	Aerobic	Terrestrial	Mesophilic (30C)
Corynebacterium diphtheriae NCTC 13129	Rod	Singles, Pairs	No	No	-	Aerobic	Multiple	Mesophilic (37C)
Frankia sp. EAN1pec	Filaments	-	-	-	-	Aerobic	Multiple	Mesophilic
Kineococcus radiotolerans SRS30216†	Coccus	Singles,Clusters	No	Yes	-	Aerobic	Multiple	Mesophilic (32C)
Kineococcus radiotolerans SRS30216 plasmids‡	-	-	-	-	-	-	-	-
Mycobacterium leprae TN	Rod	-	No	No	Non-halophilc	Aerobic	Host-associated	Mesophilic (37C)
Mycobacterium tuberculosis H37Rv	Rod	Singles	No	No	-	Aerobic	Host-associated	Mesophilic (37C)
Nocardia farcinica IFM 10152†	Rod	Filaments	Yes	No	Non-halophilc	Aerobic	Multiple	Mesophilic (37C)
Nocardioides sp. JS614†	Rod	-	No	No	Non-halophilc	Aerobic	Terrestrial	Mesophilic (30C)
Rubrobacter xylanophilus DSM 9941	Rod	Singles	No	No	-	Aerobic	Specialized	Thermophilic (60C)
Saccharopolyspora erythraea NRRL 2338	Branched filament	Filaments	Yes	No	-	Aerobic	Terrestrial	Mesophilc (28C)
Salinispora tropica CNB-440	-	Singles,Clusters,Branched mycelium	Yes	No	-	Aerobic	Aquatic	Mesophilic (28C)
Streptomyces coelicolor A3(2)†	Branched filament	Filaments	Yes	No	-	Aerobic	Multiple	Mesophilic (25-35C)
Thermobifida fusca YX	Rod	Branched hyphae	Yes	No	-	Aerobic	Multiple	Thermophilic (50-55C)
Tropheryma whipplei Twist	Rod	Singles	No	NO	-	Aerobic	Host-associated	Mesophilic (37C)
Dehalococcoides ethenogenes 195	Irregular coccus	Singles	No	Yes	-	Anaerobic	Multiple	Mesophilic (35C)
Deinococcus radiodurans R1†‡	Coccus	Pairs, Tetrads	-	-	-	Aerobic	Terrestrial	Mesophilic (30-37C)
Bacillus anthracis Ames	Rod	Singles, Pairs, Chains	Yes	Yes	-	Facultative	Multiple	Mesophilic
Bacillus cereus ATCC 14579†	Rod	Chains	Yes	Yes	-	Aerobic	Terrestrial	Mesophilic (25-35C)
Bacillus subtilis 168	Rod	Singles, Chains	Yes	Yes	-	Facultative	Terrestrial	Mesophilic (25-35C)
Clostridium botulinum A ATCC 19397	Rod	Singles, Pairs, Chains	Yes	Yes	-	Anaerobic	Multiple	Mesophilc (37C)
Clostridium difficile 630†	Rod	Singles, Pairs, Chains	Yes	Yes	Non-halophilc	Anaerobic	Multiple	Mesophilc (37C)
Lactobacillus gasseri ATCC 33323	Rod	Singles	-	-	-	Facultative	Host-associated	Mesophilic (25-35C)
Lactococcus lactis subsp. lactis Il1403	Coccus	Pairs, Chains	-	-	-	Facultative	Multiple	Mesophilic (40C)
Listeria monocytogenes EGD-e	Rod	Singles, Chains	-	Yes	-	Facultative	Multiple	Mesophilic (30-37C)
Oceanobacillus iheyensis strain HTE831	Rod	-	Yes	Yes	halophilic	Aerobic	Multiple	Mesophilic (30C)
Staphylococcus aureus Newman	Coccus	Pairs, Clusters	No	NO	-	-	Host-associated**	Mesophilic (30-37C)
Staphylococcus aureus RF122	Coccus	Pairs, Clusters	No	No	-	Facultative	Host-associated	Mesophilic (30-37C)
Streptococcus mutans UA159	Coccus	Pairs, Chains	No	No	-	Facultative	Host-associated	Mesophilc (37C)
Streptococcus pneumoniae TIGR4	Coccus	Chains	No	NO	-	Facultative	Multiple	Mesophilic (30-35C)
Caldicellulosiruptor saccharolyticus DSM 8903	Rod	-	-	-	-	Anaerobic	Specialized	Thermophilic (70C)**

† Chromosomal only

‡ Weighted averages were used for multiple plasmids or chromosomes

Data from NCBI except for: ** from Integr8- http://www.ebi.ac.uk/integr8/

L from respective literature source

Table 3.25: Average synthetic cost, GC-content, descriptions, and auxotrophy of microbes used (continued).

Gram Negative (PSORTDB)	Pathogenic - Disease	Other notes *,**
Solibacter usitatus Ellin6076	No	secrete high MW proteins
Aquifex aeolicus VF5†	No	deep branch microbe
Aquifex aeolicus VF5 plasmid ece1	-	-
Cytophaga hutchinsonii ATCC 33406	No	cellulose degradation, gliding motlity, no flagella
Flavobacterium johnsoniae UW101	No	can use chitin, cellulose, lignin
Flavobacterium psychrophilum JIP02/86	Salmoid fish - Cold water disease, Trout Fry syndrome	infection occurs 3 to 15C, secretes 13 proteases
Salinibacter ruber DSM 13855†	No	proteins have adapted to high intracellular salt
Sulcia muelleri GWSS*	-	Metabolically complements *Baumaninia*
Chlamydia trachomatis D/UW-3/CX	Human - Pharyngitis, bronchitis and pneumonitis	Intracellular active, Extraceullar inactive phase
Chlamydophila pneumoniae AR39	Human - Pharyngitis, bronchitis and pneumonitis	Intracellular active, Extraceullar inactive phase
Roseiflexus sp. RS-1	No	anoxygenic phototroph
Roseiflexus castenholzii DSM 13941	No	growth in aerobic dark conditions
Prochlorococcus marinus AS9601	No	highly abundant ocean microbe
Synechocystis sp. PCC 6803†	No	phototactic movement, oxygenic photosynthesis
Synechocystis sp. PCC 6803 plasmids‡	-	-
Thermosynechococcus elongatus BP-1	No	obligate photoautotroph, many type II introns
Trichodesmium erythraeum IMS101	No	Nitrogen fixation in diazocyte
Thermus thermophilus HB27†	No	halotolerant, Type IV pilus
Mycoplasma genitalium G37	Human, Animal, Insect - Urogenital or Respiratory Infections	obligate parasite (intracellular)
Mycoplasma pneumoniae M129	Human - Atypical pneumonia in older children & young adults	attachment organelle, can be grown alone in rich media
Fusobacterium nucleatum ATCC 25586	Human, Animal - Periodontal disease, inflamations	obligate anaerobe - associates with other organisms
Rhodopirellula baltica SH 1	No	has pirellulosome, petidoglycan production lost
Caulobacter cresentus CB15	No	asymetric lifestyle, stalk and swarmer cells
Ehrlichia chaffeensis str. Arkansas	Human,animal - Monocytic ehrlichiosis	obligate intracellular pathogen, monocyte vacuoles
Hodgkinia cicadicola Dsem*	No	Small genome yet high %GC, recoded Stop UGA to Trp
Pelagibacter ubique HTCC1062	No	smallest genome of free living org.- no transposons, HT DNA
Rhodobacter sphaeroides 2.4.1†‡	No	N fixation; aerobic,anaerobic respiration; photsyntesis
Rhodobacter sphaeroides 2.4.1 plasmids‡	-	-
Rickettsia prowazekii Madrid E	Human - Typhus, Mediterranean spotted fever	intracellular parasite in cytoplasm of macrophages
Silicibacter sp. TM1040†	No	*Ruegeria* sp. TM1040; sybiosis with dinoflagellates
Silicibacter sp. TM1040 plasmids‡	-	-
Wolbachia pipientis wMel	No	obligate intracellular symbiont of fruit fly
Bordetella pertussis Tohama I	Human - Respiratory diseases	Whopping cough
Burkholderia xenovorans strain LB400‡	Human,Plants - Oppurtunistic infections	v large genome, aerobic PCB degrader, N fixation
Neisseria gonorrhoeae FA 1090	Human - Gonorrhea	naturally competent
Ralstonia eutropha JMP134†‡	-	aerobic heterotroph bioremediation chorlianted aeromatics
Ralstonia eutropha JMP134 plasmids‡	-	-
Anaeromyxobacter dehalogenans 2CP-C	No	-
Geobacter metallireducens GS-15†	No	uses iron in environment as e- acceptor
Myxococcus xanthus DK 1622	No	1500 gene duplications
Desulfotalea psychrophila LSv54†	No	carbon and sulfur cycles

Table 3.25: Average synthetic cost, GC-content, descriptions, and auxotrophy of microbes used (continued).

Gram Negative (PSORTDB)	Pathogenic - Disease	Other notes *,**
Arcobacter butzleri RM4018	Human - Gastroenteritis and bacteremia	growth at various oxygen concentrations
Campylobacter jejuni RM1221	Human - Food poisoning	unique LOS and capsule loci
Helicobacter pylori J99	Human - Gastric inflamation, peptic ulcers	produces urease to suvrive low pH
Wolinella succinogenes DSM1740	No	non pathogenic but contains serveral virulence genes
Acinetobacter sp. ADP1	Human - Nonsocomial infection in immunocompromised	high natural competency
Baumannia cicadellinicola Hc	No	Metabolically complements *Sulcia*
Blochmannia floridanus	No	ant symbiont - provides sulfur and nitrogen
Buchnera aphidicola APS Tokyo†	No	symbiont of aphid Acyrthosiphon pisum
Buchnera aphidicola APS Tokyo plasmids‡	-	-
Carsonella ruddii PV	No	obligate endosymbion of psyllids
Colwellia psychrerythraea 34H	No	active down to -5C, extracellular proteases
Coxiella burnetii RSA 493†	Animal, Human - Q fever	obligate intracellular, replicates in phagocyte
Escherichia coli K-12 MG1655	No	derived from W1485, cured of lambda and F plasmid
Klebsiella pneumoniae MGH 78578†	Human, Animal - Pneumonia and UTIs	strain genetically unmalleable - multiple restriction systems
Legionella pneumophila Paris†	Animal - Legionnaire's diesease	Protozoa act as reservoir. *Legionella sp.* require Cys
Marinobacter aquaeolei VT8†	No	degrades hydrocarbons, iron & carbon cycles
Marinobacter aquaeolei VT8 plasmids‡	-	-
Photorhabdus luminescens laumondii TTO1	Insect - Toxemia and septicemia	symbiosis with soil entomopathogenic nematodes
Pseudomonas aeruginosa PAO1	Human - Opportunistic infections	Highly motile - single polar flagella
Pseudomonas syringae pv. tomato DC3000†	Plant - Plant rot	resistent to stresses - plant pathogen interactions
Pseudomonas syringae pv. tomato DC3000 plasmids‡	-	-
Psychromonas ingrahamii 37	No	Grows at -12C lowest to date
Psychrobacter arcticum 273-4	No	isolated from Siberian permafrost core
Psychrobacter cryohalolentis K5†	-	Growth at -10C, siberian permafrost cryopeg
Salmonella typhimurium LT2†	Human - Gastroenteritis and food poisoning	numerous prophages some with virulence genes
Salmonella typhimurium LT2 plasmid pSLT	-	-
Shewanella denitrificans OS217	No	vigorous denitrifier
Shewanella oneidensis MR-1†	Human - Rare oppurtunisitic pathogen	can use a variety of e- acceptors
Shewanella oneidensis MR-1 plasmid pMR-1	-	-
Vibrio cholerae O1 biovar eltor N16961‡	Human, Primate - Cholera	hemonlysin production
Wigglesworthia glossinidia brevipalpis†	No	Obligate endosymbiont Tsetse fly
Yersinia pestis CO92†	Human, Rodent - Bubonic plague	transferred amongst rodents by fleas
Yersinia pestis CO92 plasmids‡	-	-
Magnetococcus MC-1	No	Magnetosomes
Borrelia burgdorferi B31†	Human - Lyme Disease	Midgut of Ixodes ticks
Treponema pallidum Nichols	Human - Syphilis	Obligate human parasite
Fervidobacterium nodosum Rt17-B1	No	obligate anaerobe**
Thermosipho melanesiensis BI429	-	found on gills of deep sea mussel *Bathymodiolus brevior*
*Thermotoga petrophila*RKU-1	No	anaerotic heterotroph, reduces sulfur to H_2S
Thermotoga maritima MSB8	No	horizontal gene transfer of Archeal genes

Table 3.25: Average synthetic cost, GC-content, descriptions, and auxotrophy of microbes used (continued).

Gram Positive (PSORTDB)	Pathogenic - Disease	Other notes *,**
Arthrobacter aurescens TC1†	None	atrazine as sole C,N, energy source
Corynebacterium diphtheriae NCTC 13129	Human - Diptheria	snapping division
Frankia sp. EAN1pec	No	symbiotic relationship with plants, fix nitrogen, hyphal
Kineococcus radiotolerans SRS30216†	No	thick extracellular polymer shell
Kineococcus radiotolerans SRS30216 plasmids‡	-	-
Mycobacterium leprae TN	Human - Leprosy	massive genome reduction many pseudogenes
Mycobacterium tuberculosis H37Rv	Human - Tuberculosis	obligate aerobe
Nocardia farcinica IFM 10152†	Human, Animal - Nocardiosis	soil saprophyte
Nocardioides sp. JS614†	No	degrades vinyl chloride
Rubrobacter xylanophilus DSM 9941	No	reistant to gamma radiation
Saccharopolyspora erythraea NRRL 2338	No	obligate aerobe, produces erythromycin A
Salinispora tropica CNB-440	No	requires sea water for growth, anticancer metabolites
Streptomyces coelicolor A3(2)†	No	linear chromosome
Thermobifida fusca YX	Human - Mushroom worker's & farmer's lung disease	
Tropheryma whipplei Twist	Human - Whipple's disease	slow growth - nonculturable
Dehalococcoides ethenogenes 195	No	degrade clorinated organics
Deinococcus radiodurans R1†‡	No	most radiation resistant organism
Bacillus anthracis Ames	Animal - Anthrax	endospores germinate in alveolar macrophages
Bacillus cereus ATCC 14579†	Human - Food poisoning	two illnesses from two sources
Bacillus subtilis 168	No	well studied spore forming bacteria
Clostridium botulinum A ATCC 19397	Human - Botulism	heat resistant spores
Clostridium difficile 630†	Human - diarrhea, pseudomembraneous colitis	produces toxin A and B
Lactobacillus gasseri ATCC 33323	No	probiotic
Lactococcus lactis subsp. lactis Il1403	No	cheese production
Listeria monocytogenes EGD-e	Human - Listeriosis	enteroinvasive
Oceanobacillus iheyensis strain HTE831	No	salt tolerant alkaliphile, many H,Na,K pumps
Staphylococcus aureus Newman	Human, Animal - Skin infections, pneumonia, endocarditis	robust pathogen, four prophages, 2 PAIs
Staphylococcus aureus RF122	Human, Animal - Mastitis	bovine mastitis
Streptococcus mutans UA159	Human - Dental caries	cariogenic
Streptococcus pneumoniae TIGR4	Human - Mutiple(Pneumonia, Meningitis, others)	highly virulent capsular serotype 4 clinical isolate
Caldicellulosiruptor saccharolyticus DSM 8903	No	cellulytic, produces hydrogen, strict anaerobe

† Chromosomal only

‡ Weighted averages were used for multiple plasmids or chromosomes

Data from NCBI except for: ** from Integr8- http://www.ebi.ac.uk/integr8/

L from respective literature source

Table 3.25: Average synthetic cost, GC-content, descriptions, and auxotrophy of microbes used (continued).

	Amino Acid Auxotrophies		
Gram Negative (PSORTDB)	**Number*****	**Number**	**Source**
Solibacter usitatus Ellin6076	-	2	Ward et al AEM 2009
Aquifex aeolicus VF5†	-	None	Deckert et al Nature 1998 - growth in media with only inorganic components, autotroph - obtains all C by fixing from CO_2
Aquifex aeolicus VF5 plasmid ece1	-	-	-
Cytophaga hutchinsonii ATCC 33406	-	None	Xie et al AEM 2007 - growth on minimal media plus
Flavobacterium johnsoniae UW101	-	None	McBride et al AEM 2009 - growth on minimial media plus glucose, cellulose - genome has all enyzmes necessary
Flavobacterium psychrophilum JIP02/86	-	?	Bernadet & Kerouault 1989 - many carbohydate enzmes missing, inablity to utilize several carbon sources
Salinibacter ruber DSM 13855†	-	likely	Anton et al Int J Syst Evol Microbiol 2002 - difficulty in defining minimal media - No growth on simple sugars,organic acids
Sulcia muelleri GWSS*	-	12	McCutcheon&Moran PNAS 2007
Chlamydia trachomatis D/UW-3/CX	-	6	Heizer et al Mol Biol Evol 2006
Chlamydophila pneumoniae AR39	10	6	Heizer et al Mol Biol Evol 2006
Roseiflexus sp. RS-1	-		
Roseiflexus castenholzii DSM 13941	-	None	Hanada et al Int J Syst Evol Microbiol 2002 - growth on minimal media + citrate,lactate, or glucose
Prochlorococcus marinus AS9601	-	None	Kettler et al Plos Genetics 2007
Synechocystis sp. PCC 6803†	-	None	Heizer et al Mol Biol Evol 2006, grows in minimal media BG-11 Rippka et al J Gen Micro 1979
Synechocystis sp. PCC 6803 plasmids‡	-	-	-
Thermosynechococcus elongatus BP-1	-		
Trichodesmium erythraeum IMS101	-	None	Chen et al J Phycol 1996 - grows in defined media (without amino acids)
Thermus thermophilus HB27†	-	None	Heizer et al Mol Biol Evol 2006
Mycoplasma genitalium G37	-	~20	Fraser et al Science 1995 - has *glyA*
Mycoplasma pneumoniae M129	10		
Fusobacterium nucleatum ATCC 25586	-		
Rhodopirellula baltica SH 1	-	None	Glockner et al PNAS 2003
Caulobacter cresentus CB15	-	None	Ferber & Ely Mol Gen Genet 1982 - grows on minimal + glucose + ammonia
Ehrlichia chaffeensis str. Arkansas	9		
Hodgkinia cicadicola Dsem*	-	11	McCutcheon et al PLOS Genetics 2009
Pelagibacter ubique HTCC1062	-	None	Giovannoni et al Science 2005
Rhodobacter sphaeroides 2.4.1†‡	-	None	Gray and Escalante-Semerena JBac 2009 - growth in Sistroms minimal A - Asp&Glu and succinate as sole carbon source
Rhodobacter sphaeroides 2.4.1 plasmids‡	-	-	-
Rickettsia prowazekii Madrid E	-		
Silicibacter sp. TM1040†	-	None	Miller&Belas Enviro Micro - minimal + glycerol; Miller et al AEM 2004 - minimal + glucose,glycerol,succinate
Silicibacter sp. TM1040 plasmids‡	-	-	-
Wolbachia pipientis wMel	-		
Bordetella pertussis Tohama I	1	1	Parkhill et al Nature Genetics 2003 - cysteine biosynthesis and sulfur assimilation genes
Burkholderia xenovorans strain LB400‡	-	None	Agullo et al FEMS Microbiol Lett 2007 - growth on minimal media + glucose
Neisseria gonorrhoeae FA 1090	None	likely none	Dempsey et al JBact 1991 - requires Pro for growth, Biocyc - Pro biosyn genes present
Ralstonia eutropha JMP134†‡	-	None	Muller et al Appl Microbiol Biotechnol - growth in minimal media with pyruvate or phenol as sole carbon source
Ralstonia eutropha JMP134 plasmids‡	-	-	-
Anaeromyxobacter dehalogenans 2CP-C	-	likely none	Treude et al FEMS 2003 - related strain grown on minimal media, See Note 1
Geobacter metallireducens GS-15†	-		
Myxococcus xanthus DK 1622	-		
Desulfotalea psychrophila LSv54†	-		

Table 3.25: Average synthetic cost, GC-content, descriptions, and auxotrophy of microbes used (continued).

	Amino Acid Auxotrophies		
Gram Negative (PSORTDB)	**Number*** **	**Number**	**Source**
Arcobacter butzleri RM4018	-	None	Miller et al PLOS One 2007 - 96.5% homologs aa synthesis, Houf et al IJSEM 2009 - growh on minimal media
Campylobacter jejuni RM1221	2	likely	Grown in defined media containing several amino acids, no growth on sugars as major carbon source including glucose
Helicobacter pylori J99	7	9	Doig et al Micro & Mol Bio Review 1999
Wolinella succinogenes DSM1740	-	None	Baar et al PNAS 2003 - enzymes for all amino acids present, growth in minimal media + fumarate
Acinetobacter sp. ADP1	-	None	Metzgar et al Nucleic Acid Res 2004 - grows on minimal MS media with glucose
Baumannia cicadellinicola Hc	-	18	McCutcheon&Moran PNAS 2007
Blochmannia floridanus	-	5	Zientz et al Micro & Mol Bio Review 2004
Buchnera aphidicola APS Tokyo†	-	9	Zientz et al Micro & Mol Bio Review 2004
Buchnera aphidicola APS Tokyo plasmids‡	-	-	-
Carsonella ruddii PV	-	14-16	Tamames et al BMC Evol Bio, Nakabachi et al Science 2006
Colwellia psychrerythraea 34H	-	likely none	Methe et al PNAS 2005, Biocyc-aa biosyn genes present, Jung et al IJSEM 2006-malate,acetate sole C source
Coxiella burnetii RSA 493†	10	11	Seshadri et al PNAS 2003
Escherichia coli K-12 MG1655	None	None	Heizer et al Mol Biol Evol 2006
Klebsiella pneumoniae MGH 78578†	-	yes	Bob Bender, personal communication - will not grow on minimal media requires rich media
Legionella pneumophila Paris†	7	9	Keen & Hoffman Cur Micro 1984
Marinobacter aquaeolei VT8†	-	None	Huu et al Int J Syst Bacteriology 1999 - aerobic growth, growth on minimal media + citrate, others as sole C source
Marinobacter aquaeolei VT8 plasmids‡	-	-	-
Photorhabdus luminescens laumondii TTO1	-	None	Ciche et al AEM 2003,Fischer-Le Saux Inter J Syst Biol 1999 - minimal media various sole carbon sources, See Note 2
Pseudomonas aeruginosa PAO1	None	None	Stover et al Nature 2000 - grows on minimal media
Pseudomonas syringae pv. tomato DC3000†	-	None	Roine et al PNAS 1997 - growth in minimal media such as M9 + manitol
Pseudomonas syringae pv. tomato DC3000 plasmids‡	-	-	-
Psychromonas ingrahamii 37	-		
Psychrobacter arcticum 273-4	-		
Psychrobacter cryohalolentis K5†	-		
Salmonella typhimurium LT2†	None	None	Gutnick et al Jbac 1969 - growth on minimal media with a variety of single carbon sources
Salmonella typhimurium LT2 plasmid pSLT	-	-	-
Shewanella denitrificans OS217	-	?	Inconclusive - couldn't find appropriate literature source
Shewanella oneidensis MR-1†	-	None	Tang et al Biotechnol Bioeng 2006 - growth aerobically with lactate as sole carbon source
Shewanella oneidensis MR-1 plasmid pMR-1	-	-	-
Vibrio cholerae O1 biovar eltor N16961‡	None	None	Yildiz & Shoolnik PNAS 1999 - aerobic growth on minimal media with glucose
Wigglesworthia glossinidia brevipalpis†	-	16	Zientz et al Micro & Mol Bio Review 2004
Yersinia pestis CO92†	None	5,3	Perry & Fertherston Clin Micro Review 1997 - some revert, Parkhill et al Nature 2003
Yersinia pestis CO92 plasmids‡	-	-	-
Magnetococcus MC-1	-	None	Meldrum et al Proc R Soc Lond B 1993 - grown in O or S,O gradient w acetate or bicarbonate as sole carbon source, resp.; Williams et al AEM,2006 - chemolithoautotroph
Borrelia burgdorferi B31†	10		
Treponema pallidum Nichols	-		
Fervidobacterium nodosum Rt17-B1	-		
Thermosipho melanesiensis BI429	-		
*Thermotoga petrophila*RKU-1	-		
Thermotoga maritima MSB8	-		

Table 3.25: Average synthetic cost, GC-content, descriptions, and auxotrophy of microbes used (continued).

	Amino Acid Auxotrophies		
Gram Positive (PSORTDB)	**Number*** **	**Number**	**Source**
Arthrobacter aurescens TC1†	-		
Corynebacterium diphtheriae NCTC 13129	*2*	None	Cerdeno-Tarraga et al Nucleic Acid Research 2003
Frankia sp. EAN1pec	-		
Kineococcus radiotolerans SRS30216†	-		
Kineococcus radiotolerans SRS30216 plasmids‡	-	-	
Mycobacterium leprae TN	-	1	Cole et al Nature 2001 - *met*C pseudogene
Mycobacterium tuberculosis H37Rv	1	None	Cole et al Nature 1998, Yu et al says His lacks genes
Nocardia farcinica IFM 10152†	-		
Nocardioides sp. JS614†	-		
Rubrobacter xylanophilus DSM 9941	-		
Saccharopolyspora erythraea NRRL 2338	-		
Salinispora tropica CNB-440	-		
Streptomyces coelicolor A3(2)†	-		
Thermobifida fusca YX			
Tropheryma whipplei Twist	-		
Dehalococcoides ethenogenes 195	-		
Deinococcus radiodurans R1†‡	-		
Bacillus anthracis Ames	None		
Bacillus cereus ATCC 14579†	-		
Bacillus subtilis 168	-	None	Heizer et al Mol Biol Evol 2006
Clostridium botulinum A ATCC 19397	8		
Clostridium difficile 630†	-		
Lactobacillus gasseri ATCC 33323	-		
Lactococcus lactis subsp. lactis Il1403	-		
Listeria monocytogenes EGD-e	-	6	Premaratne et al AEM 1991
Oceanobacillus iheyensis strain HTE831	-		
Staphylococcus aureus Newman	2		
Staphylococcus aureus RF122	2		
Streptococcus mutans UA159	-		
Streptococcus pneumoniae TIGR4	3		
Caldicellulosiruptor saccharolyticus DSM 8903	-		

*** from Yu et al Infection,Genetics, and Evolution (2009) 9:514-517

Note 1: Biocyc shows all amino acid synthesis genes present except: PheA(Q2IISO in UniprotkB), SerB (many HAD hydrolase-difficult to anotate SerB), and tyrB (shares activity with ilvE (Q2IF34) & aspC (Q2IIM4))

Note 2: Biocyc shows all amino acid synthesis genes present except: PheA(Q7N793in UniprotkB)

Table 3.26: Comparison of the ASC of extracellular protein with total proteins in 26 Gram-negative bacteria.*

				Total Proteins		Extracellular Proteins							U test (ASC)		% Savings	
Organism	Group	% GC	Number Proteins	ASC per aa	Mass per aa	#	# ASC < Org.†	% ASC < Org.†	ASC per aa	Mass per aa	Avgerage Percentile	ATP Savings	vs Cyto. Proteins	vs All Proteins	ASC per aa	Mass per aa
Arcobacter butzleri RM4018	Epsilonproteobacteria	27	2259	24.608	113.851	13	12	92.31	21.736	108.451	90	2.87	1.75E-07	1.12E-07	11.67	4.74
Flavobacterium johnsoniae UW101	Bacteroidetes/Chlorobi	34.11	5017	24.443	113.337	35	29	82.86	23.245	109.862	74.5	1.2	3.27E-05	5.72E-07	4.9	3.07
Trichodesmium erythraeum IMS101	Cyanobacteria	34.14	4451	24.086	112.993	46	46	100	21.195	106.929	93	2.89	1.75E-22	9.50E-24	12	5.37
Colwellia psychrerythraea 34H	Gammaproteobacteria	38	4910	23.864	111.778	37	31	83.78	21.894	107.82	80.3	1.97	1.14E-07	2.17E-10	8.26	3.54
Cytophaga hutchinsonii ATCC 33406	Bacteroidetes/Chlorobi	38.8	3785	24.53	112.817	40	39	97.50	22.602	106.967	85.2	1.93	2.28E-12	1.67E-14	7.86	5.19
Acinetobacter sp. ADP1	Gammaproteobacteria	40.4	3325	23.947	112.141	11	10	90.91	22.095	106.614	81	1.85	1.35E-03	3.80E-04	7.73	4.93
Photorhabdus luminescens laumondii TTO1	Gammaproteobacteria	42.8	4683	23.863	112.161	33	31	93.94	22.02	108.192	83.5	1.84	2.23E-10	3.11E-11	7.72	3.54
Aquifex aeolicus VF5[1]	Aquificae	43.3	1560	24.774	114.13	11	10	90.91	23.214	111.694	81.8	1.56	7.80E-04	2.78E-04	6.3	2.13
Shewanella oneidensis MR-1	Gammaproteobacteria	45.9	4471	23.54	111.158	36	34	94.44	21.357	106.088	89.1	2.18	4.85E-16	6.09E-16	9.27	4.56
Synechocystis sp. PCC 6803	Cyanobacteria	47.3	3564	23.583	111.507	12	10	83.33	22.151	107.61	77.7	1.43	2.50E-03	9.17E-04	6.07	3.49
Vibrio cholerae O1 biovar eltor N16961	Gammaproteobacteria	47.5	3835	23.759	111.782	50	45	90.00	22.278	108.608	77.3	1.48	8.97E-09	2.96E-11	6.23	2.84
Wolinella succinogenes DSM 1740	Epsilonproteobacteria	48.5	2044	24.042	112.206	13	13	100	21.439	107.44	93.7	2.6	3.40E-08	5.48E-08	10.83	4.25
Escherichia coli K-12 MG1655	Gammaproteobacteria	50.8	4333	23.555	111.005	16	16	100	20.652	104.9	93.4	2.9	2.19E-09	1.96E-09	12.32	5.5
Salmonella typhimurium LT2	Gammaproteobacteria	52.2	4527	23.532	110.908	58	51	87.93	21.676	107.196	81.4	1.86	1.43E-14	2.22E-16	7.89	3.35
Magnetococcus MC-1	Other	54.2	3716	23.277	110.848	37	37	100	20.261	105.888	94.2	3.02	1.08E-20	2.00E-20	12.96	4.47
Rhodopirellula baltica SH 1	Plantomycetes	55.4	7325	23.011	110.75	24	21	87.50	20.872	106.693	85.9	2.14	1.54E-09	1.15E-09	9.3	3.66
Marinobacter aquaeolei VT8	Gammaproteobacteria	57	4272	23.077	110.776	19	19	100	20.671	106.006	93.2	2.41	1.95E-11	7.50E-11	10.43	4.31
Pseudomonas syringae pv. tomato DC3000	Gammaproteobacteria	58.3	5608	23.047	110.075	84	70	83.33	21.581	107.296	77.1	1.47	9.94E-18	1.25E-17	6.36	2.52
Silicibacter sp. TM1040	Alphaproteobacteria	60.1	3864	22.754	109.053	29	29	100	20.397	104.447	92.7	2.36	4.34E-16	2.22E-15	10.36	4.22
Roseiflexus castenholzii DSM 13941	Chloroflexi	60.7	4330	23.072	109.936	22	22	100	21.316	105.67	89.3	1.76	1.13E-10	2.02E-10	7.61	3.88
Burkholderia xenovorans LB400	Betaproteobacteria	62.6	8702	22.762	108.813	45	39	86.67	21.063	103.631	79.1	1.7	5.80E-12	1.45E-11	7.46	4.76
Ralstonia eutropha JMP134	Betaproteobacteria	64.5	6446	22.648	108.437	30	27	90.00	20.62	104.01	86.3	2.03	2.17E-12	6.63E-12	8.95	4.08
Pseudomonas aeruginosa PAO1	Gammaproteobacteria	66.6	5667	22.881	109.787	48	41	85.42	21.392	105.903	79.2	1.49	2.89E-12	3.56E-12	6.51	3.54
Caulobacter cresentus CB15	Alphaproteobacteria	67.2	3737	22.423	107.884	27	22	81.48	20.955	103.973	79	1.47	2.95E-07	1.95E-07	6.55	3.63
Rhodobacter sphaeroides 2.4.1	Alphaproteobacteria	68.8	4242	22.481	108.15	29	29	100	20.251	102.962	90.1	2.23	7.22E-15	9.25E-14	9.92	4.8
Anaeromyxobacter dehalogenans 2CP-C	Deltaproteobacteria	74.9	4346	21.913	106.92	28	27	96.43	20.249	102.949	84.4	1.66	2.54E-12	2.00E-10	7.59	3.71

* Each is capable of aerobic growth and has the capacity to synthesis all 20 amino acids (See notes in Table 3.25).

† The number and % of extracellular proteins that have ASC that are less than the mean organismal ASC.

[1] Not graphed in Fig. 3.5B or used in later analysis. Hyperthermophile (96C) - does not follow %GC cost trends of other organisms (See Fig. 3.5A) possibly due to different amino acid biases needed for protein folding at higher temperatures. Many thermophiles had weak cost selection relative to Mesophiles (149); Gibbs free energy of amino acid formation is radically different in hydrothermal vents (Table 3.1) (8).

References

1. Genome Sequence of the Pea Aphid Acyrthosiphon pisum. PLoS Biol **8:**e1000313.
2. http://redpoll.pharmacy.ualberta.ca/CCDB/cgi-bin/STAT_NEW.cgi.
3. **Akashi, H., and T. Gojobori.** 2002. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc Natl Acad Sci U S A **99:**3695-700.
4. **Akman Gunduz, E., and A. E. Douglas.** 2009. Symbiotic bacteria enable insect to use a nutritionally inadequate diet. Proc Biol Sci **276:**987-91.
5. **Akman, L., A. Yamashita, H. Watanabe, K. Oshima, T. Shiba, M. Hattori, and S. Aksoy.** 2002. Genome sequence of the endocellular obligate symbiont of tsetse flies, Wigglesworthia glossinidia. Nat Genet **32:**402-7.
6. **Alm, R. A., U. H. Stroeher, and P. A. Manning.** 1988. Extracellular proteins of Vibrio cholerae: nucleotide sequence of the structural gene (hlyA) for the haemolysin of the haemolytic El Tor strain 017 and characterization of the hlyA mutation in the non-haemolytic classical strain 569B. Mol Microbiol **2:**481-8.
7. **Alves, R., and M. A. Savageau.** 2005. Evidence of selection for low cognate amino acid bias in amino acid biosynthetic enzymes. Mol Microbiol **56:**1017-34.
8. **Amend, J. P., and E. L. Shock.** 1998. Energetics of amino acid synthesis in hydrothermal ecosystems. Science **281:**1659-62.
9. **Andrade, M. A., S. I. O'Donoghue, and B. Rost.** 1998. Adaptation of protein surfaces to subcellular location. J Mol Biol **276:**517-25.
10. **Arnold, R., S. Brandmaier, F. Kleine, P. Tischler, E. Heinz, S. Behrens, A. Niinikoski, H. W. Mewes, M. Horn, and T. Rattei.** 2009. Sequence-based prediction of type III secreted proteins. PLoS Pathog **5:**e1000376.
11. **Atkinson, D. E.** 1977. Cellular Energy Metabolism and Its Regulation, vol. Academic Press, New York.
12. **Baar, C., M. Eppinger, G. Raddatz, J. Simon, C. Lanz, O. Klimmek, R. Nandakumar, R. Gross, A. Rosinus, H. Keller, P. Jagtap, B. Linke, F. Meyer, H. Lederer, and S. C. Schuster.** 2003. Complete genome sequence and analysis of Wolinella succinogenes. Proc Natl Acad Sci U S A **100:**11690-5.
13. **Barnhart, M. M., and M. R. Chapman.** 2006. Curli biogenesis and function. Annu Rev Microbiol **60:**131-47.
14. **Bauchop, T., and S. R. Elsden.** 1960. The growth of micro-organisms in relation to their energy supply. J Gen Microbiol **23:**457-69.
15. **Baudouin-Cornu, P., K. Schuerer, P. Marliere, and D. Thomas.** 2004. Intimate evolution of proteins. Proteome atomic content correlates with genome base composition. J Biol Chem **279:**5421-8.
16. **Baudouin-Cornu, P., Y. Surdin-Kerjan, P. Marliere, and D. Thomas.** 2001. Molecular evolution of protein atomic composition. Science **293:**297-300.
17. **Berg, H. C.** 2000. Constraints on models for the flagellar rotary motor. Philos Trans R Soc Lond B Biol Sci **355:**491-501.
18. **Berg, H. C.** 1974. Dynamic properties of bacterial flagellar motors. Nature **249:**77-9.

19. **Berg, H. C.** 2003. The rotary motor of bacterial flagella. Annu Rev Biochem **72:**19-54.
20. **Berg, H. C., and L. Turner.** 1993. Torque generated by the flagellar motor of Escherichia coli. Biophys J **65:**2201-16.
21. **Beskhlebnaya, V. A., E. V. Trinchina, P. Aprikyan, V. Chesnokova, and E. V. Sokurenko.** 2006. Molecular genetic analysis of Escherichia coli type I adhesins. Bull Exp Biol Med **141:**339-42.
22. **Bingle, L. E., C. M. Bailey, and M. J. Pallen.** 2008. Type VI secretion: a beginner's guide. Curr Opin Microbiol **11:**3-8.
23. **Bragg, J. G., and A. Wagner.** 2007. Protein carbon content evolves in response to carbon availability and may influence the fate of duplicated genes. Proc Biol Sci **274:**1063-70.
24. **Bragg, J. G., and A. Wagner.** 2009. Protein material costs: single atoms can make an evolutionary difference. Trends Genet **25:**5-8.
25. **Brandes, N., A. Rinck, L. I. Leichert, and U. Jakob.** 2007. Nitrosative stress treatment of E. coli targets distinct set of thiol-containing proteins. Mol Microbiol **66:**901-14.
26. **Capitani, G., O. Eidam, R. Glockshuber, and M. G. Grutter.** 2006. Structural and functional insights into the assembly of type 1 pili from Escherichia coli. Microbes Infect **8:**2284-90.
27. **Cedano, J., P. Aloy, J. A. Perez-Pons, and E. Querol.** 1997. Relation between amino acid composition and cellular location of proteins. J Mol Biol **266:**594-600.
28. **Chen, X., and H. C. Berg.** 2000. Torque-speed relationship of the flagellar rotary motor of Escherichia coli. Biophys J **78:**1036-41.
29. **Chevance, F. F., and K. T. Hughes.** 2008. Coordinating assembly of a bacterial macromolecular machine. Nat Rev Microbiol **6:**455-65.
30. **Chirwa, N. T., and M. B. Herrington.** 2003. CsgD, a regulator of curli and cellulose synthesis, also regulates serine hydroxymethyltransferase synthesis in Escherichia coli K-12. Microbiology **149:**525-35.
31. **Chung, C. H., H. E. Ives, S. Almeda, and A. L. Goldberg.** 1983. Purification from Escherichia coli of a periplasmic protein that is a potent inhibitor of pancreatic proteases. J Biol Chem **258:**11032-8.
32. **Clark, M. A., N. A. Moran, and P. Baumann.** 1999. Sequence evolution in bacterial endosymbionts having extreme base compositions. Mol Biol Evol **16:**1586-98.
33. **Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, 3rd, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M. A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell.** 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature **393:**537-44.

34. **Cornelis, G. R.** 2006. The type III secretion injectisome. Nat Rev Microbiol **4:**811-25.
35. **Craig, C. L., and R. S. Weber.** 1998. Selection costs of amino acid substitutions in ColE1 and ColIa gene clusters harbored by Escherichia coli. Mol Biol Evol **15:**774-6.
36. **de Bivort, B. L., E. O. Perlstein, S. Kunes, and S. L. Schreiber.** 2009. Amino acid metabolic origin as an evolutionary influence on protein sequence in yeast. J Mol Evol **68:**490-7.
37. **DeRosier, D. J.** 1995. Spinning tails. Curr Opin Struct Biol **5:**187-93.
38. **DeRosier, D. J.** 1998. The turn of the screw: the bacterial flagellar motor. Cell **93:**17-20.
39. **Diaz-Mejia, J. J., M. Babu, and A. Emili.** 2009. Computational and experimental approaches to chart the Escherichia coli cell-envelope-associated proteome and interactome. FEMS Microbiol Rev **33:**66-97.
40. **Douglas, A. E.** 2006. Phloem-sap feeding by animals: problems and solutions. J Exp Bot **57:**747-54.
41. **Dunker, A. K., C. J. Oldfield, J. Meng, P. Romero, J. Y. Yang, J. W. Chen, V. Vacic, Z. Obradovic, and V. N. Uversky.** 2008. The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics **9 Suppl 2:**S1.
42. **Duong, F., E. Bonnet, V. Geli, A. Lazdunski, M. Murgier, and A. Filloux.** 2001. The AprX protein of Pseudomonas aeruginosa: a new substrate for the Apr type I secretion system. Gene **262:**147-53.
43. **Easom, C. A., and D. J. Clarke.** 2008. Motility is required for the competitive fitness of entomopathogenic Photorhabdus luminescens during insect infection. BMC Microbiol **8:**168.
44. **Elser, J. J., W. F. Fagan, S. Subramanian, and S. Kumar.** 2006. Signatures of ecological resource availability in the animal and plant proteomes. Mol Biol Evol **23:**1946-51.
45. **Farmer, I. S., and C. W. Jones.** 1976. The energetics of Escherichia coli during aerobic growth in continuous culture. Eur J Biochem **67:**115-22.
46. **Fauchon, M., G. Lagniel, J. C. Aude, L. Lombardia, P. Soularue, C. Petat, G. Marguerie, A. Sentenac, M. Werner, and J. Labarre.** 2002. Sulfur sparing in the yeast proteome in response to sulfur demand. Mol Cell **9:**713-23.
47. **Feist, A. M., C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. O. Palsson.** 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol **3:**121.
48. **Fukuchi, S., and K. Nishikawa.** 2001. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. J Mol Biol **309:**835-43.
49. **Gabel, C. V., and H. C. Berg.** 2003. The speed of the flagellar rotary motor of Escherichia coli varies linearly with protonmotive force. Proc Natl Acad Sci U S A **100:**8748-51.

50. **Galan, J. E., and H. Wolf-Watz.** 2006. Protein delivery into eukaryotic cells by type III secretion machines. Nature **444:**567-73.
51. **Gao, Q. B., Z. Z. Wang, C. Yan, and Y. H. Du.** 2005. Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett **579:**3444-8.
52. **Gardy, J. L., and F. S. Brinkman.** 2006. Methods for predicting bacterial protein subcellular localization. Nat Rev Microbiol **4:**741-51.
53. **Giovannoni, S. J., H. J. Tripp, S. Givan, M. Podar, K. L. Vergin, D. Baptista, L. Bibbs, J. Eads, T. H. Richardson, M. Noordewier, M. S. Rappe, J. M. Short, J. C. Carrington, and E. J. Mathur.** 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science **309:**1242-5.
54. **Gohl, O., A. Friedrich, M. Hoppert, and B. Averhoff.** 2006. The thin pili of Acinetobacter sp. strain BD413 mediate adhesion to biotic and abiotic surfaces. Appl Environ Microbiol **72:**1394-401.
55. **Guldener, U., M. Munsterkotter, G. Kastenmuller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S. J. Wodak, J. Garcia-Martinez, J. E. Perez-Ortin, H. Michael, A. Kaps, E. Talla, B. Dujon, B. Andre, J. L. Souciet, J. De Montigny, E. Bon, C. Gaillardin, and H. W. Mewes.** 2005. CYGD: the Comprehensive Yeast Genome Database. Nucleic Acids Res **33:**D364-8.
56. **Guo, M., F. Tian, Y. Wamboldt, and J. R. Alfano.** 2009. The majority of the type III effector inventory of Pseudomonas syringae pv. tomato DC3000 can suppress plant immunity. Mol Plant Microbe Interact **22:**1069-80.
57. **Hacker, J., and J. B. Kaper.** 2000. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol **54:**641-79.
58. **Haddix, P. L., S. Jones, P. Patel, S. Burnham, K. Knights, J. N. Powell, and A. LaForm.** 2008. Kinetic analysis of growth rate, ATP, and pigmentation suggests an energy-spilling function for the pigment prodigiosin of Serratia marcescens. J Bacteriol **190:**7453-63.
59. **Hahn, E., P. Wild, U. Hermanns, P. Sebbel, R. Glockshuber, M. Haner, N. Taschner, P. Burkhard, U. Aebi, and S. A. Muller.** 2002. Exploring the 3D molecular architecture of Escherichia coli type 1 pili. J Mol Biol **323:**845-57.
60. **Hasegawa, K., I. Yamashita, and K. Namba.** 1998. Quasi- and nonequivalence in the structure of bacterial flagellar filament. Biophys J **74:**569-75.
61. **Hauser, A. R.** 2009. The type III secretion system of Pseudomonas aeruginosa: infection by injection. Nat Rev Microbiol **7:**654-65.
62. **Hawkins, C. L., D. I. Pattison, and M. J. Davies.** 2003. Hypochlorite-induced oxidation of amino acids, peptides and proteins. Amino Acids **25:**259-74.
63. **Heizer, E. M., Jr., D. W. Raiford, M. L. Raymer, T. E. Doom, R. V. Miller, and D. E. Krane.** 2006. Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. Mol Biol Evol **23:**1670-80.
64. **Hempfling, W. P., and S. E. Mainzer.** 1975. Effects of varying the carbon source limiting growth on yield and maintenance characteristics of Escherichia coli in continuous culture. J Bacteriol **123:**1076-87.

65. **Henderson, I. R., F. Navarro-Garcia, M. Desvaux, R. C. Fernandez, and D. Ala'Aldeen.** 2004. Type V protein secretion pathway: the autotransporter story. Microbiol Mol Biol Rev **68:**692-744.
66. **Higgs, P. G., and R. E. Pudritz.** 2009. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. Astrobiology **9:**483-90.
67. **Horler, R. S., A. Butcher, N. Papangelopoulos, P. D. Ashton, and G. H. Thomas.** 2009. EchoLOCATION: an in silico analysis of the subcellular locations of Escherichia coli proteins and comparison with experimentally derived locations. Bioinformatics **25:**163-6.
68. **Hughes, K. T., K. L. Gillen, M. J. Semon, and J. E. Karlinsey.** 1993. Sensing structural intermediates in bacterial flagellar assembly by export of a negative regulator. Science **262:**1277-80.
69. **Iino, T.** 1969. Polarity of flagellar growth in salmonella. J Gen Microbiol **56:**227-39.
70. **Ishihama, Y., T. Schmidt, J. Rappsilber, M. Mann, F. U. Hartl, M. J. Kerner, and D. Frishman.** 2008. Protein abundance profiling of the Escherichia coli cytosol. BMC Genomics **9:**102.
71. **Jamir, Y., M. Guo, H. S. Oh, T. Petnicki-Ocwieja, S. Chen, X. Tang, M. B. Dickman, A. Collmer, and J. R. Alfano.** 2004. Identification of Pseudomonas syringae type III effectors that can suppress programmed cell death in plants and yeast. Plant J **37:**554-65.
72. **Jarrell, K. F., and M. J. McBride.** 2008. The surprisingly diverse ways that prokaryotes move. Nat Rev Microbiol **6:**466-76.
73. **Karp, P. D., I. M. Keseler, A. Shearer, M. Latendresse, M. Krummenacker, S. M. Paley, I. Paulsen, J. Collado-Vides, S. Gama-Castro, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penaloza-Spinola, C. Bonavides-Martinez, and J. Ingraham.** 2007. Multidimensional annotation of the Escherichia coli K-12 genome. Nucleic Acids Res **35:**7577-90.
74. **Khan, S., D. Pierce, and R. D. Vale.** 2000. Interactions of the chemotaxis signal protein CheY with bacterial flagellar motors visualized by evanescent wave microscopy. Curr Biol **10:**927-30.
75. **King, J. L., and T. H. Jukes.** 1969. Non-Darwinian evolution. Science **164:**788-98.
76. **Krogfelt, K. A., and P. Klemm.** 1988. Investigation of minor components of Escherichia coli type 1 fimbriae: protein chemical and immunological aspects. Microb Pathog **4:**231-8.
77. **Kutsukake, K.** 1994. Excretion of the anti-sigma factor through a flagellar substructure couples flagellar gene expression with flagellar assembly in Salmonella typhimurium. Mol Gen Genet **243:**605-12.
78. **Kutsukake, K., and T. Iino.** 1994. Role of the FliA-FlgM regulatory system on the transcriptional control of the flagellar regulon and flagellar formation in Salmonella typhimurium. J Bacteriol **176:**3598-605.
79. **Langille, M. G., W. W. Hsiao, and F. S. Brinkman.** Detecting genomic islands using bioinformatics approaches. Nat Rev Microbiol **8:**373-82.

80. **Lawrence, J. G., and H. Ochman.** 1997. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol **44:**383-97.
81. **Li, N., J. Lv, and D. K. Niu.** 2009. Low contents of carbon and nitrogen in highly abundant proteins: evidence of selection for the economy of atomic composition. J Mol Evol **68:**248-55.
82. **Link, C., S. Eickernjager, D. Porstendorfer, and B. Averhoff.** 1998. Identification and characterization of a novel competence gene, comC, required for DNA binding and uptake in Acinetobacter sp. strain BD413. J Bacteriol **180:**1592-5.
83. **Liu, D., and P. R. Reeves.** 1994. Escherichia coli K12 regains its O antigen. Microbiology **140 (Pt 1):**49-57.
84. **Lowe, G., Meister, M., & Berg, H.C.** 1987. Rapid rotation of flagellar bundles in swimming bacteria. Nature **325:**637 - 640.
85. **Lowe, M. A., S. C. Holt, and B. I. Eisenstein.** 1987. Immunoelectron microscopic analysis of elongation of type 1 fimbriae in Escherichia coli. J Bacteriol **169:**157-63.
86. **Lv, J., N. Li, and D. K. Niu.** 2008. Association between the availability of environmental resources and the atomic composition of organismal proteomes: evidence from Prochlorococcus strains living at different depths. Biochem Biophys Res Commun **375:**241-6.
87. **Macnab, R. M.** 2003. How bacteria assemble flagella. Annu Rev Microbiol **57:**77-100.
88. **Maezawa, K., S. Shigenobu, H. Taniguchi, T. Kubo, S. Aizawa, and M. Morioka.** 2006. Hundreds of flagellar basal bodies cover the cell surface of the endosymbiotic bacterium Buchnera aphidicola sp. strain APS. J Bacteriol **188:**6539-43.
89. **Masuda, T., N. Saito, M. Tomita, and Y. Ishihama.** 2009. Unbiased quantitation of Escherichia coli membrane proteome using phase-transfer surfactants. Mol Cell Proteomics.
90. **Matsuda, S., J. P. Vert, H. Saigo, N. Ueda, H. Toh, and T. Akutsu.** 2005. A novel representation of protein sequences for prediction of subcellular location using support vector machines. Protein Sci **14:**2804-13.
91. **Mazel, D., and P. Marliere.** 1989. Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. Nature **341:**245-8.
92. **McGuffin, L. J., K. Bryson, and D. T. Jones.** 2000. The PSIPRED protein structure prediction server. Bioinformatics **16:**404-5.
93. **Meister, M., S. R. Caplan, and H. C. Berg.** 1989. Dynamics of a tightly coupled mechanism for flagellar rotation. Bacterial motility, chemiosmotic coupling, protonmotive force. Biophys J **55:**905-14.
94. **Meister, M., G. Lowe, and H. C. Berg.** 1987. The proton flux through the bacterial flagellar motor. Cell **49:**643-50.
95. **Minamino, T., B. Gonzalez-Pedrajo, K. Yamaguchi, S. I. Aizawa, and R. M. Macnab.** 1999. FliK, the protein responsible for flagellar hook length control in Salmonella, is exported during hook assembly. Mol Microbiol **34:**295-304.

96. **Minamino, T., S. Yamaguchi, and R. M. Macnab.** 2000. Interaction between FliE and FlgB, a proximal rod component of the flagellar basal body of Salmonella. J Bacteriol **182:**3029-36.
97. **Misra, R. V., R. S. Horler, W. Reindl, Goryanin, II, and G. H. Thomas.** 2005. EchoBASE: an integrated post-genomic database for Escherichia coli. Nucleic Acids Res **33:**D329-33.
98. **Mongiovi, P., and R. Milo.** http://openwetware.org/wiki/BioNumber_Of_The_Month.
99. **Nakamura, S., N. Kami-ike, J. P. Yokota, S. Kudo, T. Minamino, and K. Namba.** 2009. Effect of intracellular pH on the torque-speed relationship of bacterial proton-driven flagellar motor. J Mol Biol **386:**332-8.
100. **Nakashima, H., and K. Nishikawa.** 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol **238:**54-61.
101. **Namba, K., and F. Vonderviszt.** 1997. Molecular architecture of bacterial flagellum. Q Rev Biophys **30:**1-65.
102. **Nambu, T., T. Minamino, R. M. Macnab, and K. Kutsukake.** 1999. Peptidoglycan-hydrolyzing activity of the FlgJ protein, essential for flagellar rod formation in Salmonella typhimurium. J Bacteriol **181:**1555-61.
103. **Nasset, E. S., F. P. Heald, D. H. Calloway, S. Margen, and P. Schneeman.** 1979. Amino acids in human blood plasma after single meals of meat, oil, sucrose and whiskey. J Nutr **109:**621-30.
104. **Nasset, E. S., and J. S. Ju.** 1969. Amino acids and glucose in human blood plasma after beef and nonprotein meals. Proc Soc Exp Biol Med **132:**1077-80.
105. **Neidhardt, F. C.** 1987. Escherichia coli and Salmonella typhimurium : cellular and molecular biology, vol. American Society for Microbiology, Washington, D.C.
106. **Neidhardt, F. C., R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low Jr., B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger.** 1996. Escherichia coli and Salmonella: Cellular and Molecular Biology, 2nd ed, vol. ASM Press, Washington, D.C.
107. **Neidhardt, F. C., J. L. Ingraham, and M. Schaechter.** 1990. Physiology of the bacterial cell : a molecular approach, vol. Sinauer Associates, Sunderland, Mass.
108. **Nenninger, A. A., L. S. Robinson, and S. J. Hultgren.** 2009. Localized and efficient curli nucleation requires the chaperone-like amyloid assembly protein CsgF. Proc Natl Acad Sci U S A **106:**900-5.
109. **Nishikawa, K., Y. Kubota, and T. Ooi.** 1983. Classification of proteins into groups based on amino acid composition and other characters. II. Grouping into four types. J Biochem **94:**997-1007.
110. **Nishiyama, M., M. Vetsch, C. Puorger, I. Jelesarov, and R. Glockshuber.** 2003. Identification and characterization of the chaperone-subunit complex-binding domain from the type 1 pilus assembly platform FimD. J Mol Biol **330:**513-25.

111. **Oh, H. S., B. H. Kvitko, J. E. Morello, and A. Collmer.** 2007. Pseudomonas syringae lytic transglycosylases coregulated with the type III secretion system contribute to the translocation of effector proteins into plant cells. J Bacteriol **189:**8277-89.
112. **Ohnishi, K., Y. Ohto, S. Aizawa, R. M. Macnab, and T. Iino.** 1994. FlgD is a scaffolding protein needed for flagellar hook assembly in Salmonella typhimurium. J Bacteriol **176:**2272-81.
113. **Perlstein, E. O., B. L. de Bivort, S. Kunes, and S. L. Schreiber.** 2007. Evolutionarily conserved optimization of amino acid biosynthesis. J Mol Evol **65:**186-96.
114. **Pintar, J., and W. T. Starmer.** 2003. The costs and benefits of killer toxin production by the yeast Pichia kluyveri. Antonie Van Leeuwenhoek **83:**89-97.
115. **Pirt, S. J.** 1965. The maintenance energy of bacteria in growing cultures. Proc R Soc Lond B Biol Sci **163:**224-31.
116. **Porstendorfer, D., O. Gohl, F. Mayer, and B. Averhoff.** 2000. ComP, a pilin-like protein essential for natural competence in Acinetobacter sp. Strain BD413: regulation, modification, and cellular localization. J Bacteriol **182:**3673-80.
117. **Proenza, A. M., C. Crespi, P. Roca, and A. Palou.** 2001. Gender related differences in the effect of aging on blood amino acid compartmentation*. J Nutr Biochem **12:**431-440.
118. **Qu, J., Y. Wang, G. Luo, Z. Wu, and C. Yang.** 2002. Validated quantitation of underivatized amino acids in human blood samples by volatile ion-pair reversed-phase liquid chromatography coupled to isotope dilution tandem mass spectrometry. Anal Chem **74:**2034-40.
119. **Raiford, D. W., E. M. Heizer, Jr., R. V. Miller, H. Akashi, M. L. Raymer, and D. E. Krane.** 2008. Do amino acid biosynthetic costs constrain protein evolution in Saccharomyces cerevisiae? J Mol Evol **67:**621-30.
120. **Reinhardt, A., and T. Hubbard.** 1998. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res **26:**2230-6.
121. **Reumann, S., K. Inoue, and K. Keegstra.** 2005. Evolution of the general protein import pathway of plastids (review). Mol Membr Biol **22:**73-86.
122. **Rey, S., M. Acab, J. L. Gardy, M. R. Laird, K. deFays, C. Lambert, and F. S. Brinkman.** 2005. PSORTdb: a protein subcellular localization database for bacteria. Nucleic Acids Res **33:**D164-8.
123. **Richmond, R. C.** 1970. Non-Darwinian evolution: a critique. Nature **225:**1025-8.
124. **Rosu, V., and K. T. Hughes.** 2006. sigma28-dependent transcription in Salmonella enterica is independent of flagellar shearing. J Bacteriol **188:**5196-203.
125. **Saier, M. H., Jr.** 1994. Protein uptake into E. coli during Bdellovibrio infection. A process of reverse secretion? FEBS Lett **337:**14-7.
126. **Sandström, J., and N. Moran.** 1999. How nutritionally imbalanced is phloem sap for aphids? Entomologia Experimentalis et Applicata **91:**203-210.
127. **Sandstrom, J., A. Telang, and N. A. Moran.** 2000. Nutritional enhancement of host plants by aphids - a comparison of three aphid species on grasses. J Insect Physiol **46:**33-40.

128. **Schaber, J., C. Rispe, J. Wernegreen, A. Buness, F. Delmotte, F. J. Silva, and A. Moya.** 2005. Gene expression levels influence amino acid usage and evolutionary rates in endosymbiotic bacteria. Gene **352:**109-17.
129. **Schaechter, M.** 2009, posting date. Talmudic Question #48: Can you think of an example of an intact protein molecule being taken up by a free-living bacterium? [Online.]
130. **Schaechter, M.** 2010, posting date. Talmudic Question #61: Endocytosis is said to be an exclusively eukaryotic property. Why did prokaryotes not develop this ability? [Online.]
131. **Schechter, L. M., M. Vencato, K. L. Jordan, S. E. Schneider, D. J. Schneider, and A. Collmer.** 2006. Multiple approaches to a complete inventory of Pseudomonas syringae pv. tomato DC3000 type III secretion system effector proteins. Mol Plant Microbe Interact **19:**1180-92.
132. **Schneider, G.** 1999. How many potentially secreted proteins are contained in a bacterial genome? Gene **237:**113-21.
133. **Schumann, W.** 2006. Dynamics of the Bacterial Chromosome. Structure and Function., vol. Wiley-VCH, Weinheim, Germany.
134. **Schuster, S. C., and S. Khan.** 1994. The bacterial flagellar motor. Annu Rev Biophys Biomol Struct **23:**509-39.
135. **Seligmann, H.** 2003. Cost-minimization of amino acid usage. J Mol Evol **56:**151-61.
136. **Shi, J. Y., S. W. Zhang, Q. Pan, and G. P. Zhou.** 2008. Using pseudo amino acid composition to predict protein subcellular location: approached with amino acid composition distribution. Amino Acids **35:**321-7.
137. **Shigenobu, S., H. Watanabe, M. Hattori, Y. Sakaki, and H. Ishikawa.** 2000. Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature **407:**81-6.
138. **Shimizu, K., and H. Toh.** 2009. Interaction between intrinsically disordered proteins frequently occurs in a human protein-protein interaction network. J Mol Biol **392:**1253-65.
139. **Singer, G. A., and D. A. Hickey.** 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol **17:**1581-8.
140. **Sokurenko, E. V., H. S. Courtney, J. Maslow, A. Siitonen, and D. L. Hasty.** 1995. Quantitative differences in adhesiveness of type 1 fimbriated Escherichia coli due to structural differences in fimH genes. J Bacteriol **177:**3680-6.
141. **Sowa, Y., and R. M. Berry.** 2008. Bacterial flagellar motor. Q Rev Biophys **41:**103-32.
142. **Stadtman, E. R., and B. S. Berlett.** 1991. Fenton chemistry. Amino acid oxidation. J Biol Chem **266:**17201-11.
143. **Steigmiller, S., P. Turina, and P. Graber.** 2008. The thermodynamic H+/ATP ratios of the H+-ATPsynthases from chloroplasts and Escherichia coli. Proc Natl Acad Sci U S A **105:**3745-50.

144. **Stouthamer, A. H.** 1979. The search for correlation between theoretical and experimental growth yields, p. 1-47. *In* J. R. Quayle (ed.), Int. Rev. Biochem. Microb. Biochem, vol. 21. University Park Press.
145. **Stouthamer, A. H.** 1973. A theoretical study on the amount of ATP required for synthesis of microbial cell material. Antonie Van Leeuwenhoek **39:**545-65.
146. **Stouthamer, A. H., and C. Bettenhaussen.** 1973. Utilization of energy for growth and maintenance in continuous and batch cultures of microorganisms. A reevaluation of the method for the determination of ATP production by measuring molar growth yields. Biochim Biophys Acta **301:**53-70.
147. **Stouthamer, A. H., and C. W. Bettenhaussen.** 1977. A continuous culture study of an ATPase-negative mutant of Escherichia coli. Arch Microbiol **113:**185-9.
148. **Suckow, J., P. Markiewicz, L. G. Kleina, J. Miller, B. Kisters-Woike, and B. Muller-Hill.** 1996. Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. J Mol Biol **261:**509-23.
149. **Swire, J.** 2007. Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. J Mol Evol **64:**558-71.
150. **Tamas, I., L. Klasson, B. Canback, A. K. Naslund, A. S. Eriksson, J. J. Wernegreen, J. P. Sandstrom, N. A. Moran, and S. G. Andersson.** 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science **296:**2376-9.
151. **Timmer, J. C., W. Zhu, C. Pop, T. Regan, S. J. Snipas, A. M. Eroshkin, S. J. Riedl, and G. S. Salvesen.** 2009. Structural and kinetic determinants of protease substrates. Nat Struct Mol Biol **16:**1101-8.
152. **Toh, E., H. D. Kurtz, Jr., and Y. V. Brun.** 2008. Characterization of the Caulobacter crescentus holdfast polysaccharide biosynthesis pathway reveals significant redundancy in the initiating glycosyltransferase and polymerase steps. J Bacteriol **190:**7219-31.
153. **Toh, H., B. L. Weiss, S. A. Perkin, A. Yamashita, K. Oshima, M. Hattori, and S. Aksoy.** 2006. Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of Sodalis glossinidius in the tsetse host. Genome Res **16:**149-56.
154. **Tokuda, H., and S. Matsuyama.** 2004. Sorting of lipoproteins to the outer membrane in E. coli. Biochim Biophys Acta **1694:**IN1-9.
155. **Toma, C., H. Kuroki, N. Nakasone, M. Ehara, and M. Iwanaga.** 2002. Minor pilin subunits are conserved in Vibrio cholerae type IV pili. FEMS Immunol Med Microbiol **33:**35-40.
156. **Tran, Q. H., and G. Unden.** 1998. Changes in the proton potential and the cellular energetics of Escherichia coli during growth by aerobic and anaerobic respiration or by fermentation. Eur J Biochem **251:**538-43.
157. **Trinchina, E. V.** 2003. Structural and functional study of the receptor binding site for FimH adhesin in uropathogenic strains of Escherichia coli. Bull Exp Biol Med **136:**380-4.

158. **Turina, P., D. Samoray, and P. Graber.** 2003. H+/ATP ratio of proton transport-coupled ATP synthesis and hydrolysis catalysed by CF0F1-liposomes. Embo J **22:**418-26.
159. **Turlin, E., G. Pascal, J. C. Rousselle, P. Lenormand, S. Ngo, A. Danchin, and S. Derzelle.** 2006. Proteome analysis of the phenotypic variation process in Photorhabdus luminescens. Proteomics **6:**2705-25.
160. **Wagner, A.** 2005. Energy constraints on the evolution of gene expression. Mol Biol Evol **22:**1365-74.
161. **Wagner, A.** 2007. Energy costs constrain the evolution of gene expression. J Exp Zoolog B Mol Dev Evol **308:**322-4.
162. **Walz, D., and S. R. Caplan.** 2000. An electrostatic mechanism closely reproducing observed behavior in the bacterial flagellar motor. Biophys J **78:**626-51.
163. **Ward, J. J., J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones.** 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol **337:**635-45.
164. **Waters, R. C., P. W. O'Toole, and K. A. Ryan.** 2007. The FliK protein and flagellar hook-length control. Protein Sci **16:**769-80.
165. **Wilkinson, T. L., D. Adams, L. B. Minto, and A. E. Douglas.** 2001. The impact of host plant on the abundance and function of symbiotic bacteria in an aphid. J Exp Biol **204:**3027-38.
166. **Xu, G., K. Takamoto, and M. R. Chance.** 2003. Radiolytic modification of basic amino acid residues in peptides: probes for examining protein-protein interactions. Anal Chem **75:**6995-7007.
167. **Yonekura, K., S. Maki, D. G. Morgan, D. J. DeRosier, F. Vonderviszt, K. Imada, and K. Namba.** 2000. The bacterial flagellar cap as the rotary promoter of flagellin self-assembly. Science **290:**2148-52.
168. **Yu, C. S., C. J. Lin, and J. K. Hwang.** 2004. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci **13:**1402-6.
169. **Zientz, E., T. Dandekar, and R. Gross.** 2004. Metabolic interdependence of obligate intracellular bacteria and their insect hosts. Microbiol Mol Biol Rev **68:**745-70.

Chapter IV

Examining the Kinetics and Conditions Favoring the *in Vitro* Polymerization of the Functional Amyloid Protein CsgA[1]

Abstract

Amyloids form when soluble proteins convert into biochemically and structurally distinct fibers. Although amyloid formation is traditionally associated with diseases like Alzheimer's disease, a number of biologically functional amyloids have recently been described. Curli are amyloid fibers produced by *Escherichia coli* that contribute to biofilm formation and other important physiological processes. We characterized the polymerization properties of the major curli subunit protein CsgA. CsgA polymerizes into an amyloid fiber in a sigmoidal kinetic fashion with a distinct lag, growth, and stationary phase. Adding sonicated, preformed CsgA fibers to the polymerization reaction significantly shortened the duration of the lag phase. This conversion of soluble CsgA into an insoluble fiber involves the transient formation of an intermediate similar to that characterized for several disease-associated amyloids. The CsgA core amyloid domain can be divided into 5 repeating units that share sequence and structural hallmarks. Peptides representing three of these repeating units are amyloidogenic *in vitro*. Although the defining aspects of CsgA polymerization appear conserved with disease-associated amyloids, these proteins evolved in diverse systems and for different purposes. Therefore, amyloidogenesis appears to be an innate protein folding pathway that can be capitalized on to fulfill normal physiological tasks.

---

[1] Portions of this chapter, mainly Figures 4.1-4.4 and the associated text, were published in JBC (123) with Xuan Wang. See Acknowledgements for more information regarding the proper attribution of this work.

Introduction

Amyloid formation is the hallmark of clinical disorders such as Alzheimer's disease, Huntington's disease, Parkinson's disease, and the transmissible spongiform encephalopathies (25). The root of these diseases is the uncontrolled conversion of seemingly unrelated soluble proteins into biochemically and structurally related fibers known as amyloids. Despite their diversity in size and amino acid content, all amyloid proteins assemble into 4-12nm wide fibers that are β-sheet rich and exhibit conserved tinctorial properties. Soluble pre-amyloid species also share common pore-like epitopes, and these globular species may induce cytotoxicity (51, 57, 91).

Numerous studies have revealed that amyloidogenic proteins are mostly unstructured or contain mixtures of β-sheets and α-helices in their native state, but when polymerized into fibers they invariably adopt a characteristic cross β-sheet structure (7, 8, 52). This cross β-sheet structure is common to all amyloids and is characterized by β-strands that orient perpendicular to the fiber axis. *In vitro,* disease-associated amyloids polymerize into fibers with nucleation-dependent kinetics with characteristic lag, growth and stationary phase. The lag phase is proposed to contain folding intermediates that are key to the toxicity associated with certain amyloidogenic proteins (61, 63). During the lag phase, amyloidogenic proteins adopt a transiently folded species that disrupts membrane integrity (16, 57, 61). Loss of membrane integrity is proposed to underlie the cell death and disease associated with many amyloids (16, 57). A conformational-specific antibody has been generated that recognizes a transient intermediate formed during amyloidogenesis of several disease-associated proteins (57).

A new class of amyloids has recently been found that play important physiological roles for the cell. These so-called 'functional' amyloids are found in bacteria (12, 19, 33), fungi (29, 113) and mammals (37). The first example of a functional amyloid in bacteria was curli (19). Curli compose part of the complex extracellular matrix that is required for biofilm formation (2, 135, 136), host cell adhesion (53), and invasion (44, 45), and they are proposed to be important stimulants of the host inflammatory response (10, 11,

114, 115). An intriguing question is whether these functional amyloid proteins polymerize in a manner similar to disease-associated amyloids.

Curli formation is the result of an elegant biosynthetic pathway directed by the Csg proteins in *E. coli.* The major curli subunit, CsgA, can be secreted to the cell surface as a soluble, unstructured protein (5, 19). CsgA is efficiently nucleated into an insoluble amyloid fiber in the presence of the outer membrane-associated protein, CsgB (48, 49). After nucleation, the fibers are predicted to grow by subsequent CsgA addition to the amyloid fiber's tip (5).

Both CsgA and CsgB display a remarkable five-fold internal symmetry characterized by conserved polar residues. These five 'repeating units' consist of 19-24 amino acids and align along serine, glutamine and asparagine residues (5, 28). Each repeating unit is predicted to form a strand-loop-strand motif that closely resembles the cross β-spine structure described for many disease-associated amyloids (28, 79, 80).

Here I characterize the folding of purified CsgA and show that its polymerization is similar to that of disease-associated amyloids. CsgA polymerization involves a transient structurally conserved intermediate that implies a common polymerization pathway between functional and disease-associated amyloids. I found that the conserved folding intermediate for CsgA is a monomer or low molecular weight multimer. We demonstrate that at least three of five repeating units of CsgA are amyloidogenic. These results suggest that covalent linkage of multiple amyloidogenic units facilitates efficient fiber formation.

Results

CsgA-His polymerization kinetics

To determine the polymerization kinetics of CsgA-His, an *in vitro* polymerization assay was developed. The transition of freshly purified, soluble CsgA-His to amyloid fibers was monitored using thioflavin T (ThT), an amyloid-specific dye commonly used to assay amyloid formation (64, 65). The ThT fluorescence of CsgA-His samples followed a sigmoidal curve with distinguishable lag, growth and stationary phases (Fig. 4.1A). While

the polymerization rate was concentration dependent (Fig. 4.1A and 4.24B), the lag phase and half time (t½) were not (Fig. 4.24A,C). However below 4 μM, the lag phase was concentration dependent, consistent with nucleation dependent kinetics; however, above 4 μM, the lag phase was concentration independent (Fig. 4.24A). Concentration-independent lag phases have been reported for other amyloidogenic proteins including IAPP (71, 84, 93, 94, 103). ThT fluorescence signal did not change appreciably after 8 hours, remaining at approximately the same level for over 30 days (data not shown).

The structural changes that occurred during CsgA-His amyloidogenesis were measured by circular dichroism spectroscopy and transmission electron microscopy (TEM). Circular dichroism spectrum indicated that immediately after purification CsgA-His was largely unstructured (Fig 4.1B). However, CsgA-His adopted a β-sheet-rich structure after 2 days of incubation at room temperature (Fig 4.1B). Immediately after purification there was no apparent fiber formation or aggregation by TEM (Fig 4.1C). Two hours after purification regular, unbranched fibers were readily observed (Fig. 4.1D). Dense fiber aggregates were also observed 7 days post purification (Fig. 4.1E). Therefore, the appearance of fibers observed by TEM was approximately coincident with an increase in ThT fluorescence.

The A11 antibody recognizes a transient CsgA-His folding species.

The polymerization of several amyloids found in eukaryotes involves the formation of an intermediate folding species proposed to cause amyloid-associated toxicity to host cells (51, 57). The A11 antibody recognizes an Aβ transient intermediate (57). Remarkably, this antibody also recognizes a transient intermediate formed by the amyloids IAPP, poly Q, PrP, and Sup35p, among others (57, 104). The A11 antibody recognizes only a transient intermediate species, not soluble monomers or mature amyloid fibers derived from these proteins.

We utilized the A11 antibody to determine if CsgA-His shared a common polymerization intermediate with other amyloids. I found that immediately after purification CsgA-His was recognized by the A11 antibody (Fig. 4.2A). As fiber formation

proceeded, evidenced by increased ThT fluorescence and the appearance of fiber aggregates by TEM, the A11 antibody lost no longer bound CsgA-His (Fig.2A). A polyclonal antibody generated against CsgA recognized purified CsgA-His independently of its polymerization status (Fig. 4.2A).

The observation that the A11 antibody recognized CsgA-His suggested that intermediates during CsgA-His polymerization might be structurally similar to those formed by disease-associated amyloid proteins. It also suggested that, immediately after purification, CsgA-His had already begun its transition to an amyloid fiber. To prevent CsgA-His from folding during purification, the CsgA-containing fractions were amended with 8 M Guanidine hydrochloride (GdHCl). Under these strongly denaturing conditions, the A11 antibody did not recognize CsgA-His; however, 8M GdHCl denatured CsgA-His was readily recognized by the CsgA antibody (Fig. 4.2B). Immediately after GdHCl removal with a desalting column, CsgA-His was recognized by the A11 antibody (Fig. 4.2B and Fig. 4.19).

To determine the minimum size of the CsgA-His transient intermediate, freshly purified protein was passed through Amicon Ultra centrifugal membranes with different molecular weight cutoffs. The retentate and filtrate were probed with the A11 antibody (Fig. 4.2C). The A11 antibody recognized a species in the filtrate of the 30 kD membrane, suggesting that the smallest reactive species of CsgA-His is 30 kD or less. Because CsgA-His has a predicted molecular mass of 13.9 kD, the species recognized by the A11 antibody is likely either a monomer or dimer.

CsgA-His fibers can catalyze self-polymerization

The approximately sigmoidal ThT fluorescence curve suggests that CsgA-His polymerizes by a nucleation-dependent mechanism. Therefore, the growing fiber would be expected to direct the polymerization of new CsgA molecules. We tested the hypothesis that preformed CsgA-His fibers could catalyze CsgA-His polymerization. Addition of 2.5% (w/w) sonicated CsgA-His fibers to freshly purified, soluble CsgA-His resulted in a significantly reduced lag phase (Fig. 4.3A). Coincident with the dramatically

shorter lag phase in seeded reactions, CsgA-His was recognized by the A11 antibody for a significantly shorter period of time compared to unseeded reactions (Fig. 4.3B).

CsgA-His is composed of several amyloid-forming units

The observation that CsgA-His was recognized by the A11 antibody immediately after passing through a 30 kD cutoff filter (Fig. 4.2C) was unexpected since the A11 antibody is thought to recognize an oligomeric form of amyloidogenic proteins (57, 63, 104). The number of molecules present in the oligomeric state recognized by A11 varies among amyloidogenic proteins, but Aβ oligomers have been estimated to be larger than tetramers (57, 63). However, CsgA-His is likely recognized by A11 as a monomer or at most a dimer as estimated by cutoff filtration. We postulate that thee multiple amyloidogenic domains of a single CsgA molecule collectively contribute to it folding into a state that interacts with the A11 antibody. The primary sequence of CsgA can be divided into three parts: the Sec-dependent signal sequence, the N-terminal 22 amino acids of the mature protein, and a repeat domain that contains five 19-22 amino acid repeating units (Fig. 4.4A). The five repeating units form a protease resistant structure that is proposed to be the amyloid core of CsgA (5, 28). Each repeat has four conserved polar amino acids: serine, glutamine, asparagine and glutamine (Fig. 4.4A). The regular arrangement of glutamine and asparagine residues also occurs in CsgA homologs from different *Enterobacteriaceae* (Wang and Chapman unpublished observation).

We hypothesized that these repeating units might represent single amyloid forming units. Accordingly, peptides corresponding to each repeating unit were chemically synthesized and tested for their ability to form amyloid fibers. Two independently derived preparations of each peptide were assayed. Peptides corresponding to repeating unit 1, 3 or 5 (R1, R3 or R5) produced a ThT-positive signal and self-assembled into fibers as evidenced by TEM when incubated at 0.2 mg/ml (Fig. 4.4B-E). Neither R2 nor R4 showed evidence of amyloidogenesis when resuspended at a concentration of 0.2 mg/ml, although fibers were observed by TEM when R2 or R4 were incubated at 2 mg/ml (Fig. 4.4B and data not shown). The morphology of R1 fibers was similar to those

formed by purified CsgA-His, being generally longer than 1000nm (compare Fig. 4.4C to Fig. 4.1D and 4.1E). R3 fibers were consistently shorter (ranging from 200nm-1000nm) than those formed by CsgA-His (compare Fig. 4.4D to Fig. 4.1D and E). R5 fibers appeared more rigid and aggregated than CsgA-His fibers (Fig. 4.4E). The morphologies of the fibers did not appreciably change over the course of a ten-day incubation. This analysis suggests that CsgA contains at least three highly amyloidogenic domains, R1, R3 and R5, that likely drive fiber formation *in vivo.*

CsgA-His is not non-enzymatically glycated

In both purificaiton protocols, CsgA-His that eluted from the NiNTA column was often quite brown in color, indicative of glycated proteins (Fig. 4.5B). Non-enzymatic glycation contributes to fibrilization and toxicity of several disease-associated amyloids (100, 118). The Gel filtration column used in the denaturing protocol efficiently separated from CsgA-His the source of this brown color, which was also detectable by fluorescence (Fig. 4.5E-G). The brown color is likely a small peptide present in LB that also binds to the NiNTA column with some affinity (Fig. 4.5H). Lag phases were more consistent upon gel filtration and were typically shorter than less pure preparations (data not shown) indicating the possibility that some of these small peptides may have an inhibitory role.

Effects of pH, various salts, and ionic strength CsgA-His polymerization

The effects of pH, various salts, and ionic strength on amyloid formation has proved useful in studying the fiber formation pathway of amyloids because certain conditions stabilize or destabilize transient folding intermediates that are often correlated with cellular toxicity (23, 56, 75, 76, 78, 92, 96, 107, 110, 116, 125, 132-134). Coupling the knowledge of these effects with a protein sequence can also reveal which regions or types of interactions are important for efficient amyloid fiber formation. Consequently, we examined the effects of changes to pH, ionic strength, and salts on CsgA-His polymerization.

Effects of pH on CsgA-His polymerization.

In a multiple buffer system consisting of 20 mM each of potassium borate, citrate, and phosphate (KBCP), CsgA-His had increased polymerization rates and reduced lag phases at around pH 3 and pH 8 with the latter having higher overall ThT fluorescence and relative rate of polymerization (See Fig. 4.6). Similar results were also obtained in other buffer systems including KAc (Potassium Acetate), 3M (MES, MOPS, and MOBS), and TSC (Tris, Sodium Succinate, and CAPS) (Fig. 4.7). Because ThT fluorescence of preformed CsgA-His fibers is affected by pH (58) (Fig. 4.8A), it is necessary to compare relative rates instead of slopes between pH values. Around pH 9, ThT fluorescence is maximal; whereas at pH 11 fibers start to breakdown (Fig. 4.8B,C).

Previously, different affects of pH on CsgA-observed. In earlier experiments CsgA-His was buffer exchanged using PD-10 columns, and CsgA-His polymerized better at lower pH values (Jonathon Jones, unpublished observations). However, those experiments used single component acetate or phosphate buffers. Different buffer ions can affect aggregation and amyloid formation in a pH independent fashion (56, 125, 134). For example, in potassium acetate at pH 4, CsgA-His has a much greater relative rate of fiber elongation than in any other buffer or pH assayed (Fig. 4.7A). The use of multiple buffer systems overcomes this to some degree, especially the 3M buffer which contains three compounds whose structure only differs by the number of carbons linking a morpholine to sulfonic acid.

The maximum polymerization rate of CsgA-His near pH 8 is unexpected. Most amyloids polymerize more readily near their isoelectric point (pI), where charges are balanced, thus allowing for efficient hydrophobic collapse and increased side chain interactions needed for amyloid nucleus formation (102). The theoretical pI of mature CsgA-His is 5.73 (88, 89); the pI for wild type (WT) CsgA is either 4.28 (88) or 4.51 (89). When graphing the pI versus the pH_{max} of other amyloids, CsgA-His appears to be quite the outlier (Fig. 4.9). To explore why this might be, I examined the theoretical titration curves of WT CsgA and CsgA-His (Fig. 4.10). The two curves are quite similar above pH 7, showing a shoulder that extends to pH 9, encompassing the pH range over which CsgA-

His polymerizes the fastest. Below pH 7, CsgA-His becomes more positive, relative to WT CsgA, due to the 6X-His-tag on its C-terminus. CsgA-His is expected to be neutrally charged near pH 6, yet this is where CsgA-His polymerizes the slowest. Usually, lowering the net charge of amyloidogenic proteins increases their rate of aggregation (18, 24, 112, 124).

Possible reasons for the discrepancy in the pH_{max} of CsgA-His include the proximity of the 6X-His-tag to R5, the presence of a His residue in the last beta strand of R5 (Fig. 4.29B and 4.30), or the differences in the pI of the individual repeating units of CsgA (Fig. 4.30). R5 is an important region of CsgA-His and greatly increases the aggregation of CsgA-His (122). Like R1, it is relatively uncharged and may be one reason why both R1 and R5 polymerize more efficiently than the other repeating units of CsgA (124). Perhaps the increased charge from the protonation of the 6X-His-tag or His149 (pKa ~ 6.1) at lower pH inhibits important interactions in R5 needed for efficient curli fiber formation. Histidines appear to drive the slower polymerization of CsgA-His down to around pH 4. Near pH 3, CsgA-His polymerizes more quickly, possibly because Asp (pKa 3.9) and Glu (pKa 4.1) residues become uncharged (Fig. 4.6). Wang et al 2010 (124) showed that replacing several Asp's with neutral residues greatly increased CsgA polymerization. Alternatively, R1 and R5, the two most important repeating units of CsgA (122) (Fig. 4.4), have pI's very near the two pH maximums of CsgA-His. Perhaps the increased polymerization of CsgA-His around pH 3 to 4 and around pH 8 reflects the collapse of R1 or R5, allowing them to form an amyloid nucleus that templates the other less efficient repeating units.

At both pH extremes CsgA-his polymerization is slowed down (Fig. 4.7). Below pH 3 CsgA-His is more likely to be acid hydrolyzed; the pKa of α-COOH are between 1.8 and 2.6. At pH >10, Tyr and Lys begin to be deprotonated leading to increased positive charge which should drive CsgA to a neutral charge. However, at high pH CsgA seems to polymerize more slowly and preformed fibers seem to depolymerize. Assuming high pH does not lead to increased degradation of CsgA-His (pKa of α-NH3 vary from 8.8 to 10.6), one possible way to purify unfolded CsgA-His or CsgA would be to elute into a

buffer with a pH greater than 10 and then quickly adjusting to neutral pH to start polymerization. Further study of the pH_{max} of WT CsgA and mutants in charged residues may elucidate what regions and residues are important for CsgA polymerization.

The pI's of CsgA and CsgB appear complementary.

Intriguingly, the pI of mature CsgB is quite high (either 9.3 (89) or 9.7 (88)) due to the highly basic R1, R4, and R5 repeating units (Fig. 4.30). A comparison of the theoretical titration curves of both CsgA and CsgB (Fig. 4.10) shows both would be expected to have opposite charges near neutral pH. One possible role for this difference in pI's is charge attraction. Charged amyloidogenic proteins have been shown to bind to opposite charged macromolecules resulting in increased aggregation (17). At physiological pH, the negative charge of CsgA is predicted to attract to the positive charge of CsgB, allowing heteronucleation to occur. Both R4 and R5 of CsgB are required for its proper function (Neal Hammer, Doctoral thesis and unpublished work). The importance of the charge of R4 and R5 of CsgB to attracting CsgA would be an interesting area of further study, perhaps by testing CsgA-His and WT CsgA seeding with CsgB at different pH values.

Effects of various salts and ionic strength on CsgA-His polymerization.

I also tested the effects of various salts on the polymerization of CsgA-His and found salts that are bound by His residues appear to inhibit CsgA-His polymerization. $NiSO_4$ and $ZnSO_4$ were both able to completely inhibit the polymerization of CsgA-His at 1mM (Fig. 4.11A,C). $AlCl_3$, $CaCl_2$, $MgCl_2$, and $MgSO_4$ had no effect on csgA polymerization (Fig 4.11B). Sodium salts also do not have an effect, as phosphate buffered saline and sodium phosphate buffers have both been used for CsgA-His polymerization. Addition of EDTA rescued the inhibition from $NiSO_4$ but not $ZnSO_4$ (Fig. 4.11A,C); however, EDTA itself had no apparent effect on CsgA-His polymerization (Fig. 4.11D). Many amyloids respond to salts on the Hofmeister series, with more chaotropic ions either inhibiting or promoting amyloid formation depending on the amyloidogenic protein (78, 92, 96, 107,

132); however, there are exceptions to this rule, especially at different pH values (15, 78, 92). Here, CsgA-His responded only to Zn^{+2} and Ni^{+2}, both of which are known to bind to imidazole, the functional group of histidine. A similar result was seen with human IAPP, where Zn and Cu but not Ca, Mg, Na, or NH_4 affected polymerization due to their binding to His18 (15). Thus, the effect of Zn and Ni ions could be due to their binding the 6X-His-tag or one of the three histidines in mature CsgA, one of which is in R5. Further study with WT CsgA and His mutants will hopefully elucidate which region of CsgA Zn and Ni bind.

To test the effects of ionic strength, I added different concentration of KCl to CsgA-His. Very low ionic strength resulted in increased lag phases and slower elongation rates, suggesting ionic interactions are crucial but not necessary for CsgA-His polymerization (Fig. 4.12A). Increasing ionic strength up to 100 mM resulted in decreased lag phases and increased polymerization rates. However, concentrations of KCl above 200 mM slowed polymerization rates, suggesting that fibril formation is inhibited by increasing ionic interactions (Fig. 4.12C,D). Similar results were seen for β2-microglobulin above 200 mM NaCl (92). Strangely, addition of 200 and 400 mM KCl resulted in much longer lag phases for CsgA-His than those observed with KCl concentrations greater than 600 mM. Finally, I verified that high concentrations of did not interfere with binding or quench ThT fluorescence of preformed CsgA-his fibers (Fig 4.12B-D).

Effects of temperature on CsgA-His polymerization and A11 antibody binding

I tested the effects of temperature on CsgA-His polymerization, solubility, and A11 antibody binding. Higher temperatures resulted in increased polymerization rates and decreased lag phases (Fig. 4.13A and 4.14A). At lower temperatures, CsgA-His polymerized more slowly; at 4^{o}C the lag phase was greater than 12 hours. Temperatures greater than 43^{o}C resulted in a slow breakdown of CsgA-His over time (Fig. 4.14B, 4.17, and 4.31B). At each temperature tested, A11 binding occurred during the lag phase and into the early stationary phase of CsgA-His, suggesting A11 binds to a monomer or a

multimer that is stable at multiple temperatures, including very high ones (Fig. 4.16 and 4.31C). In many cases A11 antibody binding was seen up to the point where CsgA-His was completely degraded from its apparent monomeric form on native gels (Compare Fig. 4.14A and 4.14B, See Fig. 4.17C and 4.34A,C). Previous papers have suggested that formation of fibers at different temperatures results in different morphologies which can display seeding specificity (111). Preliminary data suggested the fibers formed at 16^{o}C, 24^{o}C, 37^{o}C, and 53^{o}C all displayed the ability to seed unpolymerized CsgA at 25^{o}C (data not shown, DS52).

CsgA-His and A11 antibody binding

Amyloid formation typically involves a conversion from monomers into a multimeric nucleus; subsequently, monomer addition results in growing amyloid fibers (See Fig. 4.25). To determine if freshly purified CsgA-His was a mixture of multimeric species, I tested the mobility of CsgA-His in native gels. At multiple acrylamide percentages, CsgA-His ran as a single species on native gels (Fig 4.14C, 4.15, 4.17B, and 4.31C). Preincubation of CsgA-His at multiple temperatures up to 95^{o}C before mixing with Native Sample Buffer (30% glycerol 0.05% Bromophenol Blue) did not affect the gel mobility of CsgA-His. Furthermore, at concentrations of GdHCl that inhibited A11 binding (2M in Fig. 4.31C), CsgA-His ran at the same location on a native gel as at multiple temperatures in the absence of GdHCl (Fig. 4.31A). Intriguingly, the degradation of the single species of CsgA-His into smaller fractions was concomitant with the loss of A11 binding in dot blot assays (Fig. 4.14, 4.15; 4.17B,C; and 4.313C). Close examination of the ThT fluorescence time course assays with native gel solubility (Fig. 4.14) shows the majority of CsgA-His appears to be in a single soluble form (Fig. 4.14). As CsgA-polymerizes into curli fibers, this single species decreases in intensity (See 43C in Fig. 4.14A and 4.14B at 3 hours). Perhaps another more sensitive technique, such as SDS agarose gel electrophoresis and immunoblotting (4, 119), would show additional species. I transferred these native gels onto nitrocellulose membranes and probed them with the A11 oligomer antibody. Intriguingly, the A11 bound to a single band of CsgA-His

(Fig. 4.14); however, no A11 binding was seen in control lanes containing either bovine serum albumin (BSA) or the proteins found in Benchmark Prestained Protein Ladder (data not shown). These data suggest the A11 oligomer antibody may be binding to a single monomeric species of CsgA-His.

Assays with the A11 antibody did generate some unexpected results. I filtered CsgA-His in 8M GdHCl through a cellulose acetate membrane and then transferred it to nitrocellulose. When probing this membrane with the A11 antibody, I saw A11 binding was restored (Fig. 4.20), even if the sample was boiled in 8M GdHCl; however, similar results were not seen for other types of filters including nylon (data not shown) and Anotop γ-alumina based filters (Fig. 4.20). Like curli, cellulose is an important component of the extracellular matrix in biofilms that are produced by many *Enterobacteriaceae*; cellulose is regulated by the master curli regulator CsgD via AdrA (13, 46, 54, 99, 108, 127, 135, 136). Therefore, cellulose may accelerate the polymerization of curli. However, addition of cellulose did not affect CsgA-His fiber formation in a ThT fluorescence time course assay (data not shown, Xuan Wang unpublished results). Another unexpected result was seen when sonication of preformed fibers of CsgA-His or the R1 repeating unit resulted in A11 oligomer binding (Fig. 4.28A). This restoration of A11 binding was observed for these samples for several days (Fig. 4.28B). Whether A11 is recognizing the seed species that accelerates CsgA-His seeding reactions will require further experimentation.

The A11 oligomer antibody does not inhibit CsgA-His polymerization.

The A11 oligomer antibody recognizes an on pathway intermediate in the yeast prions Sup35 (104) and RnqPD (119). Consequently, I determined the ability of A11 antibody to inhibit CsgA-His polymerization and the concentration at which CsgA-His was no longer bound by the A11 antibody. Molar ratios up to 1:22.5 of the A11 antibody to CsgA-His had no affect the lag phase, polymerization rate, or fluorescent yield of CsgA-His. In Sup35 NM, partial inhibition was shown at molar ratios as low as 1:200, with nearly complete inhibition occurring at a 1:100 molar ratio (104). For RnqPD, a

molar ratio of 1:62 was sufficient to inhibit unseeded reactions (119). Furthermore, I tested the ability of the A11 antibody to inhibit CsgA-His below 4 µM, where the lag phase begins to increase in a concentration dependant manner (Fig. 4.1A and 4.24). A 1:45 molar ratio of A11 antibody to CsgA-his failed to inhibit polymerization of 3.6 µM of CsgA-His (Fig. 4.18C). In multiple experiments, the A11 antibody was failed to bind to CsgA-His concentrations below 4 µM (Fig. 4.21). It is possible that 4 µM CsgA-His is the detection limit of the A11 antibody, or this limit may elucidate some aspect of the mechanism of CsgA-His polymerization.

Effects of GdHCl on A11 binding and polymerization of CsgA-His

I tested the effects of GdHCl to determine the concentration at which CsgA-His polymerization was inhibited and to see if that concentration was concomitant with the loss of A11 binding. GdHCl appears to block CsgA-polymerization and ThT binding above 3.5 M (Fig. 4.23A,B); however, addition of GdHCl does affect ThT fluorescence of preformed fibers (Fig. 4.23C), complicating interpretation of this result. Concentrations of 0.8 and 1M GdHCl accelerated the half time (t½) of CsgA-His polymerization while concentrations greater than 2M had increased lag phases and slower rates of polymerization (Fig. 4.23A,B). To test which concentration of GdHCl blocked A11 binding, I diluted NiNTA elutes (8M GdHCl and higher [CsgA-His]) in H_20 or added GdHCl to gel filtered elutes (desalted into 50mM KPi pH 7.2). Diluted NiNTA elutes bound A11 in the presence of 1.6M GdHCl 30 minutes after mixing; roughly an hour later, A11 binding to CsgA-His was present in 3.2M GdHCl (Fig. 4.22A). When GdHCl was added back to gel filtered elutes, A11 antibody binding to CsgA-His was present in 2M (Fig. 4.22B) and 3M (Fig. 4.22C) GdHCl samples; however, in Fig. 4.31C, 2M GdHCl was sufficient to prevent A11 binding. There was not an apparent link between the GdHCl concentration needed to inhibit CsgA-His polymerization and A11 antibody binding. Another complicating factor in this analysis is that high GdHCl concentrations appears to inhibit CsgA-His binding to nitrocellulose membranes (See anti-CsgA blots in Fig. 4.22A,C).

Tryptophan fluorescence of CsgA-His

CsgA contains a single Trp residue in the loop between R3 and R4 (Fig. 4.4A and 4.29D); therefore, I exploited the intrinsic fluorescence and iodine quenching of this residue to analyze curli fiber formation. As little as 2.5 μM of unpolymerized CsgA-His gave a high amount of Trp 84 fluorescence with a maximum indicative of a Trp residue exposed to a polar solvent (Fig. 4.26A). When KI was added, I saw strong quenching of this residue suggesting solvent accessibility. Addition of 4 M GdHCl did not appreciably change the Trp84 fluorescence of unpolymerized CsgA-His; however, it did inhibit iodine quenching (Fig. 4.26B,C). In contrast to soluble CsgA-His, preformed CsgA-His fibers (2.5 μM) had much lower Trp 84 fluorescence and was not as accessible to iodine quenching (Fig. 4.26D-F). These results are consistent with partial burial of the Trp 84 residue in mature CsgA-His fibers.

Because the M2 plate reader can analyze multiple wavelengths nearly simultaneously, I also directly compared tryptophan fluorescence and Thioflavin T fluorescence. Strangely, tryptophan fluorescence increased in the plate reader and seemed to mimic the increasing ThT fluorescence (Fig. 4.27). This result opposes those found in a 90^{o} fluorometer above. Perhaps this result is due to difference in geometry between the two machines. The plate reader may be reading increased light scattering from more turbid CsgA-His as its geometry is more sensitive to stray light. Alternatively, CsgA-His fibers may change over time with the Trp 84 later becoming buried as fibers mature. Consistent with this possibility, CD spectra analysis showed increasing β-sheet propensity over seven days (Fig. 4.1B). Either way, tryptophan fluorescence or light scattering may facilitate studying CsgA-His polymerization at low concentrations (below 4μM) where polymerization is concentration dependent yet ThT fluorescence lacks the sensitivity to adequately assay curli fiber formation (Fig. 4.1A and 4.24).

Discussion

Amyloid formation is traditionally associated with uncontrolled protein misfolding and aggregation that results in many systemic and neurodegenerative disorders (25, 105). However, there are a growing number of functional amyloids that suggest amyloidogenesis is also a general tenet of normal cellular physiology (1, 3, 12, 19, 26, 27, 29-31, 34, 37, 38, 50, 73, 74, 81, 83, 95, 106, 120, 128-130). In fact, amyloid formation may be a common property of most proteins (36, 47).

The work presented here, as well as that published previously, demonstrates that both disease-associated and functional amyloids share a common amyloid formation pathway (43). CsgA polymerizes with nucleation-dependent kinetics, and fiber formation is ameliorated by the addition of pre-formed CsgA fibers. We also found that CsgA polymerization involves the formation of a transient species similar to that produced by other amyloidogenic proteins such as Aβ, synuclein, IAPP, insulin, lysozyme and polyglutamine (57).

The transient species that the A11 antibody recognizes during CsgA polymerization is a monomer or low-molecular weight multimer (Fig. 4.2C). The A11-recognized species of Aβ and Sup35p are thought to be large molecular weight oligomers (57, 104). Unlike Aβ and Sup35p (57, 104), CsgA was immediately recognized by the A11 antibody upon removal of strong denaturants like GdHCl or after its passage through a 30 kD Amicon filter. I also found that freshly purified CsgA heated to 99°C for 5 minutes (Fig. 4.16) or longer (Fig. 4.14) was also recognized by the A11. At least two hypotheses can be proposed to explain CsgA's ability to be recognized by A11 immediately after denaturation or passage through a 30 kD cutoff filter. First, CsgA may adopt an oligomeric conformation so quickly that our ability to measure this transition is lost in the time that it takes to immobilize CsgA on the blotting paper. Another possibility is that the CsgA species recognized by A11 is not an oligomer, but a monomer that contains multiple amyloidogenic units. In support of this hypothesis, CsgA does indeed contain at least three amyloidogenic domains (Fig. 4.4). Nonetheless, these two

hypotheses are not mutually exclusive, and there may be other plausible interpretations.

In any case, CsgA contains multiple amyloidogenic domains that may contribute to its ability to efficiently transition from a soluble protein to an amyloid fiber. Many studies have led to the proposal that amyloid fibers themselves are not toxic to cells; instead, toxicity is proposed to be caused by transient folding intermediates (32, 42, 51). Therefore, one mechanism that might be used by functional amyloids to prevent toxicity is to minimize the duration of toxic folding intermediates. This is apparently how Pmel17, an extremely rapidly forming functional amyloid found in mammalian cells, is able to assemble within the cell without eliciting a toxicity cascade (37).

CsgA has a striking primary sequence arrangement (Fig. 4.4A). The five imperfect repeats of CsgA share greater than 30% sequence identity. Each repeating unit is potentially a single amyloid domain, and R1, R3 and R5 are highly amyloidogenic *in vitro* (Fig. 4.4B-E). The covalent linkage of multiple amyloid domains may facilitate amyloid fiber formation by increasing the number of amyloidogenic building blocks, and may also result in rapid formation of the intermediate recognized by the A11 antibody. Other amyloidogenic proteins contain repeat sequences that have been postulated to facilitate fiber formation (97, 131). For instance, the N-terminal prion-determining domain of Sup35p has five imperfect oligopeptide repeats and certain deletions of the repeats are defective in propagation of Sup35p fibrils. Moreover, *in vitro*, repeat-expansion peptides (with 2 extra repeats) are more amyloidogenic than wild-type peptides (69). The most amyloidogenic domains of CsgA were reported to be contained in the hexapeptides GHGGGN and QFGGGN, which are present in R2 and R4 respectively (22). However, our analysis suggests that R1, R3 and R5 contain the more highly amyloidogenic sequences. A more thorough examination of was performed in our lab, confirming the importance of R1, R3, and R5 (121, 122, 124).

The amyloidogenic peptides R1 and R5 contain sequences that contribute significantly to CsgA's ability to bind human proteins such as fibronectin, plasminogen, tissue plasminogen activator, and β2-microglobulin (82). This correlation suggests

amyloidogenicity of CsgA may be directly linked to these biological activities. In fact, Gebbink *et al* suggested that curli contribute to colonization in animal hosts by activating host proteases that are involved in haemostasis (41).

Curli can also enhance amyloid protein A amyloidosis in mice (72). It is proposed that cross-seeding may play a role in the development of amyloid diseases (59, 72). The *in vitro* system that we have established here provides an ideal vehicle to test the specificity of curli seeding with other amyloids. Understanding how functional amyloid proteins interact with other host proteins may lead to new ideas about cellular physiology and the processes that promote the toxicity associated with many amyloids.

Most amyloids are known to self-propagate in a process called seeding. In prion diseases such as Bovine Spongiform Encephalopathy, seeding underlies protein infectivity (90). Amyloid self-propagation is also critical to disease development in the non-transmissible amyloid diseases (59, 60). Our demonstration of CsgA seeding suggests that functional amyloids also utilize a controlled self-propagation process to fulfill their biological function. *In vivo,* CsgA polymerization is nucleated by the outer membrane associated protein CsgB, which shares nearly 49% sequence similarity with CsgA (48). One proposed model of nucleation is that CsgB provides an amyloid-like template that initiates CsgA polymerization (5, 48). The growing fiber tip could then act as a template to direct subsequent CsgA polymerization.

Proteins that are not predicted to form stable globular folds may be prone to aggregation and amyloid formation. Indeed, most functional amyloid proteins have natively disordered segments (79, 117). Consistent with this, some proteins have been shown to form amyloid fibers only after the native, globular fold has been compromised by chemical denaturants or by mutations (14, 36, 47). Circular dichroism studies presented here suggest that CsgA is unstructured after secretion (Fig. 4.1B). Mature CsgA is also predicted to be natively unfolded (Fig. 4.29) by the algorithms of Uversky, Galzitskaya, and others (39, 117). The natively unfolded segments of CsgA may facilitate amyloidogenesis indirectly by preventing formation of stable globular structures that would be less likely to aggregate and precipitate into a fiber. Alternatively, the unfolded

regions of CsgA may actively direct amyloidogenesis by presenting specific aggregative surfaces to neighboring molecules. In this context, 'natively unfolded' would be a transition state during the formation of a stably-folded amyloid fiber. Importantly, in the case of functional amyloids, the amyloid fiber would not be the product of protein misfolding, but that of protein folding. Certainly, the growing number of functional amyloids suggests that the amyloid fiber is an evolutionarily conserved structure. The selective processes that have been employed by functional amyloids to limit cellular toxicity provide a unique context from which to investigate disease-associated amyloidogenesis.

Proposed mechanism of CsgA-His polymerization and A11 binding.

Based on published work and the evidence presented here, I propose the following mechanisms for CsgA-His polymerization. While fiber elongation is concentration independent (Fig. 4.24B), the lag phase and t½ were not (Fig. 4.24A,C), suggesting that CsgA-His fiber formation involves two lag phases. One lag phase is only present when CsgA-His is below approximately 4μM, is concentration dependent, and appears to follow nucleation dependent kinetics. This lag phase likely represents the time required to form an on-pathway multimeric nucleus (n_a) (Fig. 4.25). The second lag phase is concentration independent and is the same for all concentrations above 4μM (Fig. 4.24). This step likely represents a maturation of the on-pathway nucleus (n_a) into another nucleus (n_b) of approximately the same stoichiometry (Fig. 4.25). While some amyloids follow nucleation dependent kinetics (9), many do not and instead display a concentration independent step (62, 71, 84, 94, 98, 103). Various explanations have been put forward for this concentration independent step, including the formation of micelles (62, 70, 71, 93, 94, 98, 103, 109), the presence of off-pathway detergent soluble oligomers (101), phase partitioning (84), and supercritical concentration (86). Following the formation of the n_b nucleus, CsgA-His monomers polymerize on the template provided by the nucleus to form long fibers which associate into curli fibers.

While the intermediate recognized by the A11 antibody is observed during the polymerization of many amyloid proteins, the A11 antibody binding experiments discussed above indicate distinct differences regarding this common intermediate in CsgA-His polymerization. The A11 antibody may bind to any number of potential CsgA-His intermediates including an off pathway monomer, an off-pathway multimer, or an on-pathway multimer. Alternatively, the A11 antibody may recognize a micelle-like form of CsgA-His that is present in concentrations above 4 μM, explaining both the lack of binding of A11 to unpolymerized CsgA-His below this concentration (Fig. 4.21) and the concentration independence of CsgA-His polymerization above this concentration (Fig 4.1A and 4.24). Collectively, I believe our data suggest the A11 antibody binds to an off-pathway monomer. Consistent with this idea, the A11 oligomer antibody reacts with CsgA-His in a monomeric state (See Fig. 4.2C and 4.15). Also, the A11 antibody does not inhibit CsgA-His polymerization (Fig. 4.18A) even at concentrations below 4 μM (Fig 4.18B), consistent with an off-pathway interaction. Whatever the case, the multiple repeating units of CsgA may allow it to quickly adopt a form recognized by the A11 antibody. This quick adoption of an intermediate should speed up CsgA-His amyloid formation thus lowering host toxicity. Fast cross linking of CsgA-His and the A11 anti body using PICUP (35) may aid in elucidating the stoichiometry and importance of the A11 bound intermediate in CsgA-His polymerization. Future work should aim to characterize this and other intermediate species in CsgA-His fiber formation which may differ from disease-associated amyloids due to either the evolved nature of the functional amyloid fold of CsgA or the requirement of an *in vivo* interaction of CsgA with CsgB for efficient curli fiber formation.

Acknowledgements

Portions of this chapter, mainly Figures 4.1 through 4.4 and the associated text, were published in *JBC* (123). Wang et al 2007 (123) was a collaboration between Xuan Wang, Jonathon Jones, and I. Xuan Wang provided the largest portion of the manuscript and figures sent to *JBC.* The authorship of that paper properly reflects his role; however, much of the original manuscript was unpublished and is presented here for posterity. Xuan provided the work and design for Fig 4.1C-E, 4.3, and 4.4 provided here. Jonathon Jones provided the data for Fig. 4.1B; I designed the figure. The experiments leading to Fig. 4.2A were collectively performed several times by both Xuan and I. The non denaturing purification scheme used was developed by Jonathon Jones, Xuan Wang, and I. The denaturing purification scheme for CsgA-His was my development. I introduced the peptide purification scheme using TFA/HFIP (20) and did early fiber formation, cross seeding, and TEM work with the repeating units R1 and R3 of CsgA; Xuan followed up on and greatly extended this work. All the work and design for the remainder of the figures in the Wang et al 2007 (123), the work and design of Figs. 4.5 through 4.30, and the unpublished portions of this Chapter are solely my work. Additionally, I have used different pronouns in this chapter when appropriate. Work done collectively or by others uses the pronoun we. My contributions are indicated using the pronoun I.

Methods

CsgA-His Purification.

CsgG and CsgA-His were over expressed in LSR12 (C600::*Δcsg*) as previously described (19). Following centrifugation for 15 minutes at 10,000g, the supernatant was clarified by filtration though a 0.22 µm PES bottle-top filter (Corning, Acton, MA). Filtrates containing CsgA-His were passed over a HIS-Select™ HF NiNTA (Sigma Aldrich, Atlanta, GA) column, washed with 10 volumes of 10 mM potassium phosphate buffer (KPi) pH 7.2, and eluted with 10 mM KPi 100 mM imidazole pH 7.2. CsgA-containing fractions were combined and passed through a 0.02 µm Anotop 10 filter (Whatman, Maidstone, England). A modified protocol using guanidine hydrochloride (GdHCl) was employed to fully denature CsgA-His (Fig. 4.5). Following the first wash, the column was equilibrated with 5 volumes 10 mM KPi 8 M GdHCl pH 7.2 and eluted with 50 mM KPi 8 M GdHCl pH 2. Sephadex G25 was used for desalting/buffer exchange. To create CsgA-His seeds, two week old fibers were sonicated using a Sonic Dismembrator (Fisher Model 100, Fisher, Pittsburg, PA) for three 15-second bursts on ice. Where indicated, CsgA-His samples were filtered through a prewashed Amicon Ultra-4 (Millipore, Bedford, MA) centrifugal filter devices. Samples were centrifuged at 4,000g for 2 minutes and the retentate and filtrate fractions were collected. A plasmid encoding CsgA-His can complement *ΔcsgA* cells *in vivo* and purified CsgA-His polymerizes into an amyloid fiber with similar kinetics as Wt CsgA (19).

Thioflavin T (ThT) assay

Following desalting to remove imidazole or GdHCl, CsgA-His was incubated at room temperature. At different time intervals, CsgA-His samples were mixed with 20 µM ThT and fluorescence was measured using a Spectramax M2 plate reader (Molecular Devices, Sunnyvale, CA) set to 438nm excitation and 495nm emission with a 475nm cutoff. Alternatively, samples amended with 25 µM ThT were assayed directly in the Spectramax M2 plate reader every 10 minutes after shaking for 5 seconds. Graphs were made in Kaleidagraph 4.0 and 4.1. ThT assays graphs were smoothed for images. Lag

phase, relative rates, and t½ were determined using Origin 6 and 7 using raw unsmoothed data.

Circular Dichroism (CD) Spectroscopy

CsgA-His samples (10 μM CsgA-His in 50 mM KPi, pH 7.2) were assayed in a Jasco J-810 spectropolarimeter from 190 to 250nm in a quartz cell with 1-mm path length at 25°C.

Blot assay with A11 and anti-CsgA

CsgA-His samples were dripped onto 0.2μm Transblot Nitrocellulose membranes (Bio-Rad, Hercules, CA) as described (57) and allowed to dry for 5 minutes. The membrane was blocked in 5% milk in TBS-T (0.01% Tween-20) for at least 1 hour. The dot blots were washed 3 times in TBS-T before and after incubating with 1:10000 dilutions of A11 primary antibody (BioSource, Camarillo, California) and goat anti-rabbit-HRP (Sigma-Aldrich) in 3%BSA/TBS-T. The blots were developed using the SuperSignal® West Dura system (Pierce, Rockford, IL). Blots were stripped and reprobed with a 1:10000 dilution of rabbit anti-CsgA antibody (6).

Electron microscopy

Philips CM12 Scanning Transmission Electron Microscope was used to visualize the fiber aggregates. Samples (10 μl) were placed on formvar coated copper grids (Ernest F. Fullam, Inc, Latham, NY) for 2 min, washed with deionized water, and negatively stained with 2% uranyl acetate for 90 seconds.

Peptide preparation

Peptides were chemically synthesized by Proteintech Group Inc., Chicago. Purity was greater than 90% by HPLC and size was confirmed by mass spectroscopy. To equilibrate the pH of each sample and to remove any potential seed in the peptide preparations, the peptides were denatured using a modified protocol described previously (20).

Briefly, peptides were dissolved to 0.5 mg/ml in TFA/HFIP (1:1 v/v) and sonicated for 10 min. The suspensions were incubated at room temperature until they visibly cleared. The solvent was then removed by vacuum. Peptides were then dissolved in cold 2 mM HCl and centrifuged at 100,000 X g in a TLA-55 (Beckman, Fullerton, CA) for 1 hour at 4°C. The supernatants were equilibrated to 50mM K-Pi pH7.2 by 200mM K-Pi pH7.2 on ice. When the samples were shifted to room temperature, the polymerization was measured by ThT.

CsgA-His polymerization in different pH buffers

CsgA-His was purified using the alternative GdHCl method described above. Two Sephadex G-25 columns were used in tandem to desalting CsgA-His into MilliQ H_20 after first equilibrating with 10mM potassium phosphate pH 7.2 and then 3 column volumes of MilliQ H_20. The resulting desalted protein was added to black 96 well plates containing 5X concentration of different buffer at the given pHs. The 5X concentrated buffers tested: KBCP is 100mM each of potassium borate, citrate, and phosphate; KAc is 100mM potassium acetate; and 3M is 100mM each of MES, MOBS, and MOPS. Thioflavin T (ThioT) was present at a final concentration of 20 µM and timecourse measurements of ThioT fluorescent were taken as described above. Relative rates were calculated from the ThioT curve by taking the maximum rate divided by the total change in fluorescence and multiplying the result by 1000. Half time or t½ was determined to be the time at which half of the growth phase of CsgA-His polymerization was completed.

Native Gels and Native Western blots

CsgA-His samples were mixed with 1:2 with 3X Native Sample buffer (30% glycerol 0.05% Bromophenol Blue) and loaded onto 8 or 12% Native PAGE gels (3 or 4% PAGE stacking) and ran at 20mA each for 40 minutes at 4°C in Native Running buffer (25mM Tris 187.5mM Glycine pH 8.6 at 25°C; stored at 4°C). For Native Western blots gels were incubated in Native Transfer Buffer (0.8X Native Running Buffer) for 10 minutes and then transferred onto Nitrocellulose membranes (BioRad 162-0147 0.2µM) in using a

semidry apparatus at 10V for 30 minutes at 4°C. The Western blot stack consisted of 3 Whatman papers, gel, Nitrocellulose membrane, three Whatman papers. The Whatman papers and Nitrocellulose gels were soaked in Native Transfer Buffer for more than 10 minutes at 4°C. Following transfer the membrane was blocked at 4°C overnight in 5% Milk in TBST (0.01% Tween-20). The membranes were probed with 1:10000 dilution of A11 Oligomer antibody and 1:10000 Goat Anti Rabbit IgG HRP antibody both in 3% BSA/TBST (0.01% Tween-20) as described above for A11 Dot blots. TBST with 0.01% Tween-20 was used for the three 5 minute washes between before and after each antibody wash and before developing using the Pierce SuperSignal® West Dura system.

Figure Legends

Fig. 4.1. *In vitro* polymerization of CsgA-His measured by ThT fluorescence, CD and TEM. (A) The fluorescence of freshly purified CsgA-His mixed with 25 μM ThT was measured in 10-minute intervals at 495nm after excitation at 438nm. (B) Circular dichroism analysis of 10 μM CsgA-His immediately after purification, 2 days post-purification, 7 days post-purification and 15 days post-purification. CsgA-His was incubated at room temperature without shaking after purification. (C-E) Transmission Electron Microscopy (TEM) micrographs of 30 μM CsgA-His after incubation at room temperature for the indicated times. (Scale bar: 500nm)

Fig. 4.2. Detection of transient conserved intermediate species during CsgA-His polymerization. (A) ThT fluorescence (top) and immunoblotting (bottom) of 80 μM CsgA-His incubated for the indicated time post-purification. At the indicated times, samples were removed, ThT was added to a final concentration of 20 μM, and fluorescence was measured. Samples were blotted onto nitrocellulose membrane and probed with the A11 antibody, and after stripping, with the anti-CsgA antibody. (B) CsgA-His denatured with 8 M GdHCl was blotted onto nitrocellulose and probed with the A11 and anti-CsgA antibodies (left). GdHCl was removed using a Sephadex G25 column (final buffer: 50 mM KPi pH 7.2) and then immediately blotted onto nitrocellulose and probed with the A11 and anti-CsgA antibodies (right). (C) Amicon ultra filters were used to separate CsgA-His solutions prior to probing with the A11 and anti-CsgA antibodies. The molecular weight cutoff of the filters is indicated. Retentates and filtrates were immediately blotted onto nitrocellulose and probed with the A11 antibody, and after stripping, with the anti-CsgA antibody.

Fig. 4.3. CsgA-His fibers can catalyze self-polymerization. (A) CsgA-His (40 μM) fluorescence in the absence (solid line) or presence of 2.5% by weight of sonicated CsgA-His fibers (dashed line). Samples were incubated at room temperature, collected at the indicated times, and amended with 20 μM ThT prior to excitation at 438nm and

measurement at 495nm. (B) Samples were removed at the indicated times and immediately blotted to nitrocellulose. Blots were probed with the A11 antibody and after stripping with an anti-CsgA antibody.

Fig. 4.4. Three CsgA intramolecular peptide repeats can assemble into amyloid fibers.
(A) Alignment of internally conserved residues. CsgA primary sequence shows the repeated consensus sequences. The identical amino acid residues of five repeats are in gray color and the conserved polar amino acid residues are enclosed in 4 boxes.
(B) Oligopeptides of R1, R2, R3, R4 and R5 at 0.2 mg/ml in KPi were incubated at room temperature for 5 days before ThT fluorescence measurements were taken. Error bar indicates the standard error mean for at least three separate experiments. CsgA-His fibers were diluted to 0.2 mg/ml and assayed for ThT fluorescence. (C-E) Transmission Electron Microscopy (TEM). 0.5 mg/ml of R1, R3 and R5 in pH7.2 KPi were incubated at room temperature for 5 days. Samples of different peptide solutions were directly applied on formvar-coated grids and visualized with negative staining electron microscopy. Scale bar is equal to 500nm.

Fig. 4.5. Denatured CsgA-His purification. (A) Overview of denatured CsgA-His purification protocol. (B-G) DS47 was one such denaturing preparation of CsgA-His. (B-D) NiNTA column and elutes: After binding and washing the CsgA-His on 7.5mL NiNTA, the column was equilibrated in 8M GdHCl (Wash 2) and then eluted with 50mM KPi 8M GdHCl pH 2. Twenty five ~1.2 mL elutes were collected and tested for their (B) color; (C) absorbance at 280nM (green), pH (red), glycated protein fluorescence (blue), concentration (BCA Assay); and (D) anti-CsgA and A11 Oligomer antibody binding. (D) Abbreviations: Sup = supernatant from pelleted CsgA-His expressing cells, FT = Flowthrough, W1 = Wash 1, 40mL of 10mM KPi pH 7.2, W2 = Wash 2, 20mL of 10mM KPi 8M GdHCl pH 7.2, E = Elutes; Suffixes for Flowthroughs and Washes = collected at the **B**eginning, **M**iddle, or **E**nd. (B-C) Collectively, the NiNTA elutions shows early brown elutes which come off the column with high concentrations and may indicated glycated

proteins are present Further elution results in even lower pH changes bumping off additional brown elutes which peak around Elute 15. These later elutes also have high absorbance at 280nm and fluorescence similar to glycated proteins; however, according to the BCA Assay their concentrations are quite low. The browning of proteins is indicative of non-enzymatic protein glycation, which has a characteristic fluorescence emission around 450 nm when excited at 360 nm (55) and has been shown to affect the formation of many disease related amyloids (100, 118). Glycation fluorescence was measured on the M2 plate reader as the average of two reads at 360nm excitation, 440nm emission, and 420nm cutoff. (E-G) Gel filtration of a combination of elutes E6 though E11 of the NiNTA column (4.7mLs) after passage though a 0.1µm Anotop filter (Whatman). The Gel filtration column was 6 grams of swelled Sephadex G-25 Fine; it was equilibrated in and eluted with 50mM KPi pH 7.2. Five ~1.3mL flowthrough fractions and forty ~1.3mL elute fractions were collected. Each fraction was tested for (E) absorbance at 280nM (green), pH (red), glycated protein fluorescence (blue), and concentration (BCA Assay) as well as dot blotting (F) onto Nitrocellulose membranes to test for CsgA and A11 binding. (E) Abbreviations: FT = Flowthrough, E = Elutes. (G) Flowthroughs and elutes were also subjected to SDS-PAGE on 13% Acrylamide Gels. Elute 16 was the first elute that showed precipitation due to the presence of GdHCl. Just 20mM GdHCl is sufficient to precipitate Sample Buffer with 2% SDS (85). Collectively, the Gel filtration shows most of CsgA-His elutes quickly through the G-25 column while GdHCl, low pH, and the source of the glycated protein fluorescence are all are small enough to enter the matrix and elute in later fractions. Note: The A11 antibody background signal was high and is not shown here; however, none of the NiNTA elutes had any signal due to the presence of 8M GdHCl. Elutes 5-9 of the gel filtration column had high A11 binding; Elutes 10-13 had weak A11 binding. (H) In a non-denaturing preparation of CsgA-His, passage through a PD-10 gel filtration column separated the small peptide seen in the NINTA elutes on a 13% SDS-PAGE gel stained with Coomassie Blue. The brown color and glycated protein fluorescence in CsgA-His preparations may be due to a small peptide (less than 3kD, the cutoff range for Sephadex G-25) present in

Lysis Broth (LB). This peptide did not readily dialyze away even after two 1000X dialysis dilutions (data not shown) and may associate with CsgA-His and inhibit polymerization: gel filtration columns always removed this brown peptide, but also resulted in decreased lag phases (data not shown).

Fig. 4.6. Effects of pH on CsgA-His polymerization. (A) Effects of pH on CsgA-His polymerization in KBCP buffer as measured by ThT fluorescence. Quantification of CsgA-His polymerization in KPCP buffer: (B) Lag phase (blue) and Half time (t½, orange) = when half of the CsgA-His is polymerized. (C) Delta fluorescence (purple) = change in ThT Fluorescence from beginning of the lag phase to the average fluorescence of stationary phase. (D) Rates of polymerization: Slope (red) = ΔThT fluorescence/min and Relative rate (green) = Slope /Delta fluorescence. CsgA-His was denatured in GdHCl and desalted into Milli-Q H_20 before addition of KBCP buffer to a final concentration of the 20mM each potassium borate, citrate, and phosphate.

Fig. 4.7. Effects of pH on relative rates of CsgA-His polymerization in different buffer systems. (A) Relative rates of CsgA-His polymerization in different buffers as measured with ThioT fluorescence. CsgA-His (final concentration 40µM) was denatured in GdHCl and desalted into Milli-Q H_20 before adding to a final concentration the following buffers: KBCP (20mM each potassium borate, citrate, and phosphate), KAc (20mM potassium acetate), or 3M (20mM each MES, MOPS, MOBS) buffers. (B) Lag phase and half time (t½) of CsgA-His polymerization in different buffers. Similar results were seen in TSC buffer (25mM each of Tris, sodium succinate, and CAPS – data not shown, 2:115-8).

Fig 4.8. Effects of pH on the ThT fluorescence and stability of preformed CsgA-His fibers. (A) ThT fluorescence is lower in more acidic pHs. Preformed fibers were from previous preparations dialyzed in 100 µM KPi pH 7.2 and then mixed with KBCP (here final concentration of potassium borate, citrate, and phosphate is 25mM). Similar

results were seen in TSC buffer (25mM each of Tris, Sodium succinate, and CAPS – data not shown, 2:115-8). (B) Preformed fibers incubated in KBPC pH 11 rapidly lost ThT fluorescence. (C) Absorbance of preformed fibers (green line) incubated in KBPC pH 11 was identical to soluble CsgA-His that was allowed to 'polymerize' in KBPC pH 11 (blue line). Unlike soluble CsgA-His that polymerized in KBPC pH 7 (red line), the pH 11 fibers appear to have very little absorbance near 400nm where absorbance from amyloid fiber aggregates is typically found.

Fig. 4.9. The pH of maximum amyloid formation (pH_{max}) typically occurs near a protein's pI; not so in CsgA-His. Many proteins form amyloids more readily near the pH of their pI (102). At the pI the resulting net charge is nearly zero, solubility decreases, and hydrophobic collapse occurs more readily, thus increasing backbone interactions needed for efficient amyloid nucleus formation. Other amyloids (102) are indicated (black closed circles). Both HEWL and RNase Sa 5Kg (black open circles) have pH_{max} greater than 10 and were not be determined with the ThT method (102). Above pH 10 ThT is either unable to efficiently bind to amyloid fibers or leads to their degradation (See Fig. 4.7 and 4.8). Unlike other amyloids, CsgA-His measured pH_{max} of 8.0 (red closed circle) is far from its pI of 5.73 (red X) (See Fig. 4.6 and Fig. 4.7). The theoretical pH_{max} and pI of mature WT CsgA are indicated (red open circle) for comparison.

Fig. 4.10. Theoretical titration curves of mature CsgA-His, WT CsgA, and CsgB. Titration curves from http://www.iut-arles.up.univ-mrs.fr/w3bb/d_abim/compo-p.html (A) CsgA-His (red) and wild type CsgA (blue). The 6X Histidine tag on the C-terminus of CsgA-His shifts the protein charge towards neutral and then positive, relative to wild type CsgA, at more acidic pH values. (B) CsgB (green) is relatively positively charged at near neutral pH values.

Fig. 4.11. Effects of various salts on CsgA-His polymerization. (A) $NiSO_4$ (blue lines) is a potent inhibitor of CsgA-His polymerization at 1mM. However, addition of 1mM EDTA

restores CsgA-His polymerization (red lines). (B) $AlCl_3$ (blue line) does not inhibit CsgA-His polymerization, neither did $CaCl_2$, $MgCl_2$, or $MgS0_4$ (data not shown, DS7-8). (C) $ZnS0_4$ is also a potent inhibitor of CsgA-His polymerization at 0.1mM. Addition of 0.1 mM EDTA did not rescue CsgA-His polymerization. $ZnCl_2$ also inhibited CsgA-His polymerization but not as effectively as $ZnS0_4$ (data not shown, DS7-8). (D) Addition of EDTA alone does not seem to affect CsgA-His polymerization.

Fig. 4.12. Effects of ionic strength on CsgA-His polymerization and preformed CsgA-His fibers. (A-B) Soluble unfolded CsgA-His was dialyzed for 3½ hours at 4°C in 1mM KPi pH 7.2 using a 3500 MWCO Dialysis membrane. Concentrated KCl was added to each well to the get final KCl concentrations indicated. (data not shown, 2:76-8). (C-D) CsgA-His fibers were dialyzed against 1mM KPi pH 7.2. Concentrated KCl was added to each well to the get final KCl concentrations indicated.

Fig. 4.13. Effects of temperature on polymerization and solubility of non-denaturing preparations of CsgA-His. (A) ThT time course of polymerization of DS37 preparation of CsgA-His at different temperatures. (B) SDS solubility of CsgA-His preparation DS37 before incubation at various temperatures. (C). SDS solubility off CsgA-His preparation DS37 after incubation at various temperatures. 2X SDS sample buffer was mixed equally and not boiled before loading onto 13% SDS-PAGE gels. RT = room temperature, approximately 24°C. DS37 was dialyzed against 50mM KPi pH 7.2. for 4 hours at 4°C after elution from NiNTA under non denaturing conditions and filtered through Anotop 0.2 and 0.02 μm filter.

Fig. 4.14. Effects of temperature on the polymerization, solubility, and A11 Oligomer binding of denaturing preparation of CsgA-His. (A) ThT fluorescence of CsgA-His polymerization incubated at different temperatures. Start of polymerization is 5 minutes after temperature shift. (B) Dot blot time course of CsgA-His incubated at different temperatures probed with A11 Oligomer antibody; following A11 signal detection the

dot blots were stripped and reprobed with Anti-CsgA antibody. Both antibodies were detected using an Alpha Innotech FluorChem imager with all the dotblots being imaged at the same time which allowed equal background. Both resulting images were auto leveled using Adobe Photoshop. (C) Effects of temperature on CsgA-His solubility assayed with Native gels over 3 hours. Each Native gel was auto leveled using Adobe Photoshop except 12% + 3 hours which was manually adjusted to show faint bands. 12% Native gels were ran for 1:25 hour at 25mA. 12% Native gels had 1.8cm of 4% stacking gel; 8% Native gels had 1.8cm of 3% stacking gel.

Fig. 4.15. The A11 Oligomer antibody recognizes a single species of CsgA-His at multiple temperatures. CsgA-His runs as a single band on a native gel which is recognized by the A11 Oligomer antibody. A denaturing prep of CsgA-His was incubated at the different temperatures indicated for 5 minutes and then ran on a 12% Native gel and dot blotted onto Nitrocellulose membranes. The Native gels were natively transferred to Nitrocellulose membranes. Both the Native westerns and the dot blots were probed with the A11 Oligomer antibody and anti CsgA antibodies. Native transfer worked best on 12% Native gel (8% Native gel were unsuccessful and had only faint A11 signal). The CsgA-His sample used for this figure is the same as the one used in Fig. 4.14.

Fig. 4.16. The A11 Oligomer signal is seen even after 5 minutes incubation of CsgA-His at 99°C. A denaturing prep of CsgA-His was incubated at the different temperatures indicated and dot blotted onto Nitrocellulose membranes. Even at 99°C, some A11 Oligomer binding is seen. However, after 1 hour no signal is seen possibly due to hydrolysis of CsgA-His (See Fig. 4.14C and Fig. 4.17). For each time point, the entire dot blot was adjusted to the same levels using Adobe Photoshop so each temperature's signal was treated equally.

Fig. 4.17. High temperatures lead to rapid hydrolysis of CsgA-His. (A) In a non denaturing preparation of CsgA-His, boiling the protein in SDS Sample buffer resulted in

higher MW bands appearing in 13% SDS-PAGE gels. CsgA-His also had more smearing on SDS-PAGE gels if it was preincubated at higher temperatures even if the samples were not boiled in SDS Sample Buffer. (B) For the Native Gels, CsgA-His was prepared by the denaturing protocol. Following shift to higher temperatures, CsgA-His was mixed with Native Gel Sample Buffer and ran on a 12% Native PAGE gel and stained with Coomassie Brilliant Blue. Within 5 minutes at 95^{o}C multiple breakdown products of CsgA-His are present. Within 1 hour hydrolysis is also apparent at 54^{o}C and 65^{o}C. The CsgA-His sample used in the native gels shown is the same as the one used in Fig 4.15. For each time point the entire native gel was auto leveled in Adobe Photoshop. The break seen in each time points gel band represents a cropped control band of BSA that was used to distinguish between different gels. See Fig. 4.14C for more examples of CsgA-His hydrolysis at high temperatures. (C) Time course of denaturation and hydrolysis of a denaturing prep of CsgA-His (DS69) at 95^{o}C. Multiple breakdown products are present on a Novex 4-12% Tris Glycine Gel (Invitrogen) ran with Native Running Buffer; however, A11 binding is present in all samples. Native Mark sizing is consistent with a 200kD protein; however, charge and conformation can have large affects on native gel migration especially if the protein is in an unfolded conformation as CsgA-His is following a denaturing prep. Lysozyme did not run into the native gel; it has a pI of 9.3.

Fig. 4.18. The A11 Oligomer antibody does not inhibit CsgA-His polymerization. (A) When A11 antibody was added to a denaturing prep of CsgA-His, no inhibition was seen even at a molar ratio of 1:30 A11 Oligomer antibody:CsgA; similarly, Anti-His antibody did not inhibit CsgA-His polymerization. This result contrast with the following papers: unseeded but not seeded Sup35 NM showed partial inhibition (~50% less CR binding) at 1:200 and nearly complete inhibition at a 1:100 ratio of A11 Oligomer antibody to Sup35 NM (104); unseeded but not seeded RnqPD showed inhibition at a 1:62 ratio of A11 Oligomer antibody to RnqPD (119). (B) Similar results were seen at lower concentrations of a denaturing prep of CsgA-His (here 3.6µM) where lag phases begin to increase and CsgA-His polymerization more closely follows Nucleation Dependent Polymerization

kinetics. (C) A non-denaturing prep of CsgA-His was also not affected by a 1:22.5 or 1:45 ratios of A11 Oligomer antibody:CsgA-His. No change in polymerization of ThT fluorescence was seen when 1mM total DTT was added to a reaction with 1:25 A11 antibody at the time indicated (black arrow).

Fig. 4.19. The A11 oligomer antibody binds CsgA-His directly after coming off the Sephadex G-25 fine desalting column. (A) Absorbance at 280nm of ~1.3 mL fractions of a Sephadex G-25 fine desalting column (6.1g dry Sephadex). CsgA-His in 8M GdHcl 50mM KPi pH ~5 was eluted using 50mM KPi pH 7.2. (B) Dot blots of elutes of CsgA-His as it directly comes off the Sephadex G-25 desalting column probed with A11 Oligomer antibody and CsgA antibody.

Fig. 4.20. Effects of filtering CsgA-His in GdHCl through a 0.22 µm Cellulose Acetate filter and an Anotop 0.1 µm γ-Alumina based filter on A11 antibody signal. NiNTA elutes are CsgA-His eluted into 50mM KPi 8M GdHCl pH 2 (final pH 4 to 5). NiNTA diluted were diluted by 25% with MilliQ-H20. After filtration through a 0.22µm Cellulose Acetate (CA) membrane, CsgA-His bound A11 antibody in the presence of 8M GdHCl even after further boiling for 5 minutes at 99°C. Unlike the CA membrane the Anotop (Whatman) 0.1 µm γ-Alumina based filter had no effect on A11 binding to CsgA-His in 8M GdHCl.

Fig. 4.21. The A11 oligomer antibody does not bind CsgA-His below 4 µM. Denaturing preps of CsgA-His were diluted in 50mM KPi pH 7.2 and dot blotted onto Nitrocellulose membranes. No A11 antibody signal was seen for concentrations of CsgA-His below 4 µM. (A) Prep DS64 (B) Prep DS65 (C) Prep DS68

Fig. 4.22. Effects of different GdHCl concentrations on A11 antibody binding to CsgA-His. (A) In prep. DS64. CsgA-His was eluted from NiNTA using 50mM KPi 8 M GdHCl pH 2 (final pH was 4.86) and diluted 1:10 or 1:5 with water or 8 M GdHCl to reach the final

concentrations of GdHCl indicated. Afterward mixing 3 µL was dotted onto a nitrocellulose membrane. The mixtures were allowed to incubate for the times indicated at room temperature and again dotted blotted. All the dot blots were probed with A11 Oligomer antibody and an antibody raised against CsgA. Over time A11 is seen binding to CsgA-His in up to 3.2 M GdHCl. In (B) prep. DS65 (C) prep. DS69, a denaturing preparation of CsgA-His was desalted using a gel filtration column and GdHCl was added to the concentrations indicated. At each time point after mixing, a portion of each reaction was dot blotted onto a nitrocellulose membrane, allowed to dry, and probed with the indicated antibodies. A11 antibody is shown binding to CsgA-His with GdHCl concentrations up to 3 M. Higher concentrations of GdHCl does seem to affect CsgA-His binding to nitrocellulose.

Fig. 4.23. Effects of GdHCl on CsgA-His polymerization. GdHCl concentrations above 2.5 M inhibit CsgA-His polymerization resulting in longer lag phases. However, these results are complicated by high concentrations of GdHCl quenched ThT fluorescence of preformed CsgA-His fibers. (A) ThT fluorescence time course of a denaturing prep of CsgA-His (DS64) with GdHCl added to final concentrations indicated. Not shown: 1.6M GdHCl was able to block 6 µM CsgA-His; while 0.8 M GdHCl doubled the lag phase. (B) ThT fluorescence time course of a denaturing prep of CsgA-His (DS65) with GdHCl added to final concentrations indicated. (C) ThT fluorescence of preformed CsgA-His fibers with GdHCl added to the final concentrations indicated.

Fig. 4.24. Effects of CsgA-His concentration on the half time (t½), relative rate, and lag phase of polymerization. (A) The polymerizations of various concentrations of a denaturing prep of CsgA-His (DS51) were monitored with ThT fluorescence. Half time (t½) is indicated with blue dots and is the point at which half of the CsgA-His is polymerized into fibers. (B) The rate of CsgA-His polymerization is linear with concentration. (C) The lag phase of CsgA-His is concentration independent above 4µm.

In multiple other preps, CsgA-His concentrations below ~4µm showed increasing lag phase and concentrations up to 150um showed concentration independent lag phases.

Fig. 4.25. Proposed mechanism of CsgA-His polymerization and A11 Oligomer antibody binding. (A) There are two lag phases in CsgA-His polymerization (see Fig. 4.24). The first lag phase is present below 4µM where lag phase is concentration dependent and CsgA-His appears to follow nucleation dependent kinetics. This lag phase likely represents the time required to form an on-pathway multimeric nucleus (n_a). The second lag phase is concentration independent and is the same for all concentrations above 4µM. This step is likely a maturation of the on-pathway nucleus (n_a) into another nucleus (n_b) of approximately the same stoichiometry. Following n_b formation or further nucleus maturation, CsgA-His monomer are templated to form long amyloid fibers. (B) The A11 Oligomer antibody (purple Y shape) may bind in one of three ways: (1) to an off-pathway monomer; (2) to an off-pathway multimer; (3) to an on-pathway multimer. The A11 Oligomer antibody most likely binds to an off-pathway monomer (1).

Fig. 4.26. Tryptophan fluorescence spectra and iodine quenching of CsgA-His and CsgA-His fibers. (A-F) 2.5µM of a denaturing prep of CsgA-His in 50mM KPi pH 7.2 was tested for changes in Tryptophan fluorescence in a Hitachi F-4500 FL Spectrophotometer. A wavelength scan was done with an excitation wavelength of 295 nm sweeping at 1200nm/min. Both excitation and emissions slit widths were set to 5nm. The GdHCl used was from Fluka Biochemika Ultra pure 99.5% Pure (50935); it is spectral quality, which has low light absorption at 260 and 280nM. (A) Unpolymerized CsgA-His shows typical tryptophan fluorescence from a tryptophan in a polar solvent (maximum emission around 350nM). The Trp 84 of unpolymerized CsgA-His also appears readily available for fluorescent quenching with potassium iodine in the solvent. Potassium chloride does not readily quench intrinsic fluorescence of tryptophan and was used as a control. (B) Unpolymerized CsgA-His in 4M GdHCl does not as readily quench with KI. (C) Comparison of tryptophan fluorescence of unpolymerized CsgA-His with and without

4M GdHCl: GdHCl increases CsgA-His tryptophan fluorescence. 0.2M KI less readily quenches the tryptophan of CsgA-His in 4M GdHCl. (D) Comparison of tryptophan fluorescence of unpolymerized CsgA-His and CsgA-His fibers. CsgA-His fibers have lower intensity of tryptophan fluorescence, consistent with partial burying of Trp 84 or possible quenching from nearby residues (21). Addition of 4M GdHCl has the opposite effect on tryptophan fluorescent of CsgA-His fibers. CsgA-His fibers with (E) and without (F) 4M GdHCl showed some quenching with the addition of 0.2M KI consistent with Trp84 being partially solvent exposed. Like tryptophan fluorescence, light scattering at 360nm could be used to assay low concentrations of CsgA-His polymerization. A few early experiments were done which appeared to show rapid polymerization of CsgA-His by the R1 subunit fibers within minutes of addition to the reaction; however, these early experiments need to be repeated with better controls.

Fig. 4.27. CsgA-His polymerization time course assayed with tryptophan and ThT fluorescence in the Molecular Devices Spectramax M2 plate reader. The tryptophan fluorescence (excitation 295 nm and emission 360 nm) of CsgA-His appears to increase in a manner similar to ThT fluorescence but with a lower sensitivity. These results oppose those seen in Fig. 4.26 and may be due to the following. (1) The M2 plate reader does not read emission at a 90º angle to the incident light and also lacks selectable slits to reduce stray light. The increased Trp fluorescence may therefore reflect the geometry of M2's fluorometer: increased light scattering from more turbid CsgA-His fibers results in increased scattering of incident light. (2) CsgA-His Trp Fluorescence originally increases with fiber formation but may decrease over time as the fibers mature (See the changes in CD spectra over 15 days in Fig. 4.1B). Attempts to read ThT and Trp fluorescence in the same sample failed. Addition of ThT interferes with Trp fluorescence and resulted in small non-reproducible changes.

Fig. 4.28. Sonication of CsgA-His fibers results in A11 oligomer antibody binding. (A) Sonication of preformed fibers of CsgA-His and a peptide corresponding to the R1

repeating units resulted in A11 oligomer antibody binding. Sonication also resulted in A11 binding to CsgA-His that was rapidly concentrated in a 10 NMWL Amicon Ultra-4 filtration device (Millipore) immediately following purification. The rapid concentration prevented amyloid fibers from forming, pushing CsgA-His down an amorphous polymerization pathway. (B) Once sonicated, these fiber preps appear to retain their A11 antibody signal for several days.

Fig. 4.29. CsgA is a predicted to be a natively unfolded protein. (A) The Uversky et al 2000 (117) algorithm was incorporated into Foldindex (87) which is available online at (http://bip.weizmann.ac.il/fldbin/findex#info). Foldindex settings were window size = 10 step size = 1. Data from the "show windows' verbose detail" setting were graphed in Kaleidagraph 4.1; Adobe Photoshop was used to fill in regions where Foldindex scores were positive (green) for folded regions and negative (red) for unfolded regions. The overall folding index score for mature CsgA (without the Sec secretion signal region) is -0.027 (charge: 0.046, phobic: 0.420) suggesting CsgA is an unfolded protein (87). (B) At default settings, Globprot 2 (67) which is available at (http://globplot.embl.de/) predicted most of CsgA is disordered. (C) At default settings Disembl 1.5, which is available at (http://dis.embl.de/), also predicted CsgA is mostly disordered according to a "loops or coil" definition (blue line) (66). Linding et al 2004 (68) discusses the propensity of many natively unfolded and globular proteins to aggregate and form amyloids. (D) Disopred (126) and Psipred (77) predictions of the disordered and secondary structure of CsgA, respectively. Disopred2 predictions were performed at (http://bioinf.cs.ucl.ac.uk/at) using false positive rate thresholds of 2% and 10%. (*) represent disorder predictions and (.) prediction of order. The confidence estimates (graphed linearly as bars from 0-9) give a rough indication of the probability that each residue is disordered. Disopred results were affected by the presence or absence of the Sec signal region of CsgA; the results of CsgA without the sec signal are indicated as "Dis-Sec 2%" and "Dis-Sec 10%". Psipred version 2.6 was performed using a filtered data set: See Chapter III and Appendix A for more details. For Psipred, removal of the Sec signal

sequence did not appreciably affect the predictions. Unlike the other disorder predictors, Disopred did not predict CsgA to be particularly disordered and may reflect the training set on which it was made (87). Note: FoldUnfold (40) which is based on Galzitskaya & Garbuzynskiy 2006 (39) was originally used to show CsgA-His was predicted to be unstructured (123). However, as of this writing the site at (http://skuld.protres.ru/~mlobanov/ogu/) was not available. There are many other protein disorder predictors; Disprot.org lists 19 such predictors at (http://www.disprot.org/predictors.php).

Fig. 4.30. Theoretical Isoelectric points (pI) of Repeating Units of CsgA and CsgB. Theoretical pI's are from Proteine (88) and Protparam (89). Acidic residues are colored in Red, Basic in Blue, Tyrosine and Histidine (pka 10.07 and 6.10 respectively) in purple. Gray bars represent aligned stacks of residues found in at least four repeating units.

Fig. 4.31. The effects of temperature and GdHCl on the mobility of CsgA-His in Native or SDS-PAGE gels and on A11 Oligomer antibody binding. A denaturing preparation of CsgA-His (DS65) was subjected to Native gels, SDS-PAGE gels, and A11 Oligomer antibody binding. (A) 13% Native gel showing CsgA-His runs primarily as a single species even at elevated temperatures. The addition of 2M GdHCl did not affect the mobility of CsgA-His. (B) 13% SDS-PAGE samples of the reaction above. The breakdown of CsgA-His is apparent at 95^{o}C especially at 35 minutes. (C) Dot blots show A11 oligomer binding is present at multiple temperatures but is absent in 8M GdHCl just off the NiNTA column and when 2M GdHCl is added back to CsgA-His that has been desalted on a Sephadex G-25 Fine column into 50mM KPi pH 7.2.

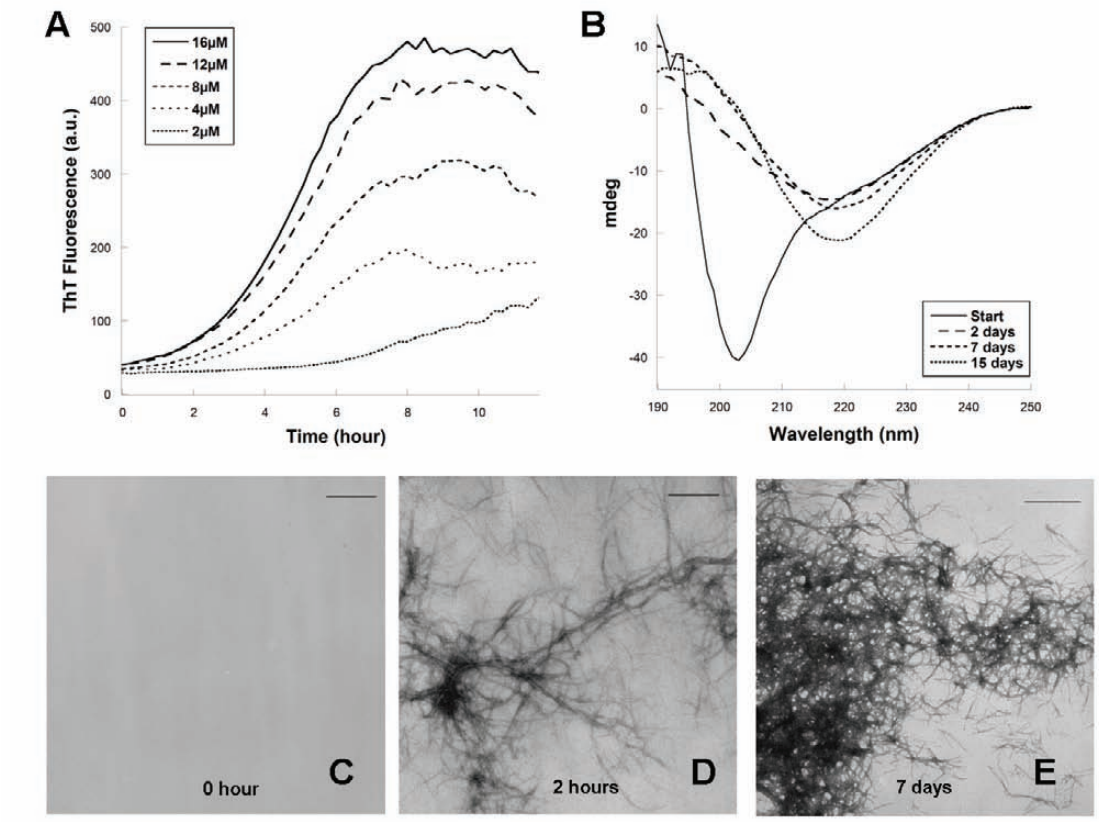


Figure 4.1. *In vitro* polymerization of CsgA-His measured by ThT fluorescence, CD and TEM.

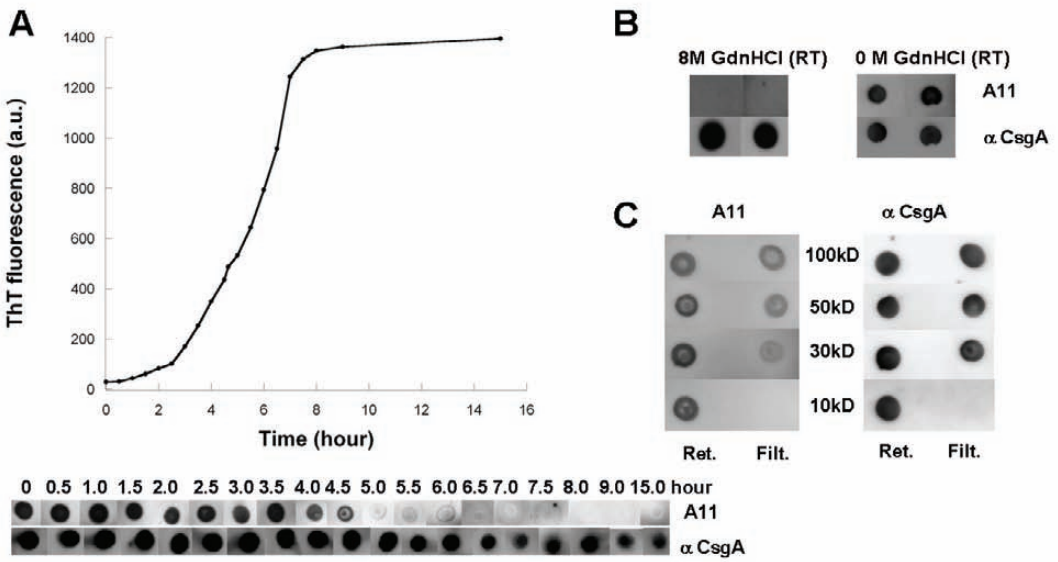


Figure 4.2. Detection of transient conserved intermediate species during CsgA-His polymerization.

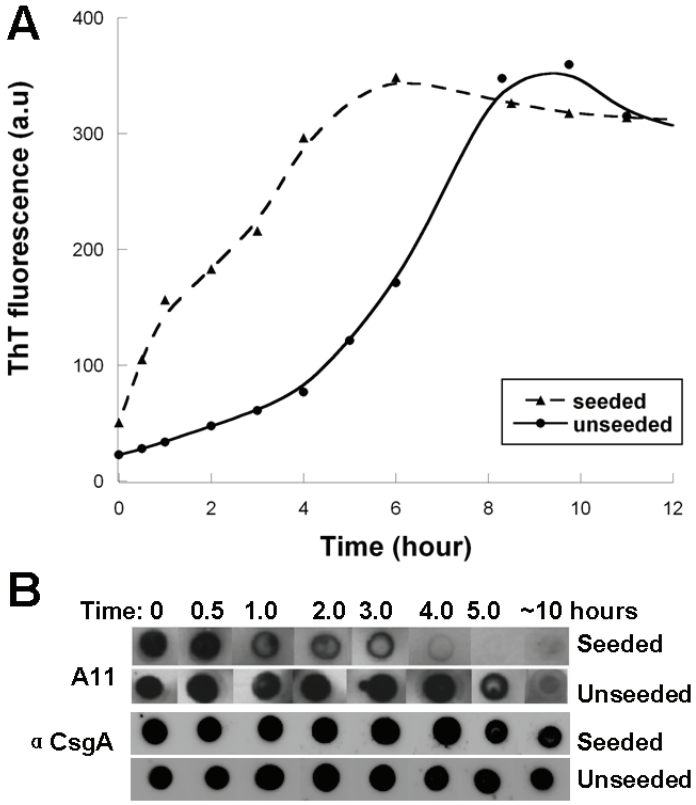


Figure 4.3. CsgA-His fibers can catalyze self-polymerization.

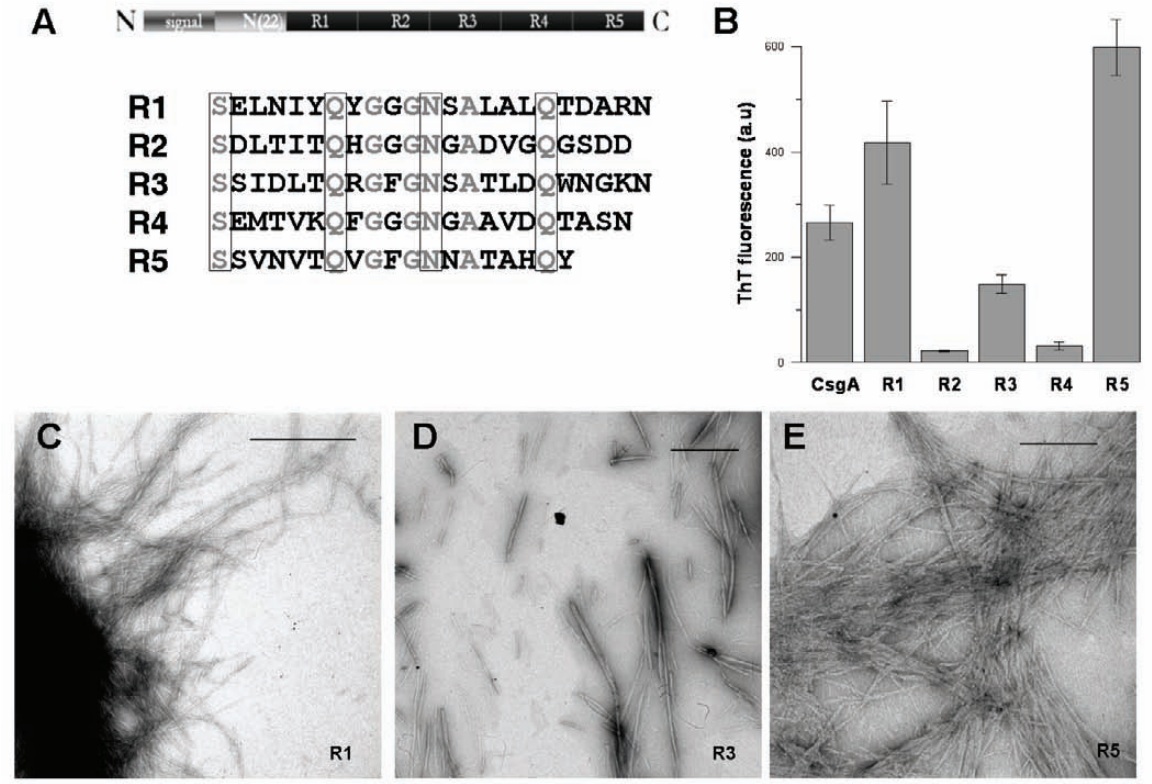


Figure 4.4. Three CsgA intramolecular peptide repeats can assemble into amyloid fibers.

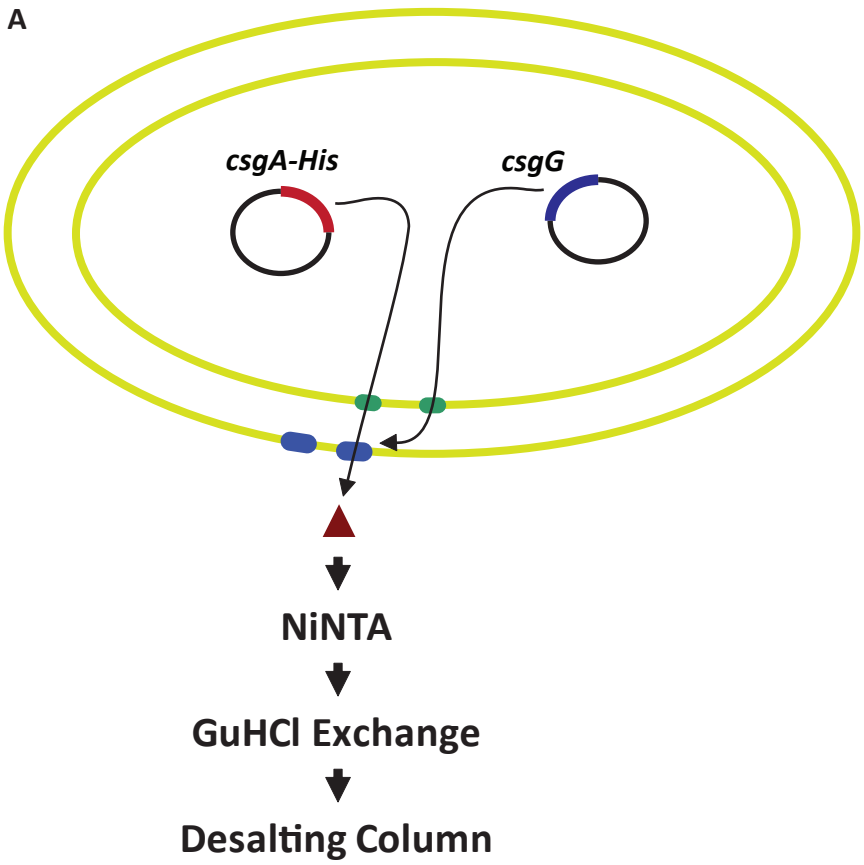


Figure 4.5. Denatured CsgA-His purification.

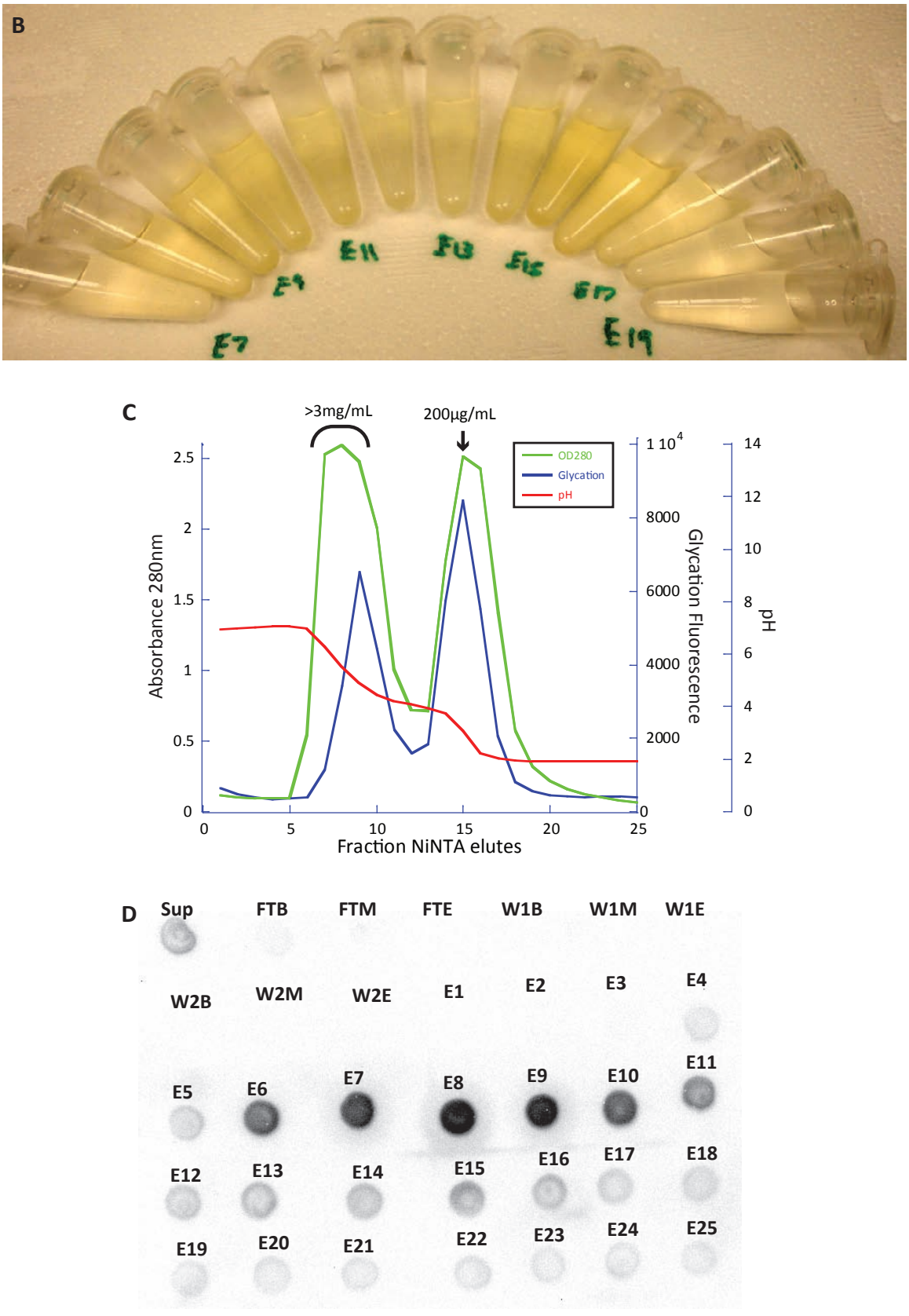
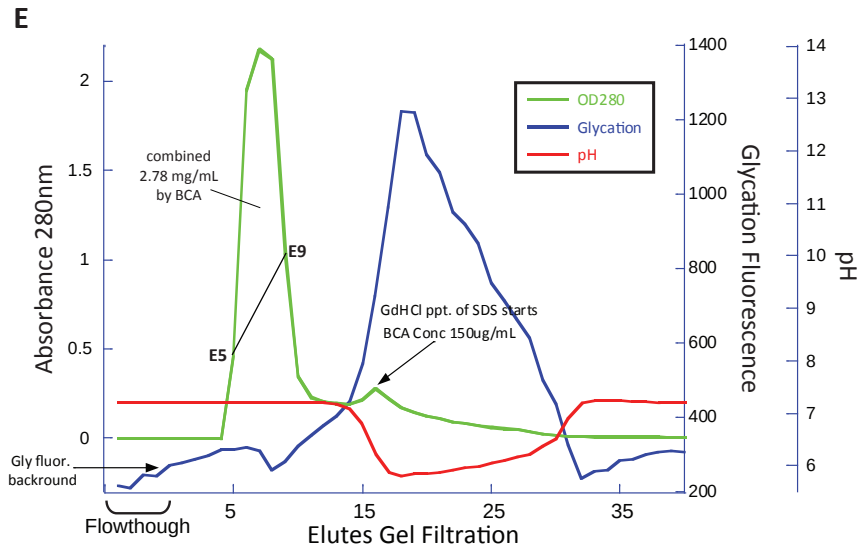
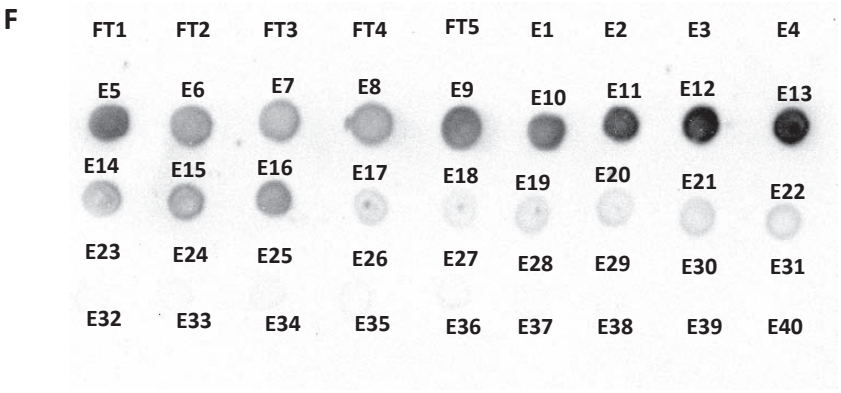
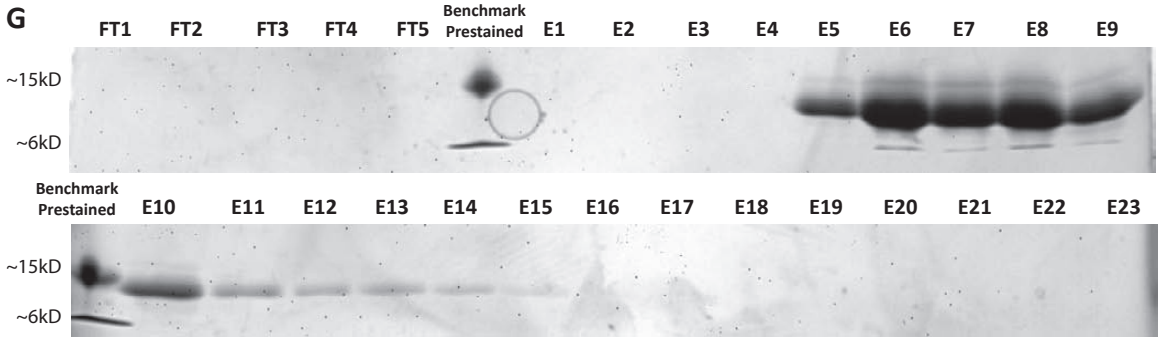


Figure 4.5 (continued). Denatured CsgA-His purification.







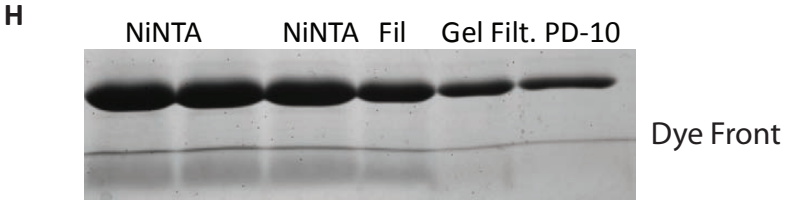


Figure 4.5 (continued). Denatured CsgA-His purification.

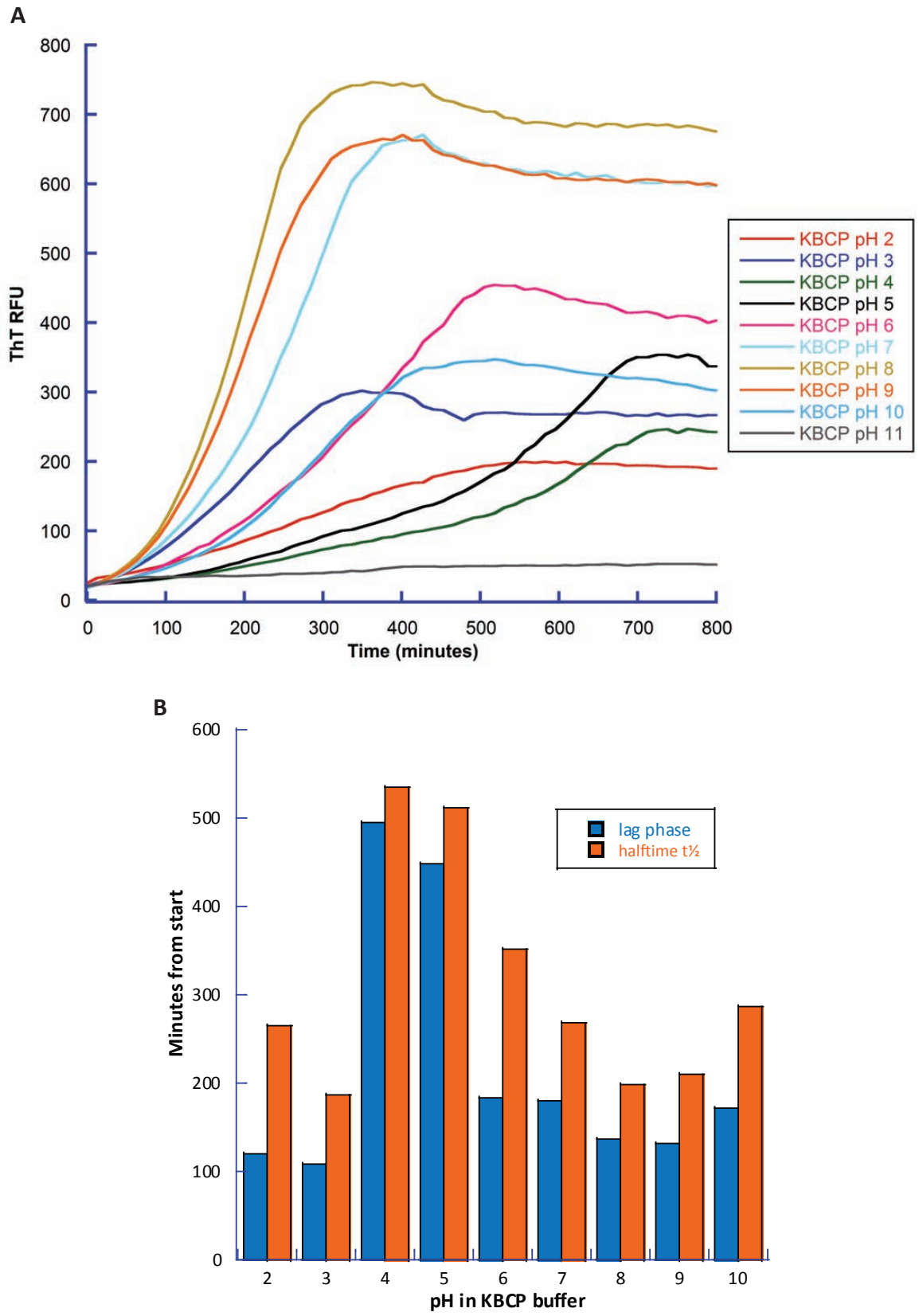
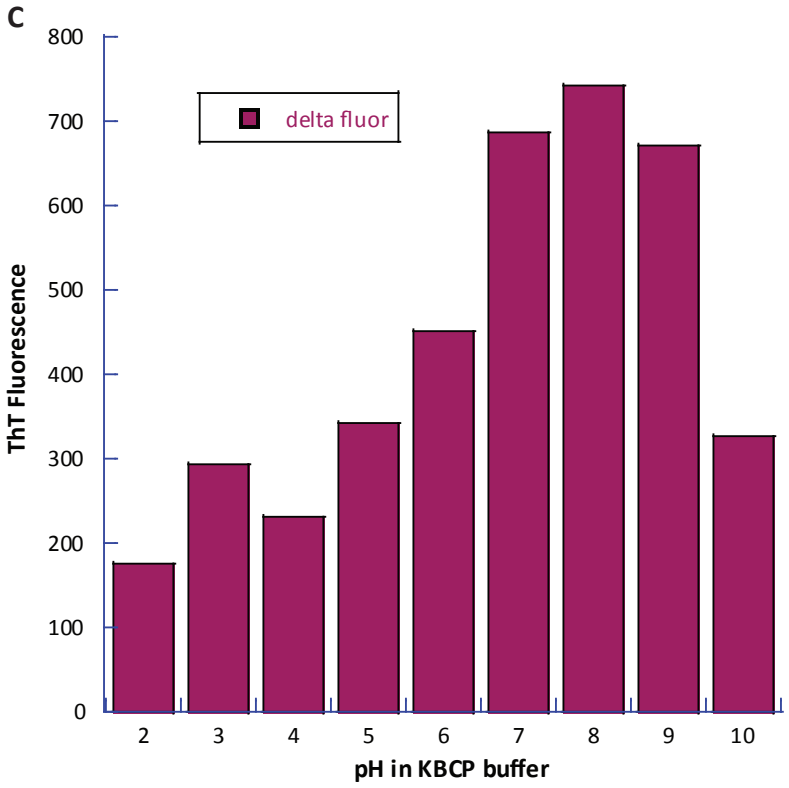


Figure 4.6. Effects of pH on CsgA-His polymerization.



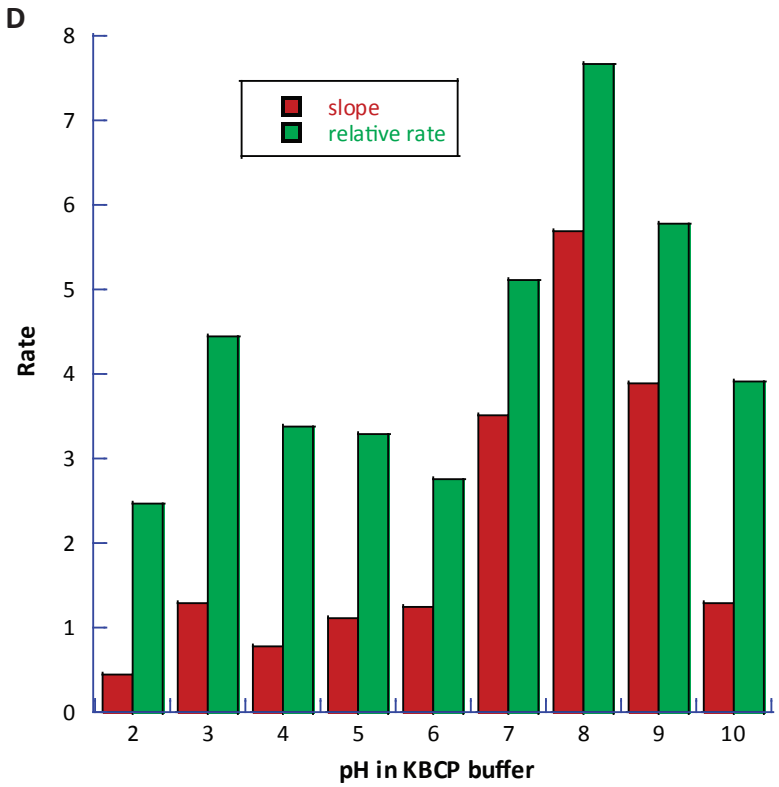


Figure 4.6 (continued). Effects of pH on CsgA-His polymerization

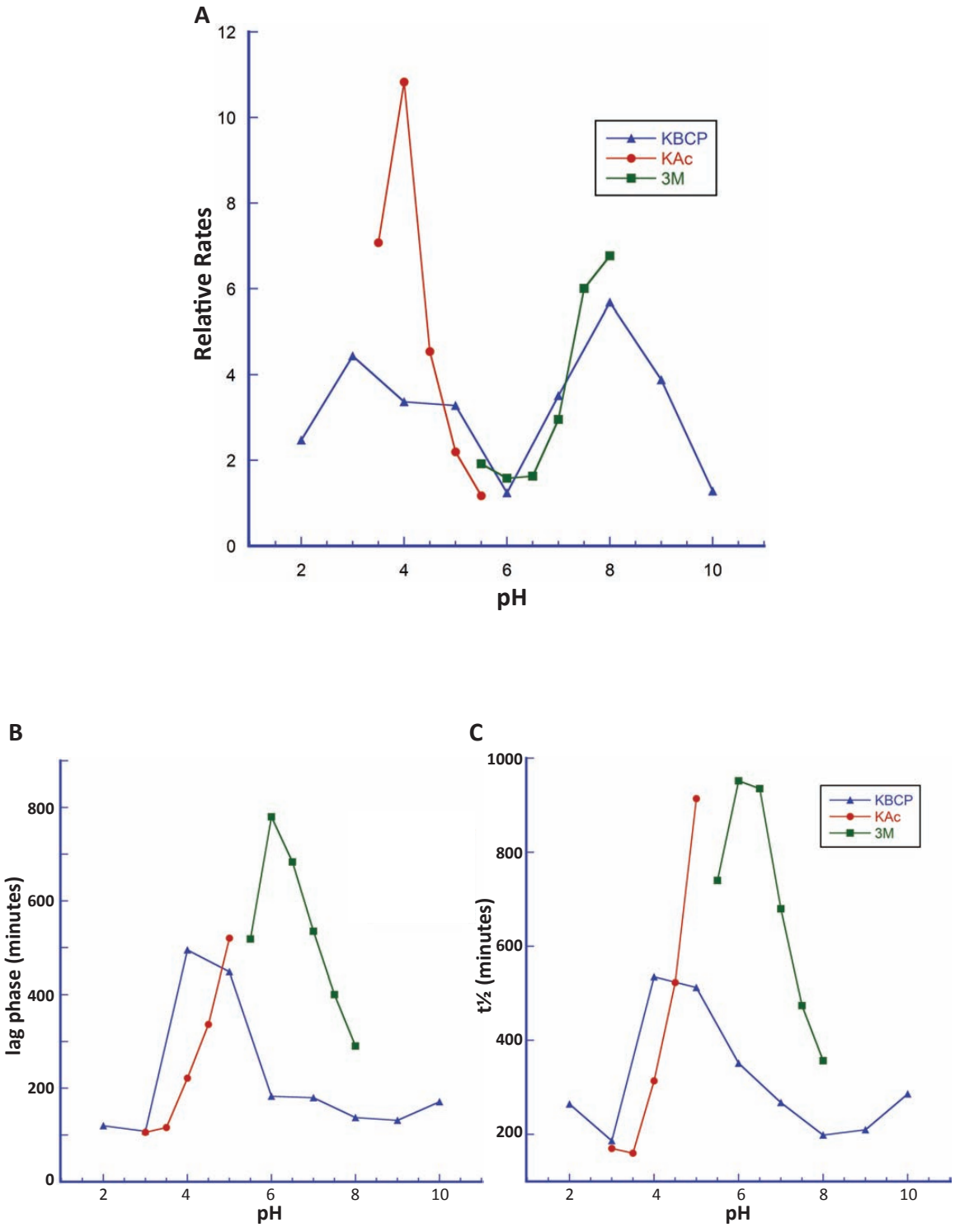


Figure 4.7. Effects of pH on relative rates of CsgA-His polymerization in different buffer systems.

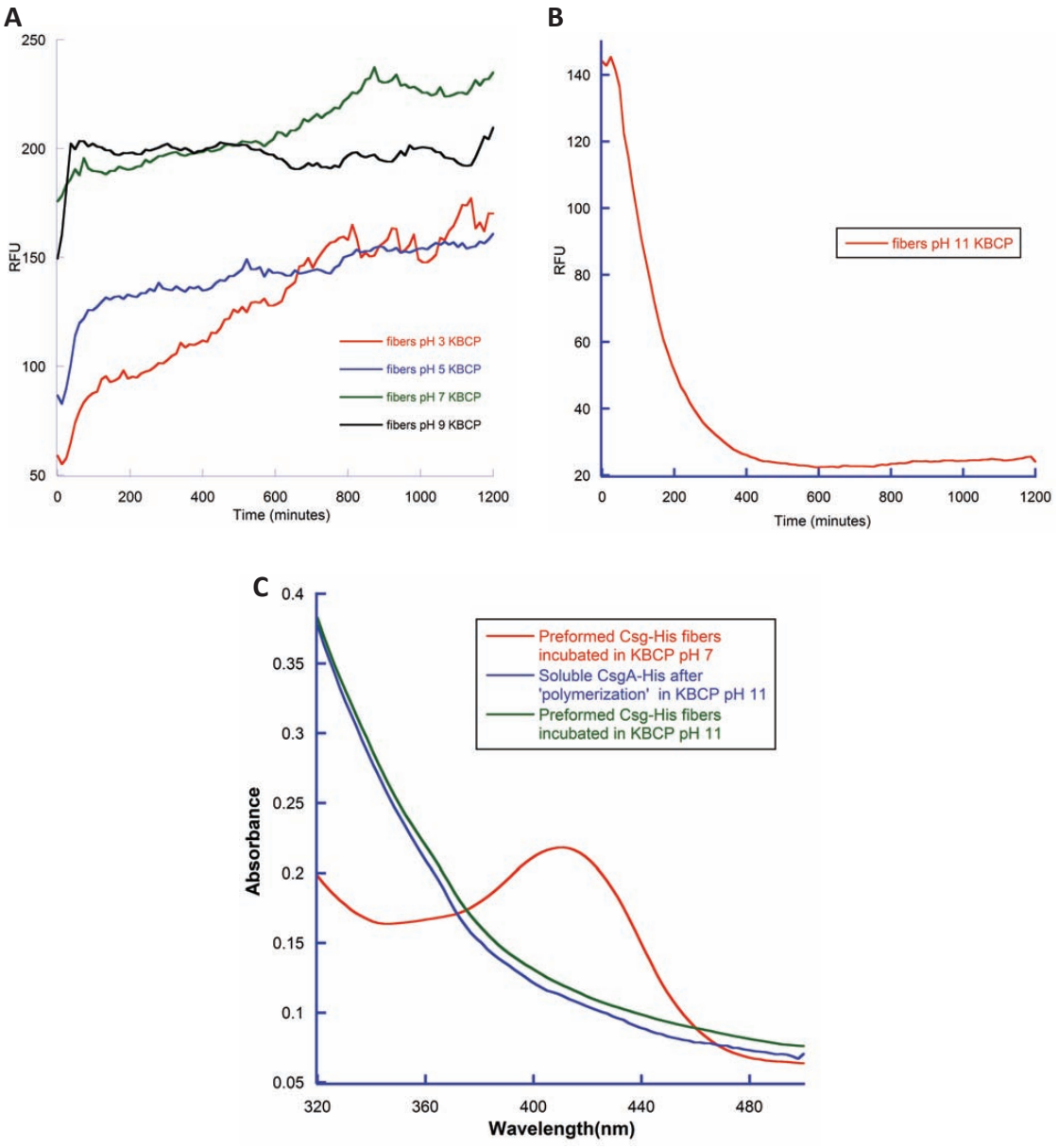


Figure 4.8. Effects of pH on the ThT fluorescence and stability of preformed CsgA-His fibers.

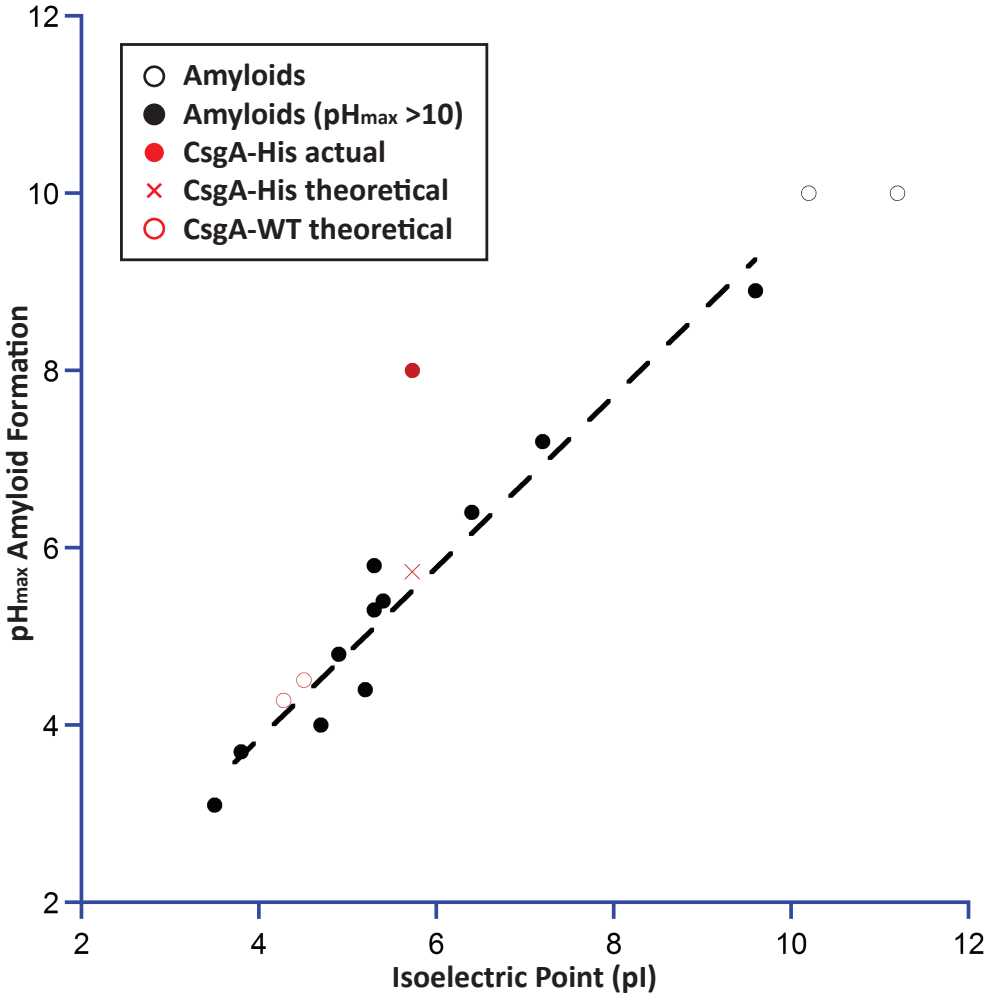


Figure 4.9. The pH of maximum amyloid formation (pH_{max}) typically occurs near a protein's pI; not so in CsgA-His.

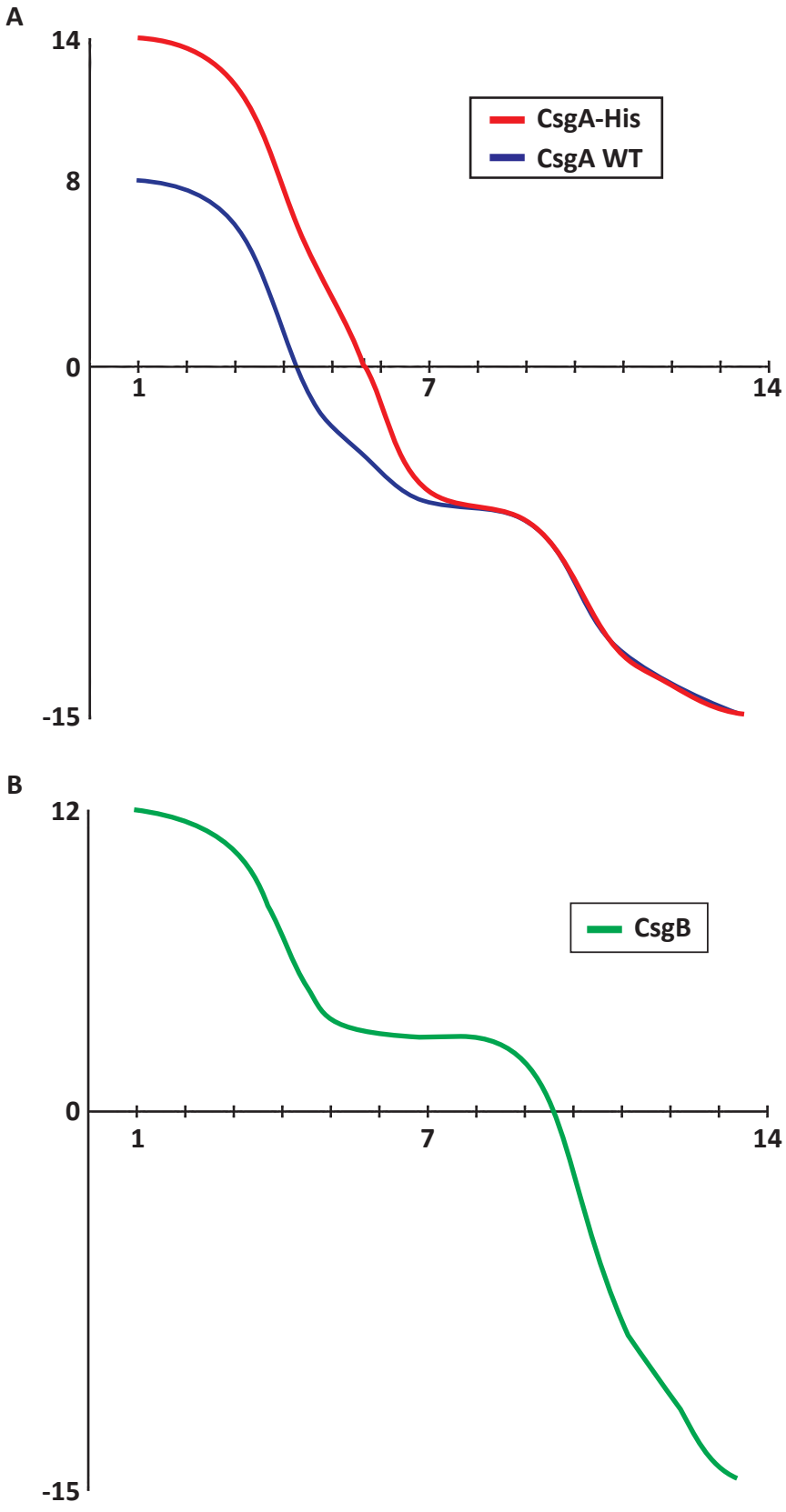


Figure 4.10. Theoretical titration curves of mature CsgA-His, WT CsgA, and CsgB

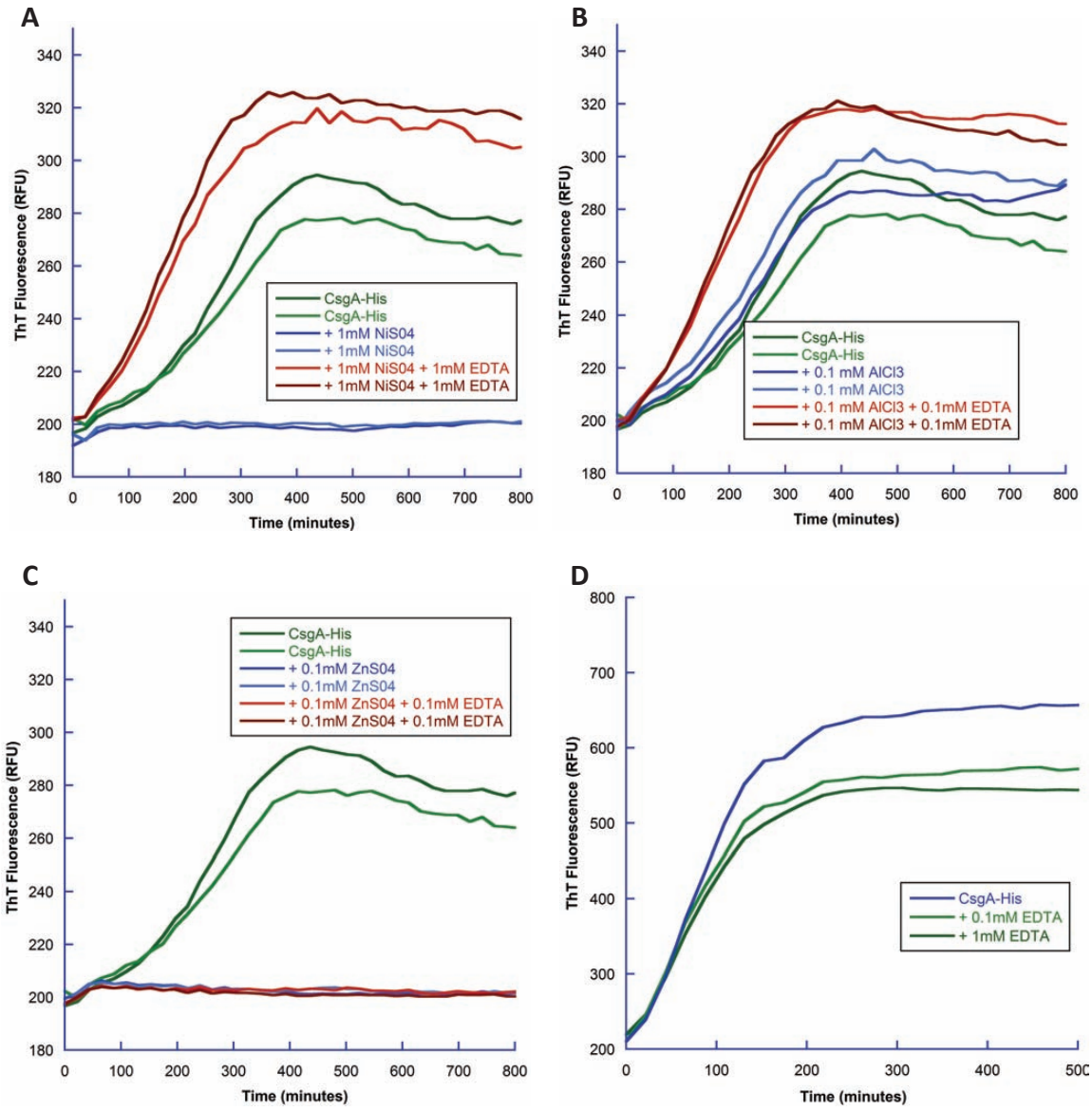


Figure 4.11. Effects of various salts on CsgA-His polymerization.

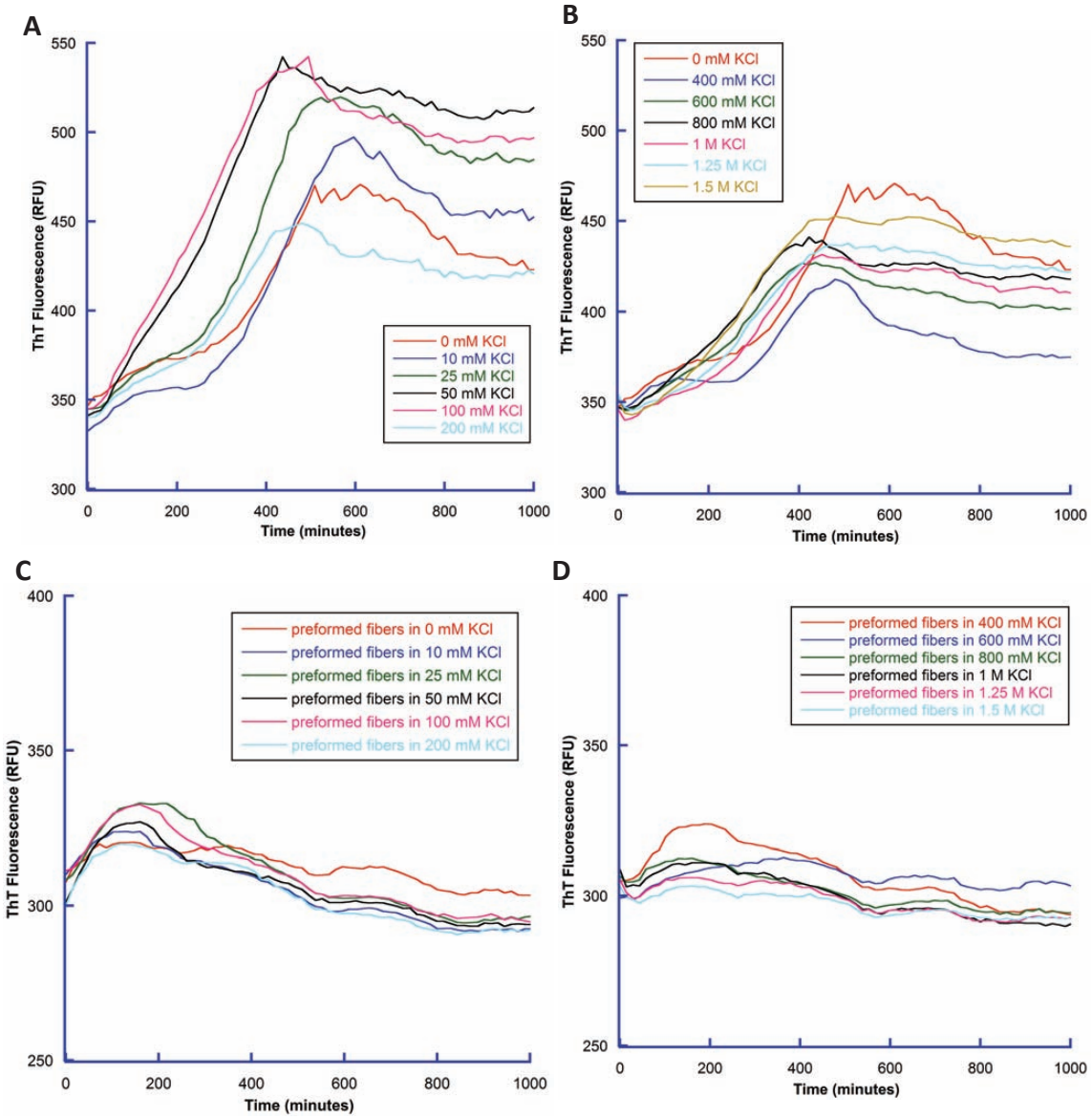


Figure 4.12. Effects of ionic strength on CsgA-His polymerization and preformed CsgA-His fibers.

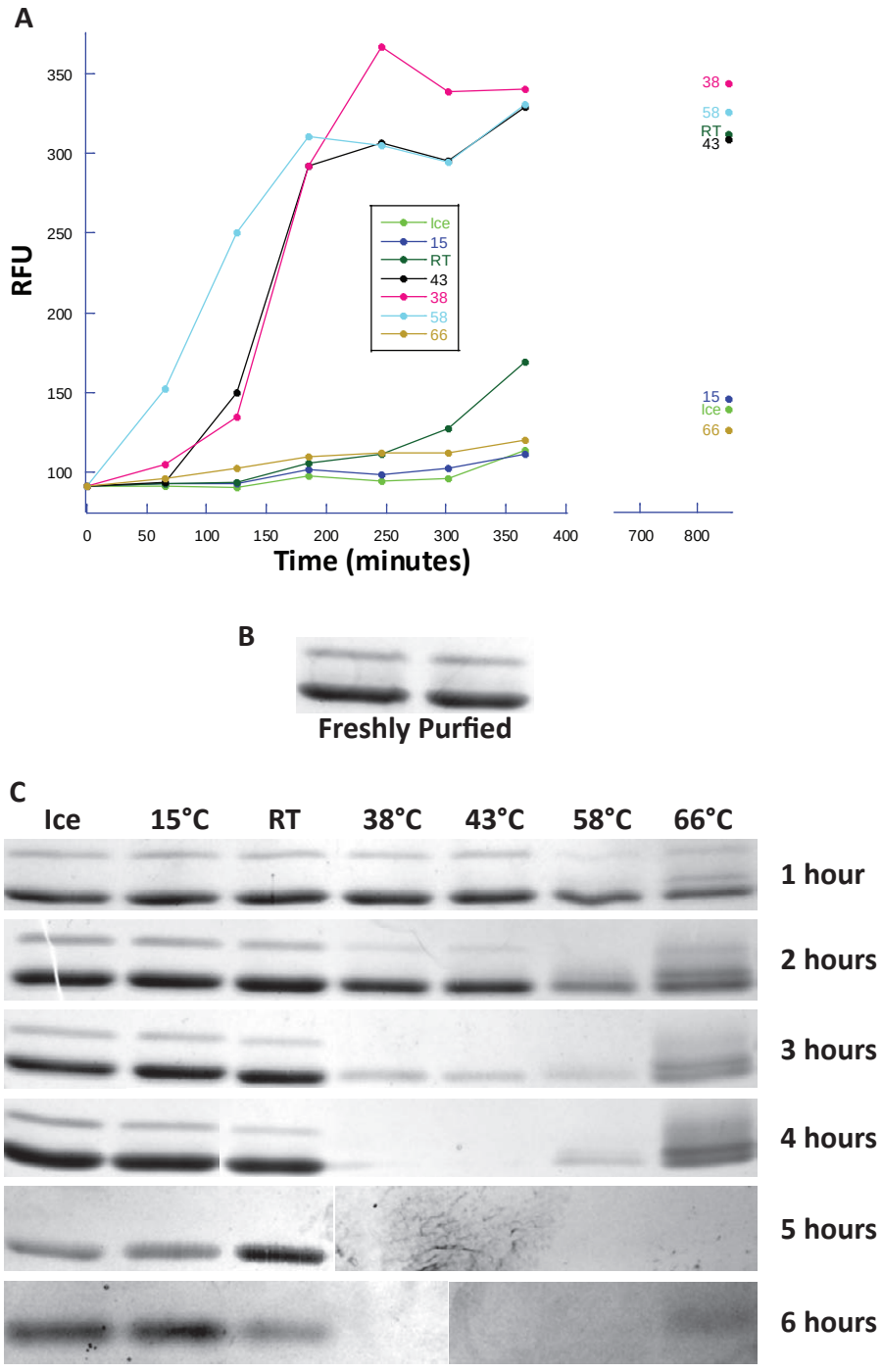


Figure 4.13. Effects of temperature on polymerization and solubility of non-denaturing preparations of CsgA-His.

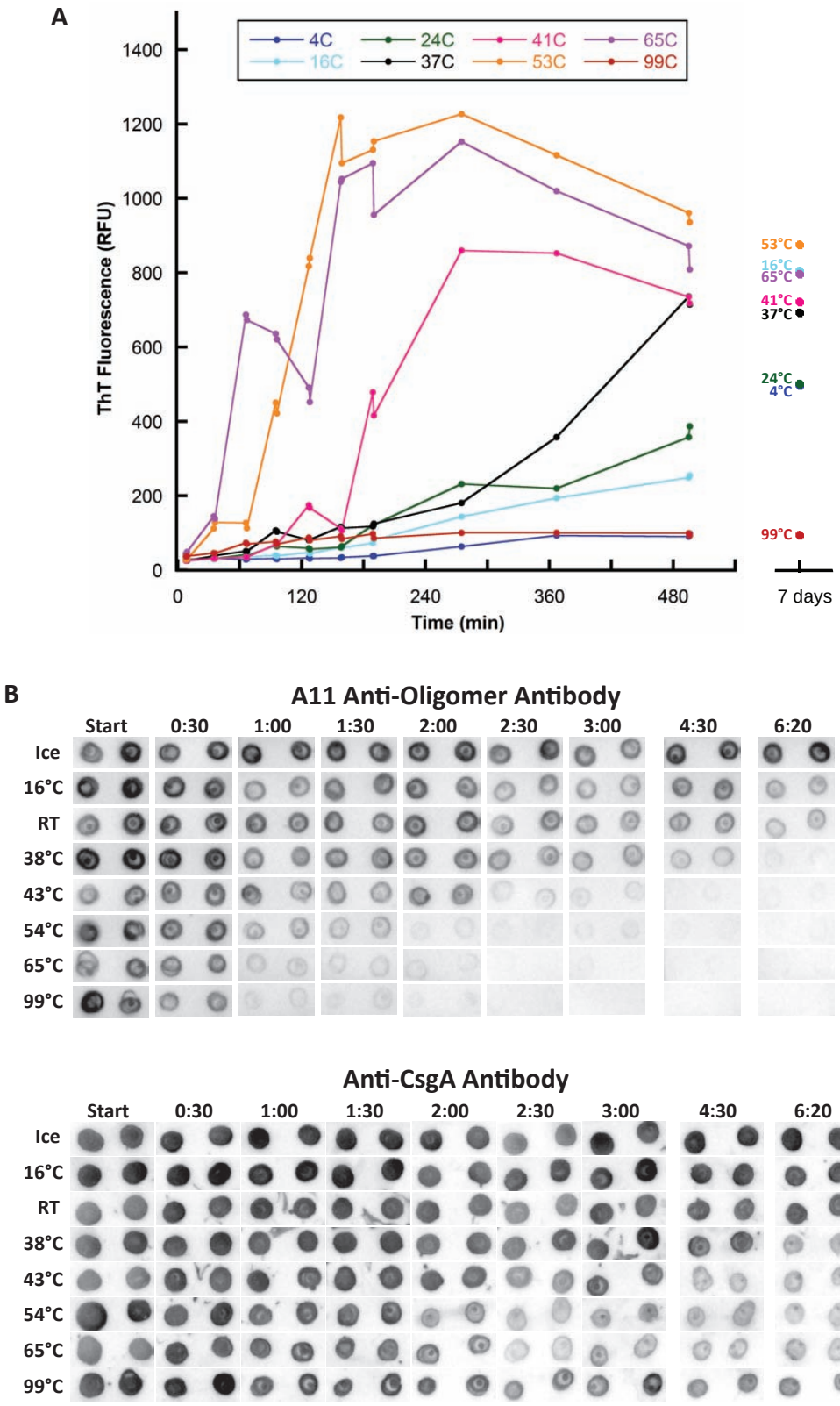


Figure 4.14. Effects of temperature on the polymerization, solubility, and A11 Oligomer binding of denaturing preparation of CsgA-His.

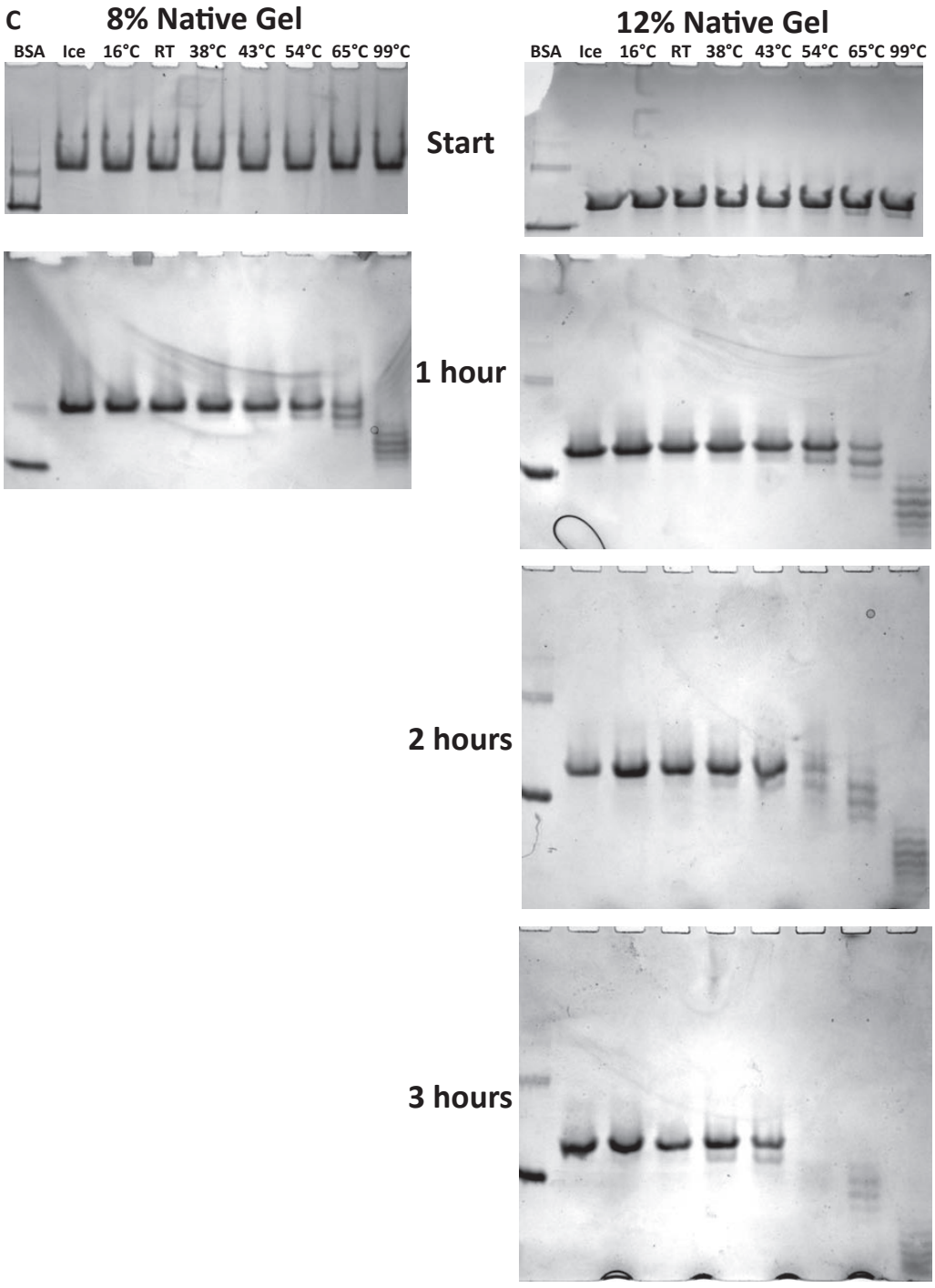


Figure 4.14 (continued). Effects of temperature on the polymerization, solubility, and A11 Oligomer binding of denaturing preparation of CsgA-His.

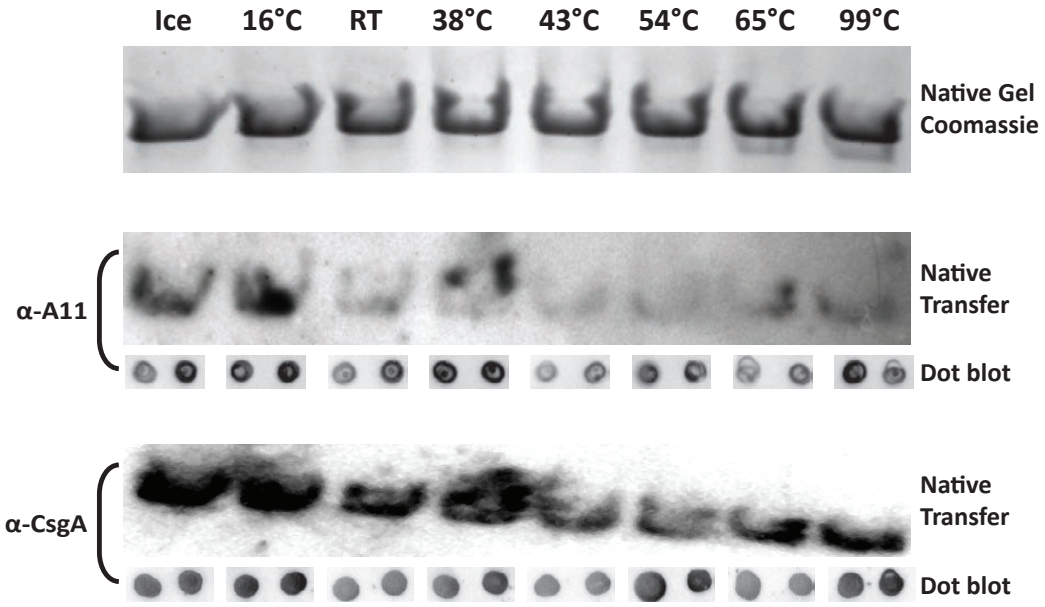


Figure 4.15. The A11 Oligomer antibody recognizes a single species of CsgA-His at multiple temperatures.

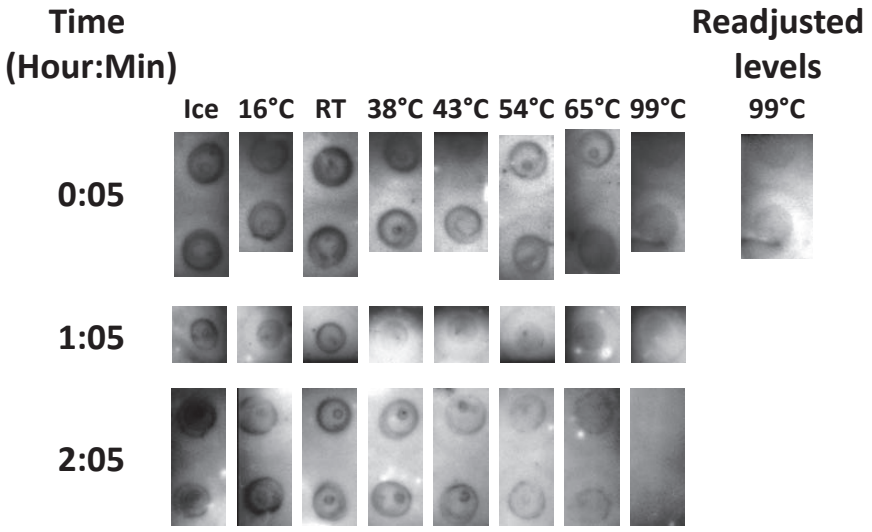


Figure 4.16. A11 Oligomer signal seen after 5 minutes incubation of CsgA-His at 99°C.

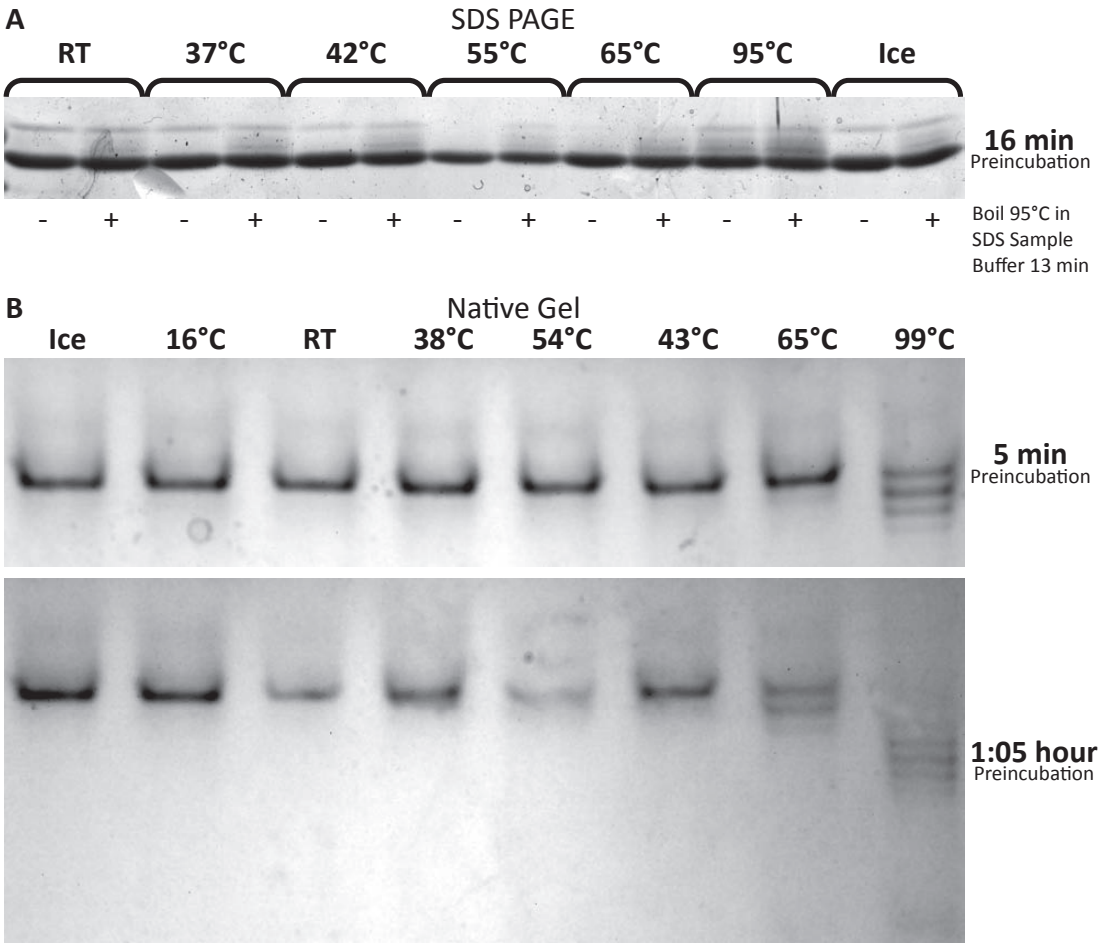


Figure 4.17. High temperatures lead to rapid hydrolysis of CsgA-His

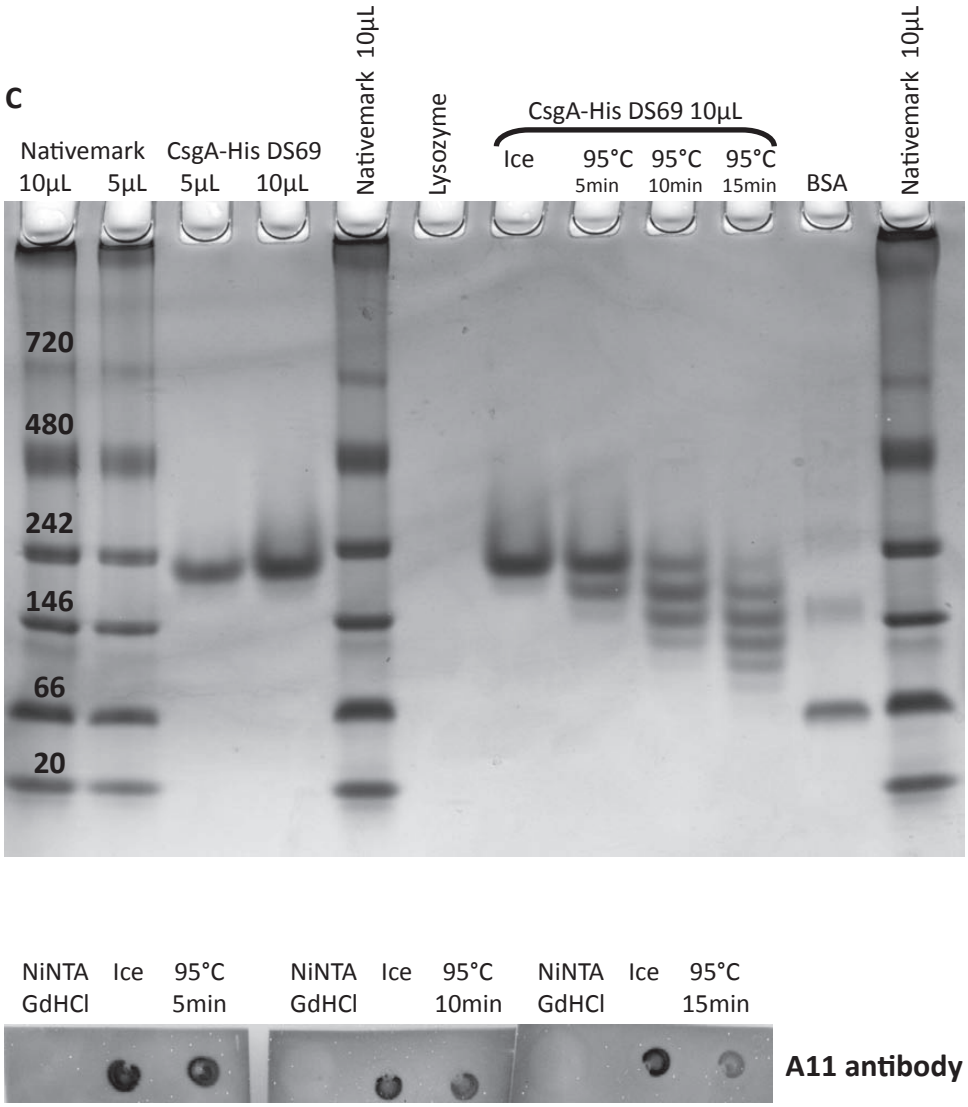


Figure 4.17 (continued). High temperatures lead to rapid hydrolysis of CsgA-His

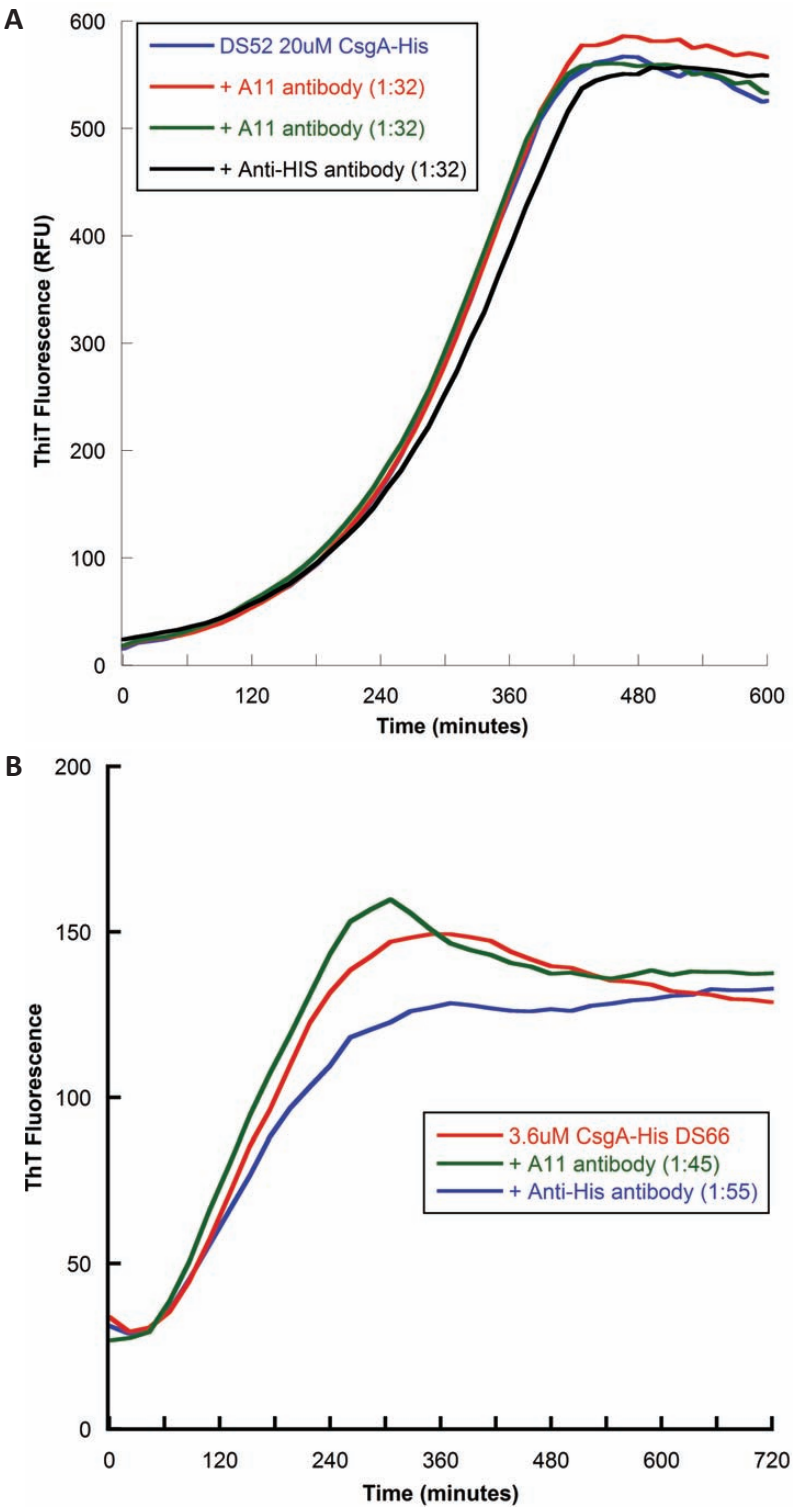


Figure 4.18. The A11 Oligomer antibody does not inhibit CsgA-His polymerization.

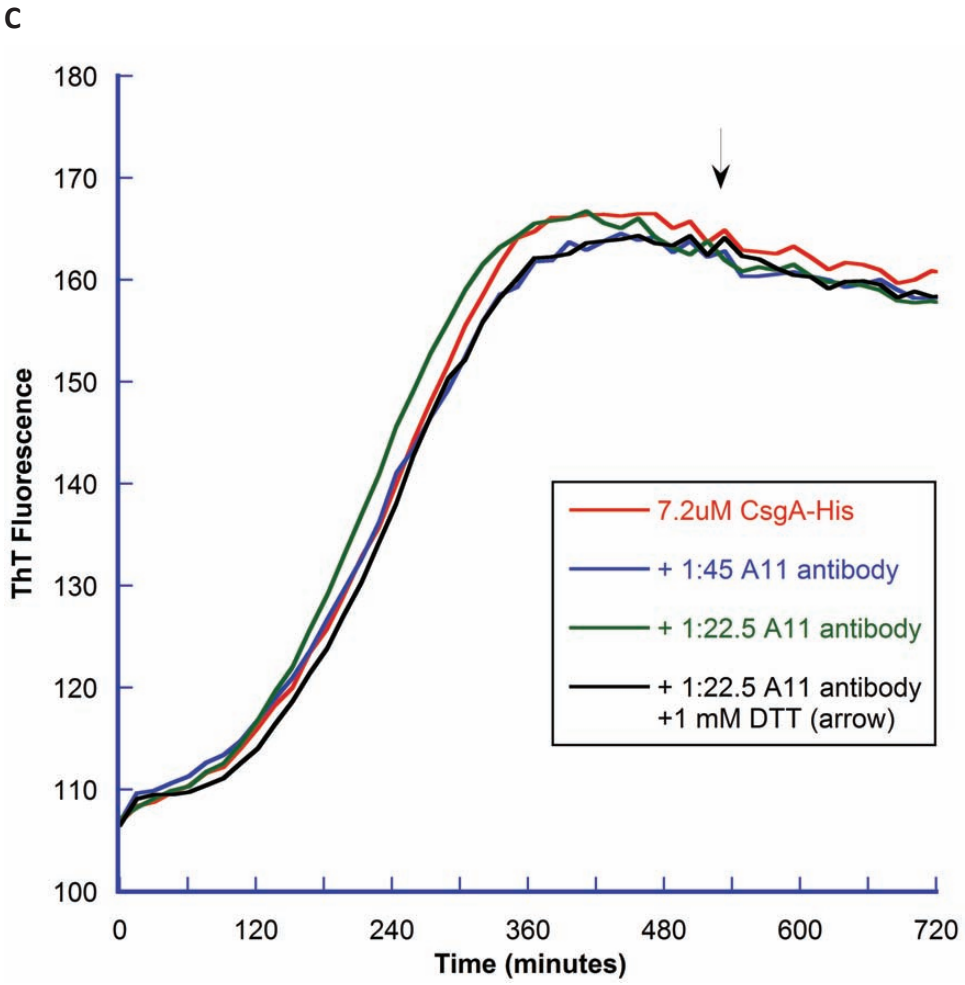


Figure 4.18 (continued). The A11 Oligomer antibody does not inhibit CsgA-His polymerization.

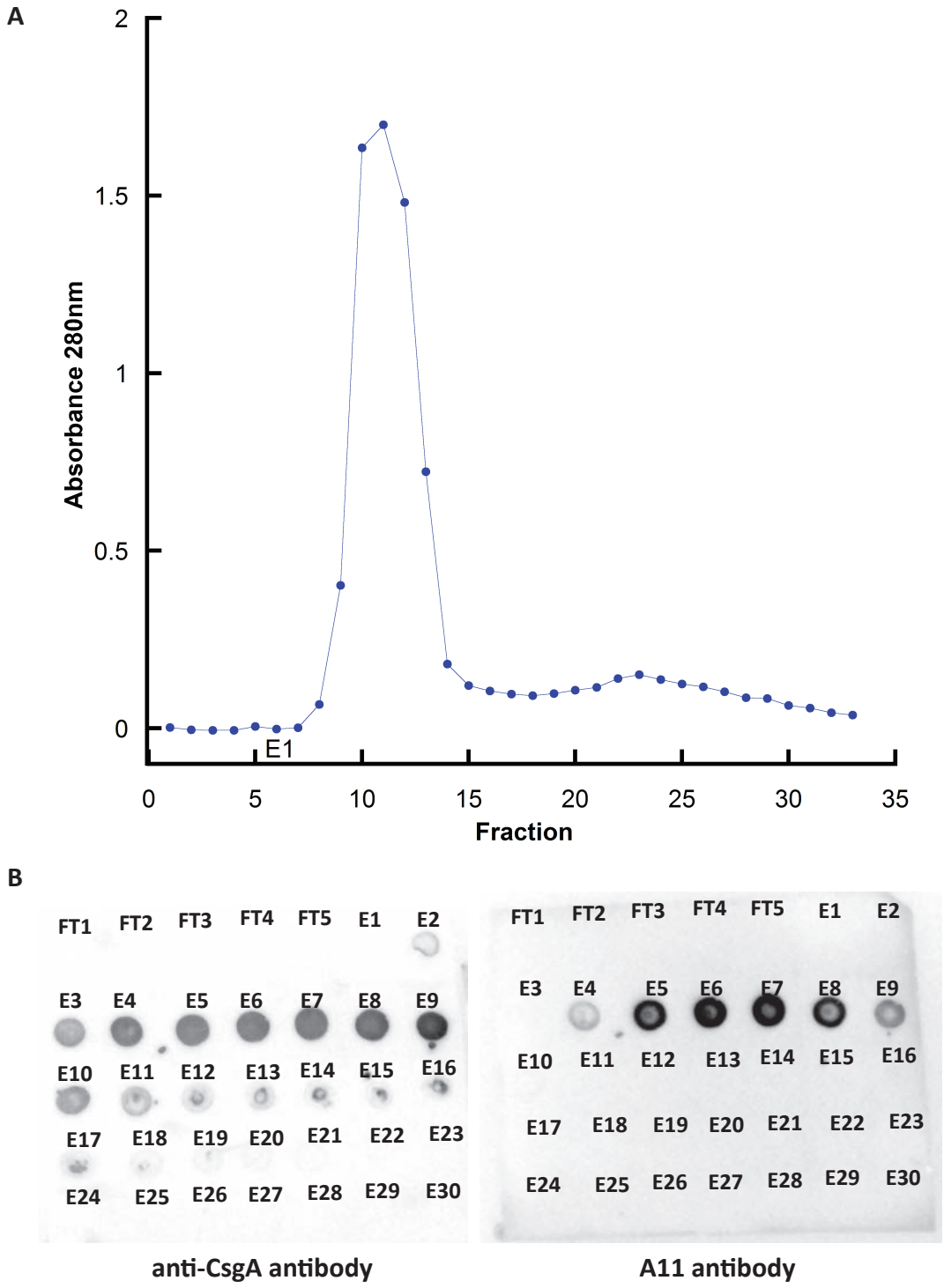


Figure 4.19. The A11 oligomer antibody binds CsgA-His directly after coming off the Sephadex G-25 fine desalting column.

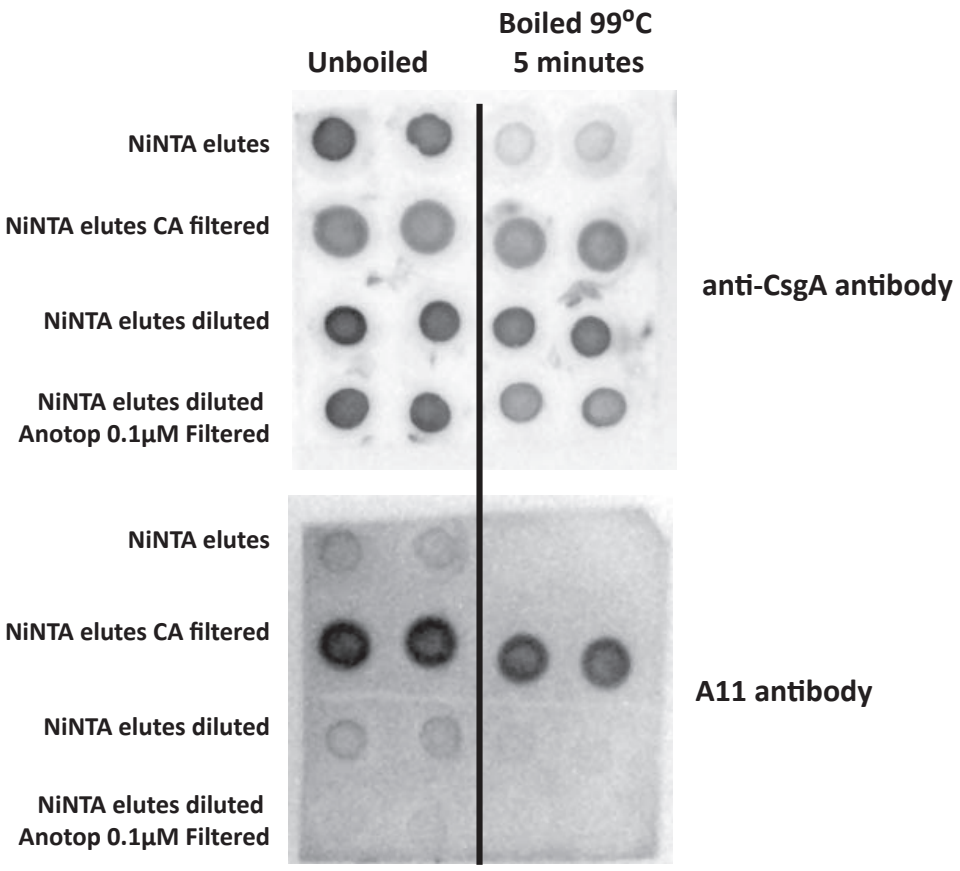


Figure 4.20. Effects of filtering CsgA-His in GdHCl through a 0.22µm Cellulose Acetate filter and an Anotop 0.1 µm γ-Alumina based filter on A11 antibody signal.

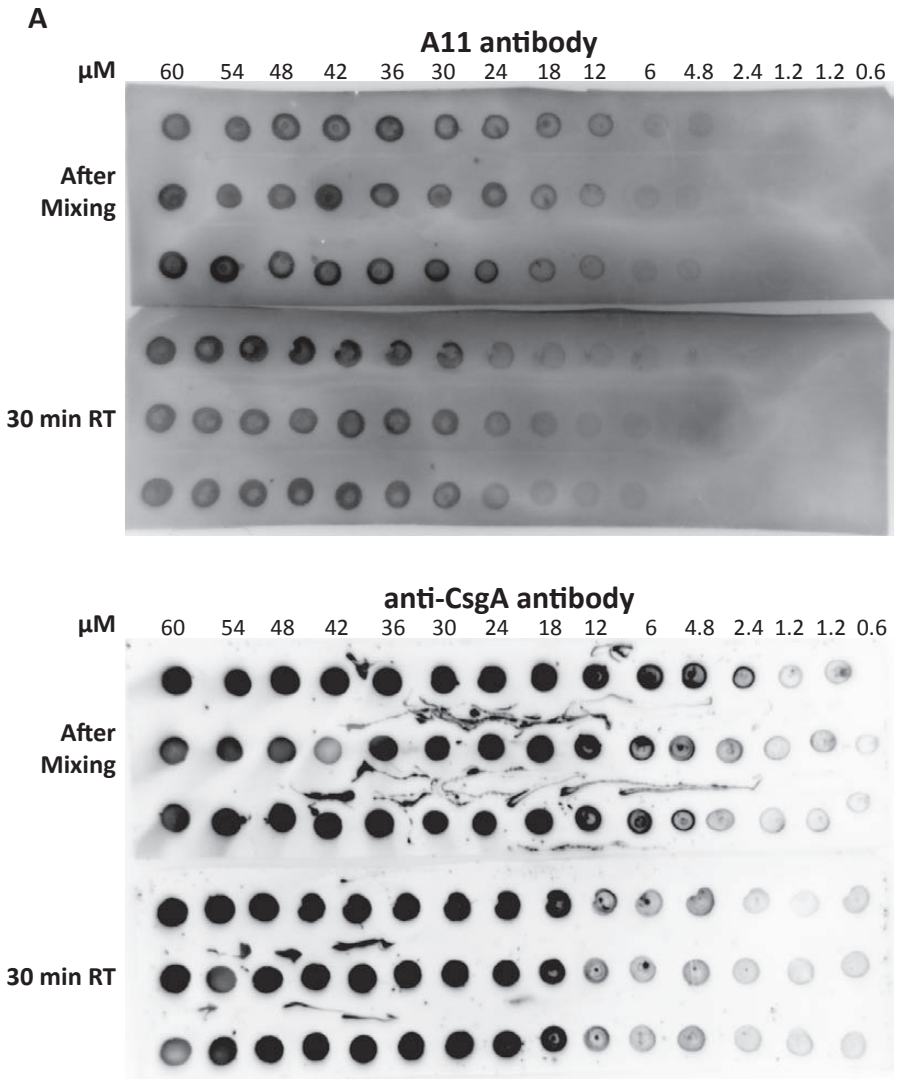


Figure 4.21. The A11 oligomer antibody does not bind CsgA-His below 4µM.

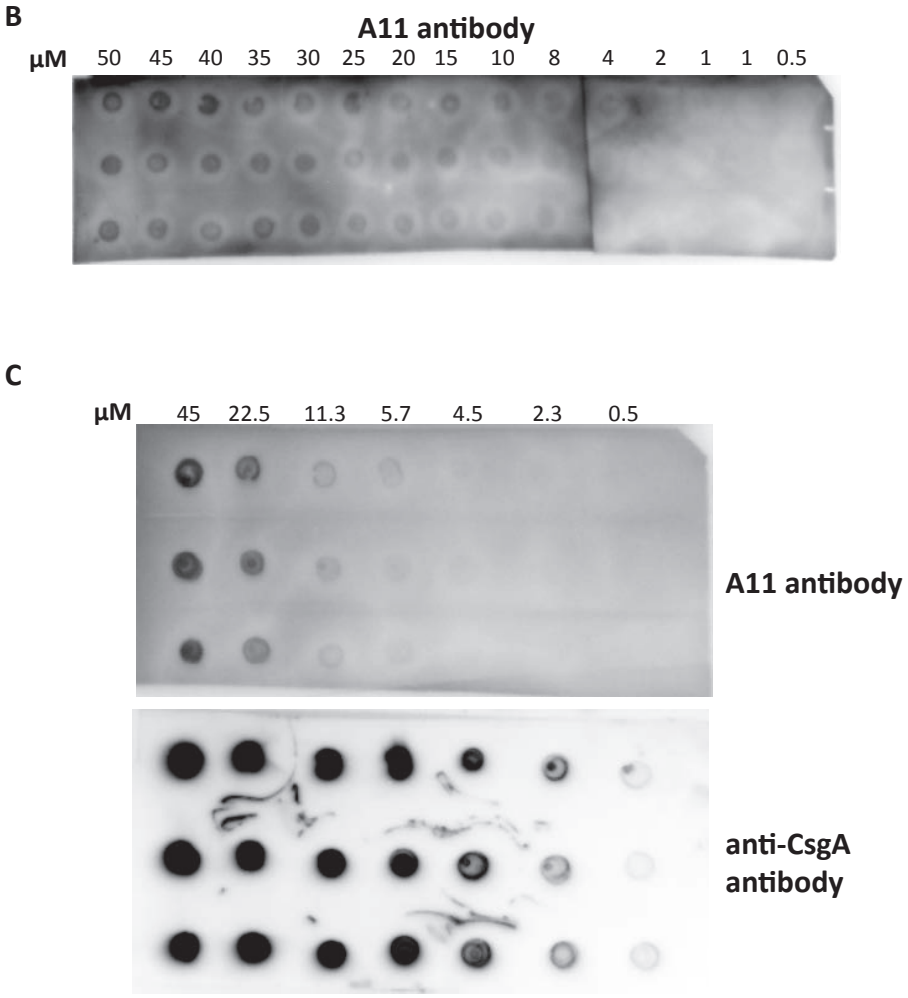


Figure 4.21 (continued). The A11 oligomer antibody does not bind CsgA-His below 4µM.

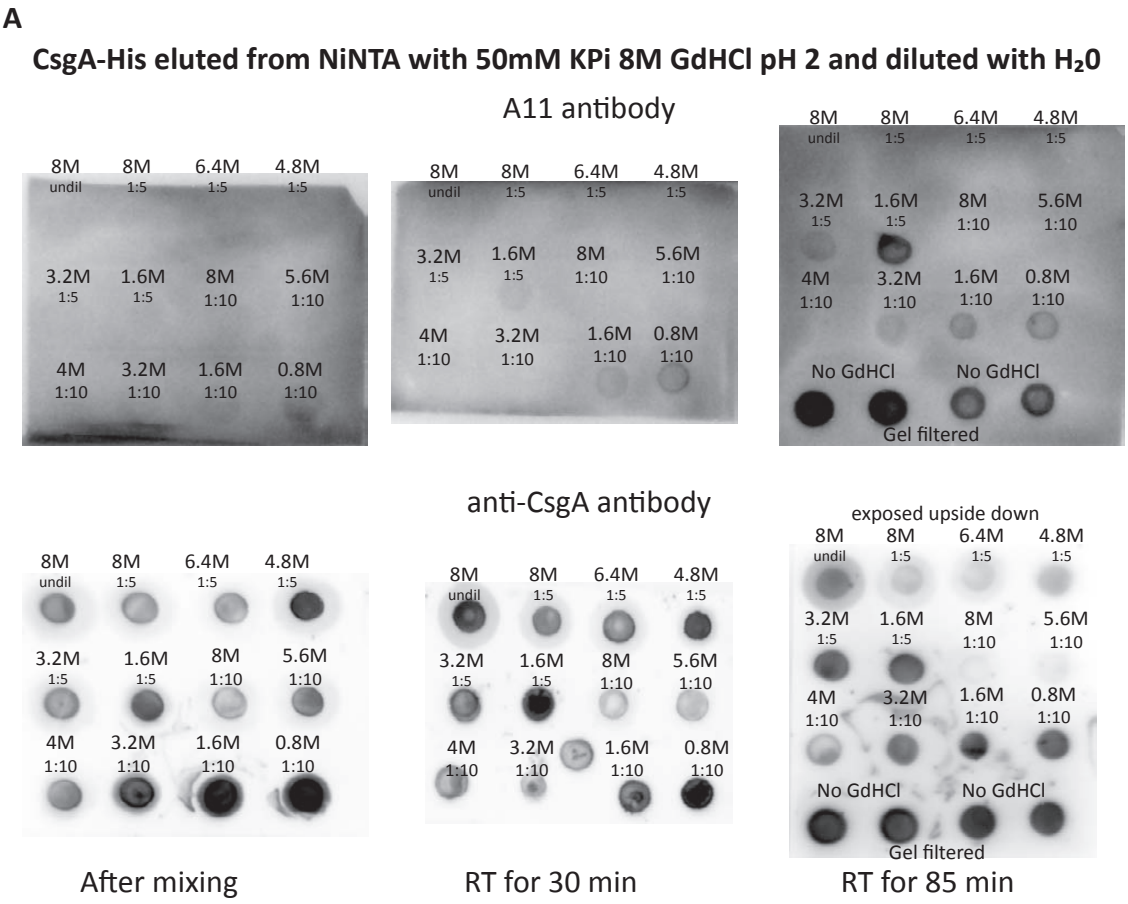
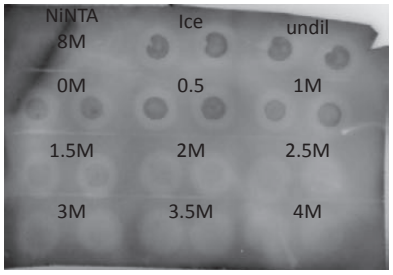


Figure 4.22. Effects of different GdHCl concentrations on A11 antibody binding to CsgA-His.

B

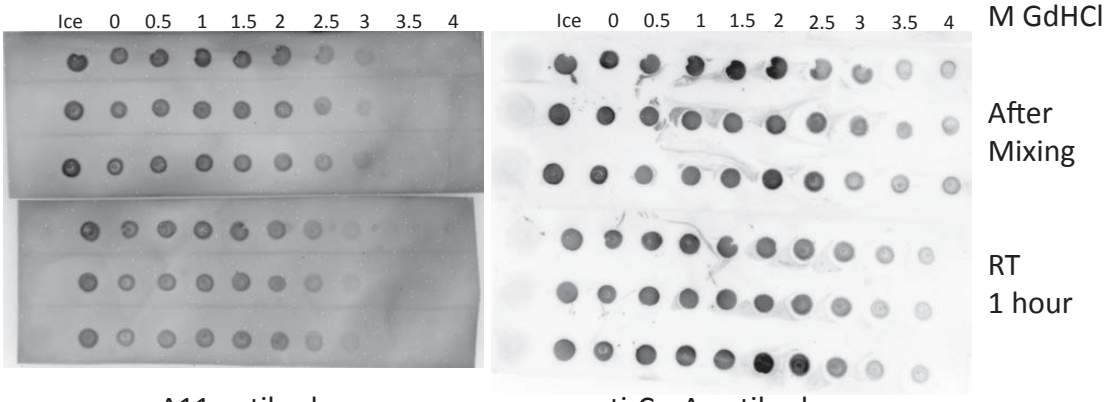
25µM CsgA-His with GdHCl added



A11 antibody

C

35µM CsgA-His with GdHCl added



A11 antibody anti-CsgA antibody

Figure 4.22 (continued). Effects of different GdHCl concentrations on A11 antibody binding to CsgA-His.

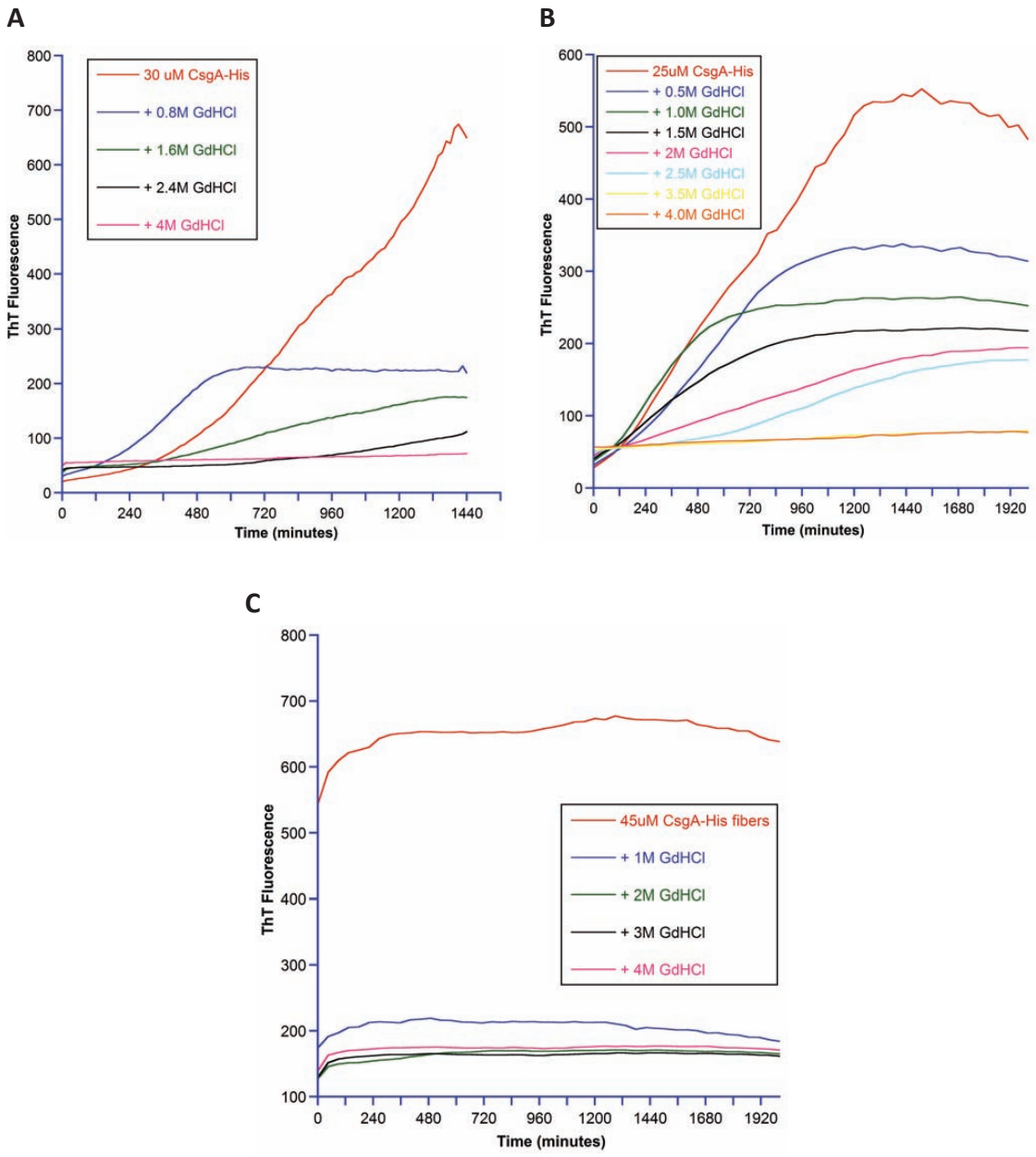


Figure 4.23. Effects of GdHCl on CsgA-His polymerization.

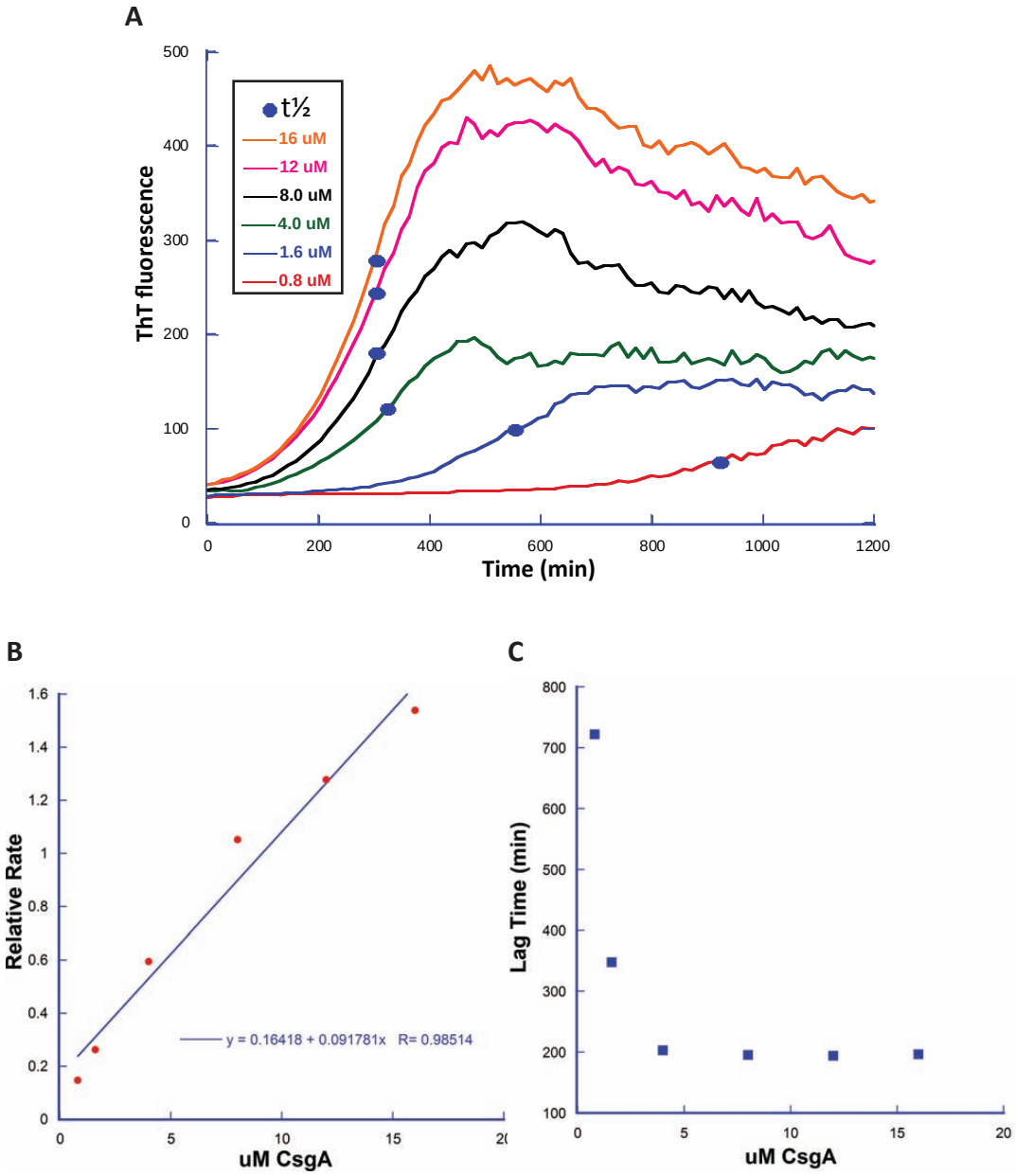


Figure 4.24. Effects of CsgA-His concentration on the half time (t½), relative rate, and lag phase of polymerization.

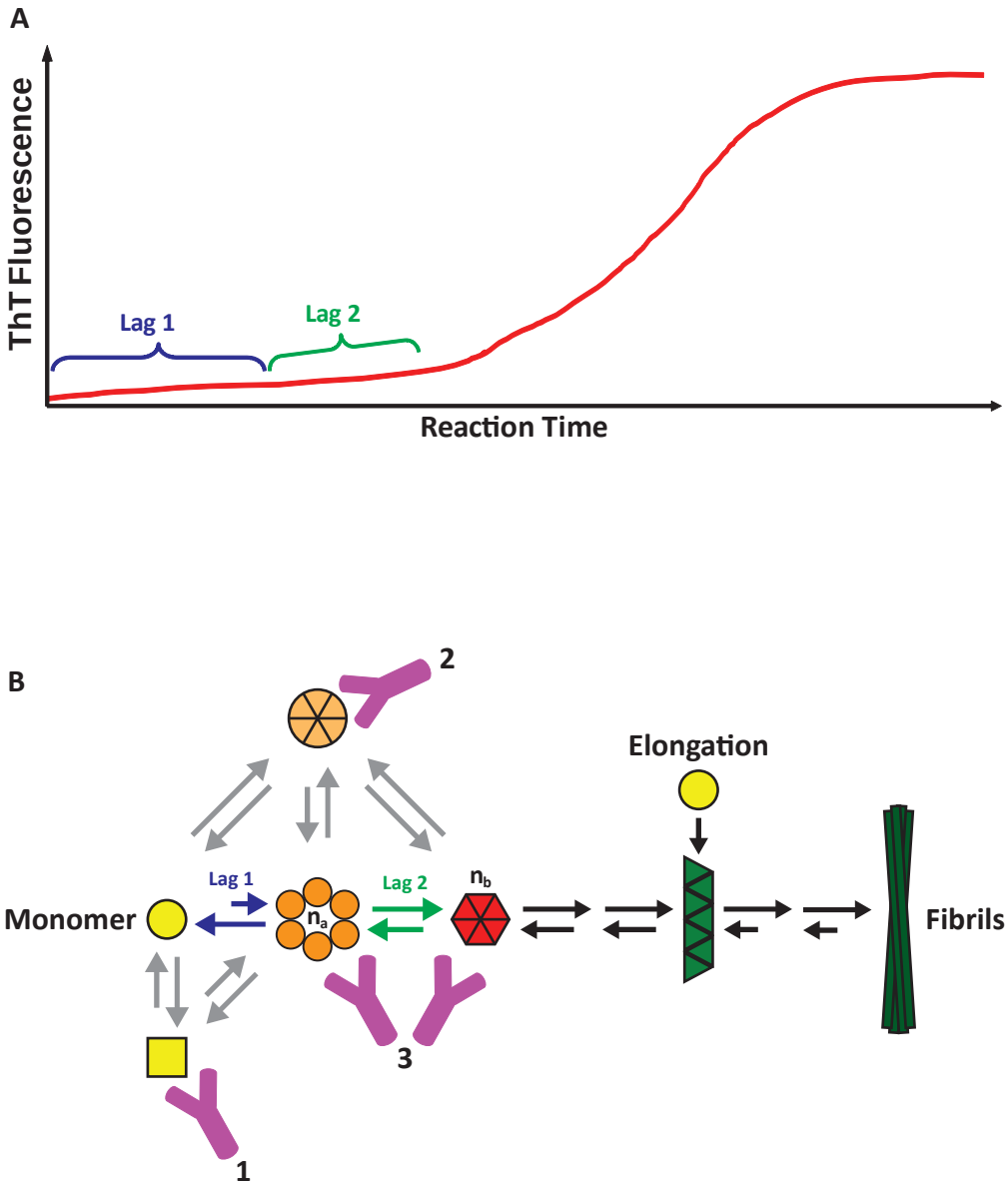


Figure 4.25. Proposed mechanism of CsgA-His polymerization and A11 Oligomer antibody binding.

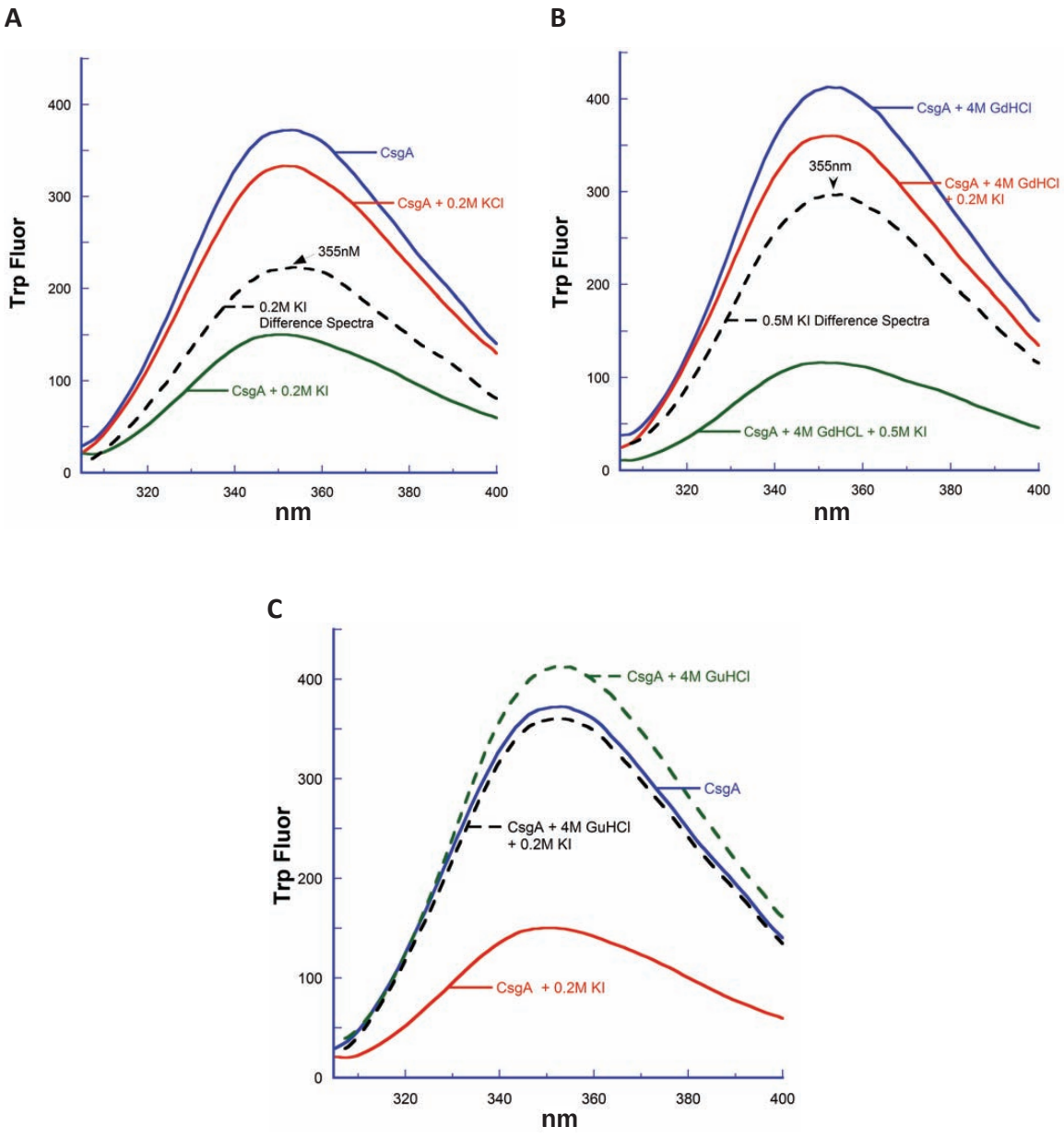


Figure 4.26. Tryptophan fluorescence spectra and iodine quenching of CsgA-His and CsgA-His fibers.

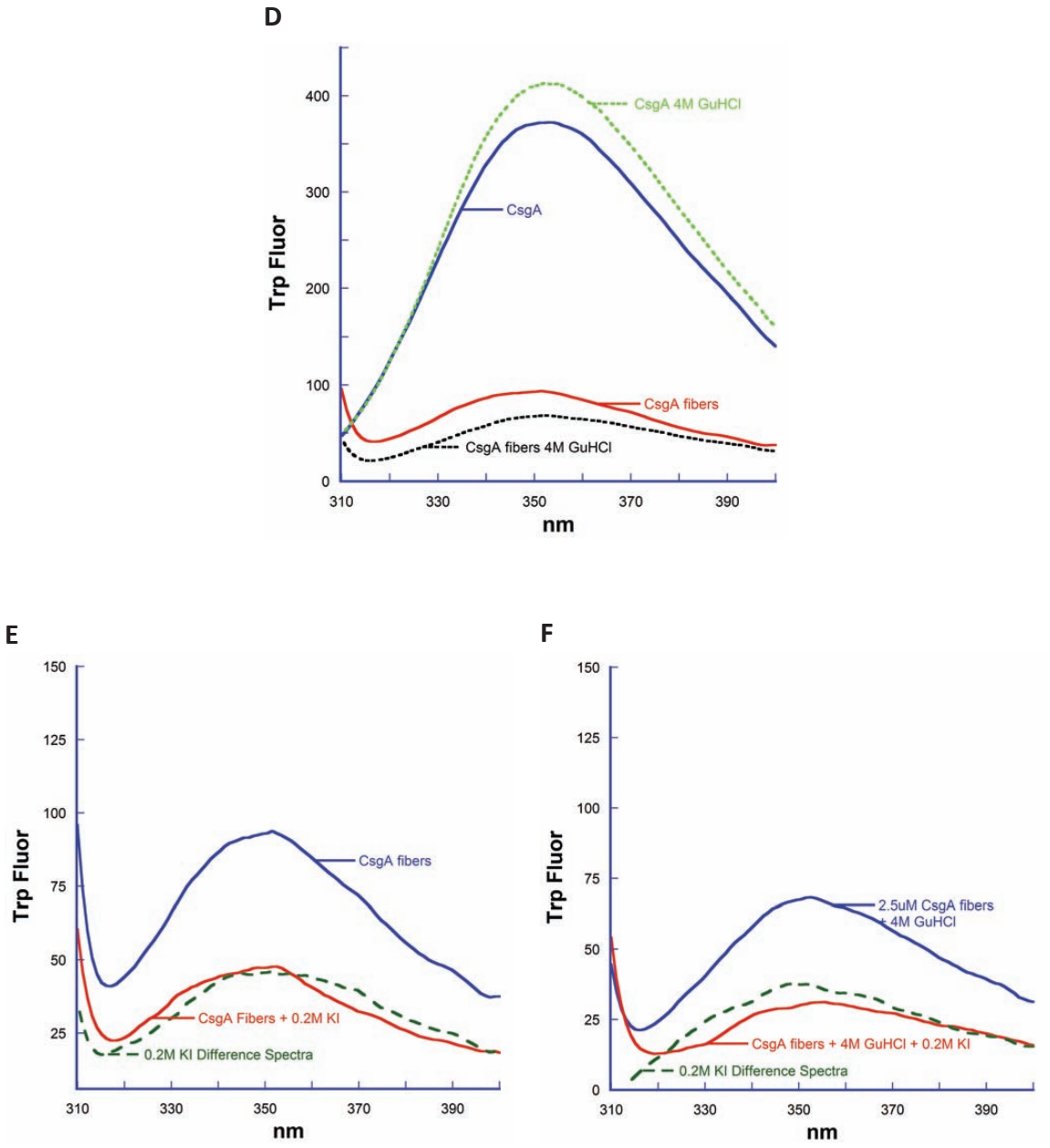
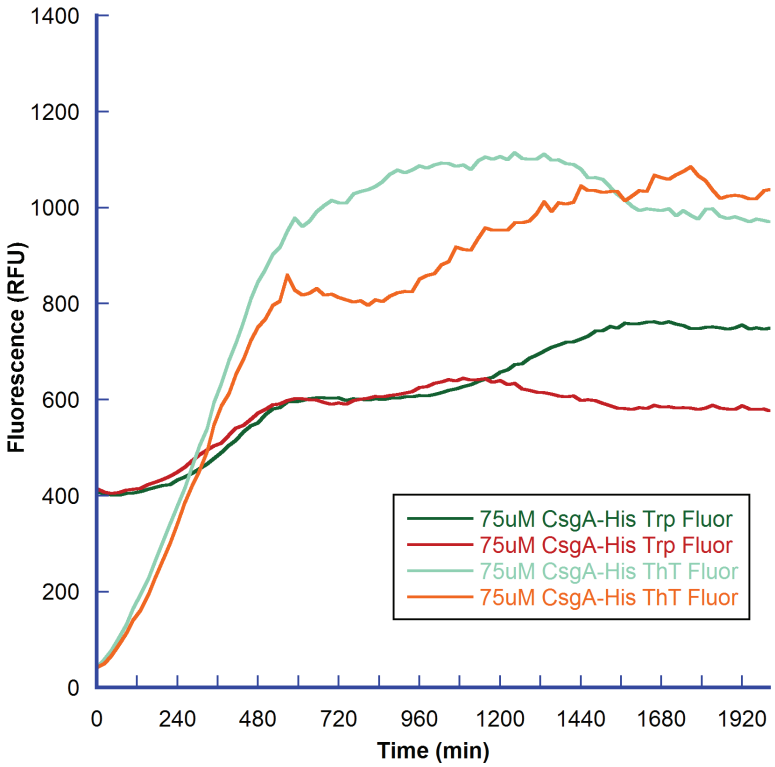


Figure 4.26 (continued). Tryptophan fluorescence spectra and iodine quenching of CsgA-His and CsgA-His fibers.



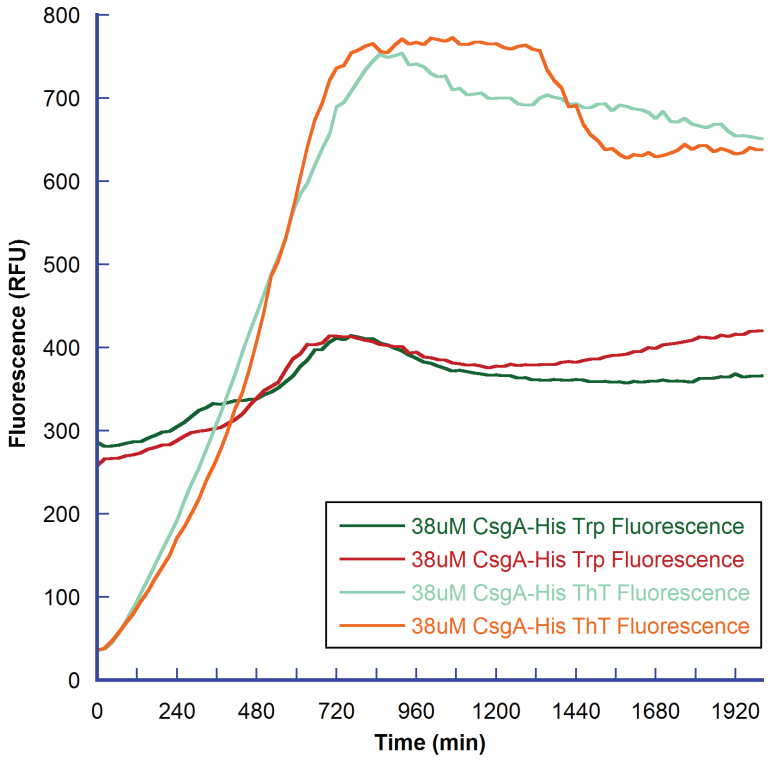
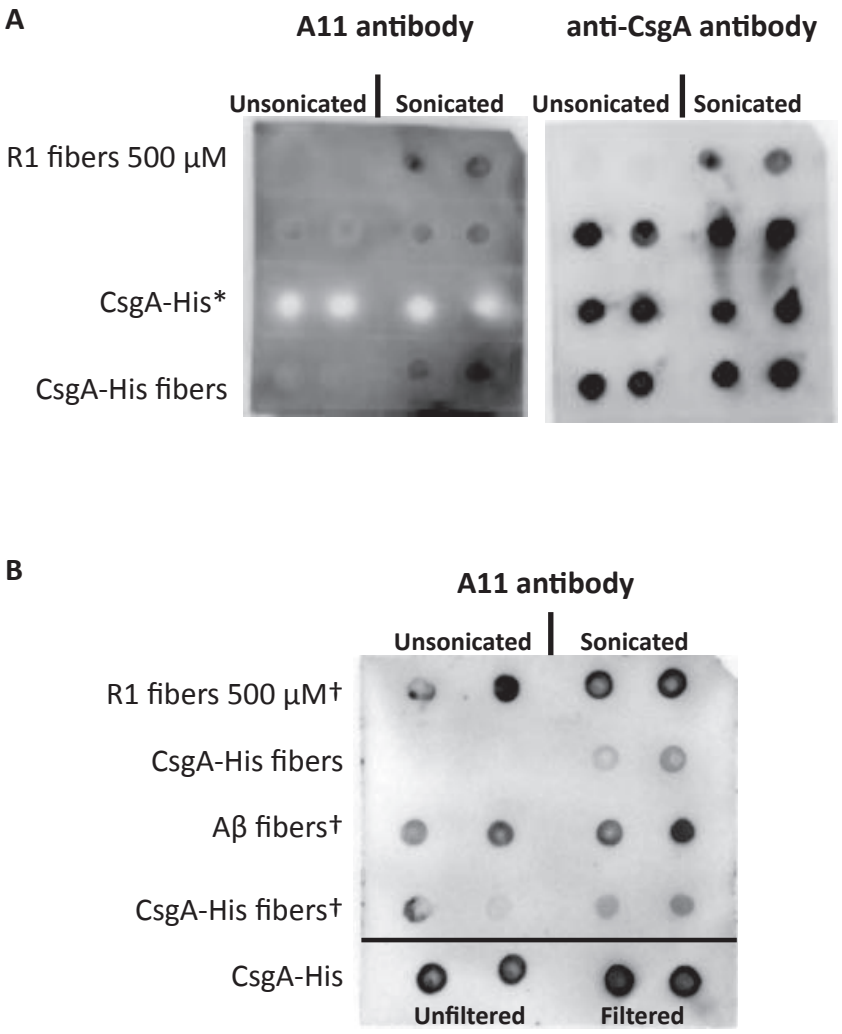


Figure 4.27. CsgA-His polymerization timecourse assayed with tryptophan and ThT fluorescence in the Molecular Devices Spectramax M2 plate reader.



* Unpolyerized CsgA-His was rapidly concentrated in Amicon Ultra filter.

† Previously sonicated.

Figure 4.28. Sonciation of CsgA-His fibers results in A11 oligomer antibody binding.

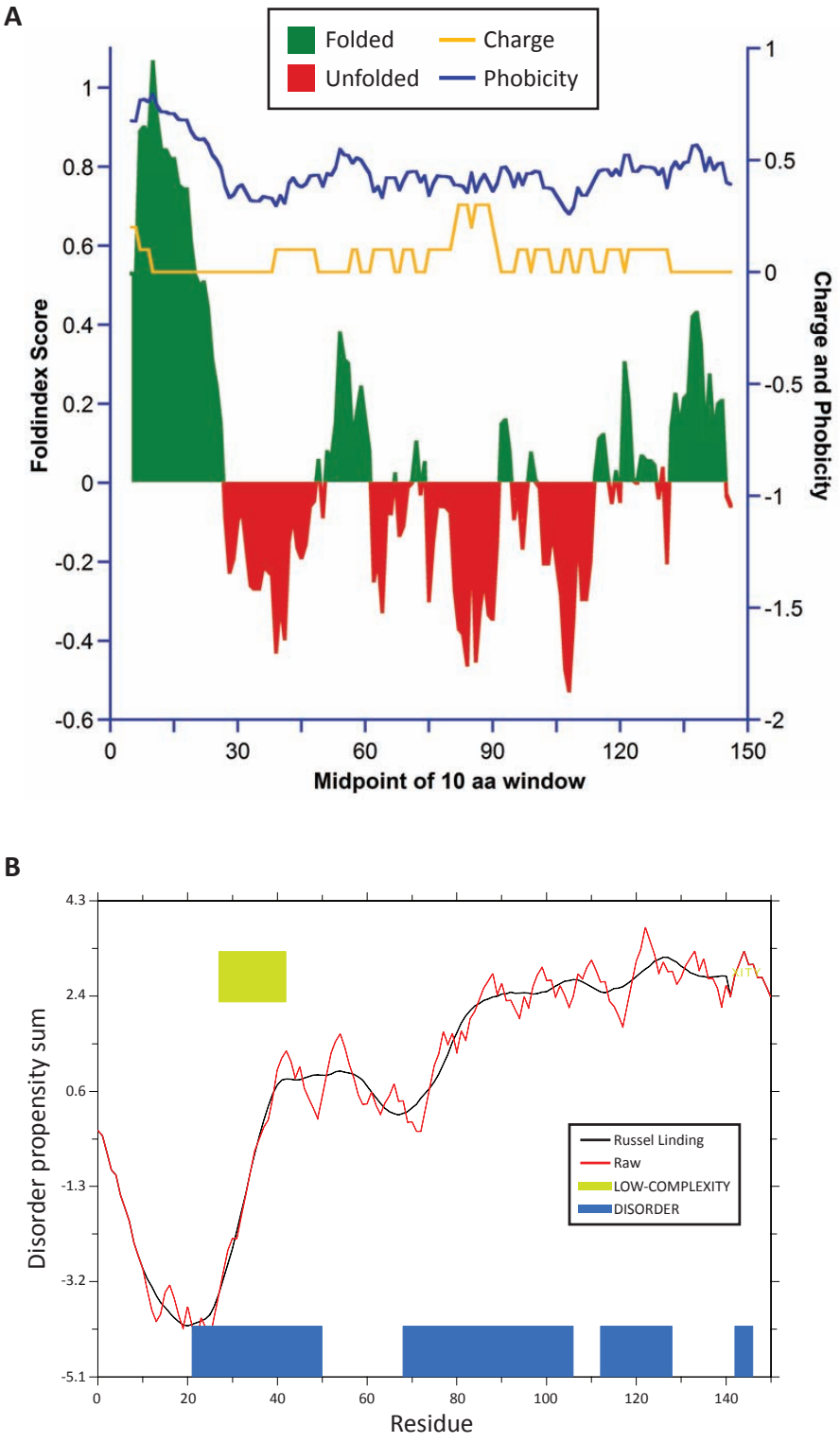


Figure 4.29. CsgA is predicted to be a natively unfolded protein.

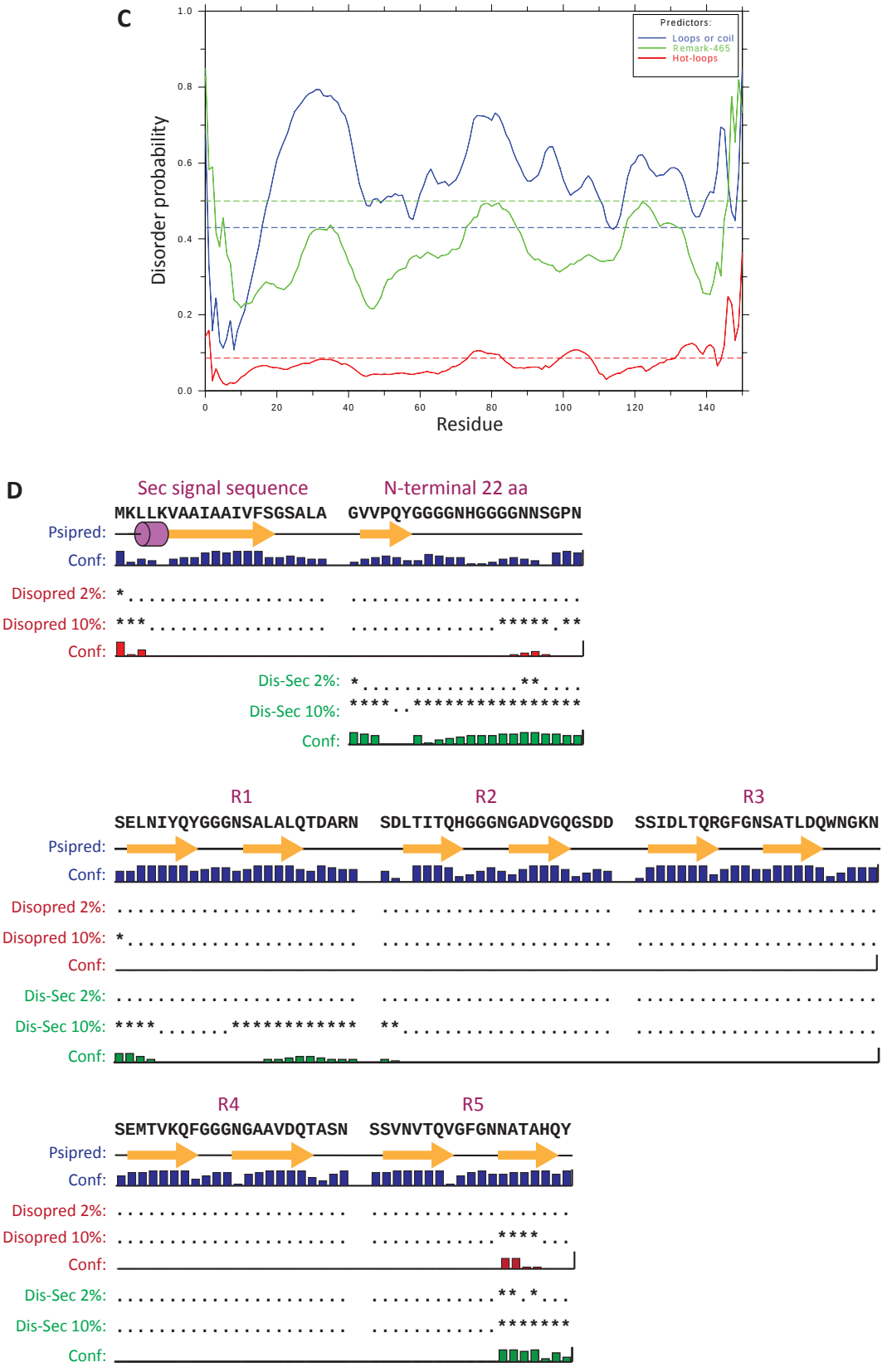


Figure 4.29 (continued). CsgA is predicted to be a natively unfolded protein.

CsgA

		Protéine	Protparam
SEC	MKLLKVAAIAAIVFSGSALA	10.50	10.00
N22	GVVPQYGGGGNHGGGGNNSGPN	7.84	6.74
R1	SELNIYQYGGGNSALALQTDARN	4.05	4.37
R2	SDLTITQHGGGNGADVGQGSDD	3.47	3.96
R3	SSIDLTQRGFGNSATLDQWNGKN	6.93	5.68
R4	SEMTVKQFGGGNGAAVDQTASN	4.05	4.37
R5	SSVNVTQVGFGNNATAHQY	7.84	6.46

CsgB

		Protéine	Protparam
SEC	MKNKLLFMMLTILGAPGIAAA	10.50	10.00
N23	AGYDLANSEYNFAVNELSKSSFN	3.90	4.14
R1	QAAIIGQAGTNNSAQLRQGGSK	11.45	11.00
R2	LLAVVAQEGSSNRAKIDQTGDY	4.21	4.56
R3	NLAYIDQAGSANDASISQGAYG	2.95	3.56
R4	NTAMIIQKGSGNKANITQYGTQ	10.10	9.70
R5	KTAIVVQRQSQMAIRVTQR	12.64	12.30

Figure 4.30. Theoretical Isoelectic points (pI) of Repeating Units of CsgA and CsgB.

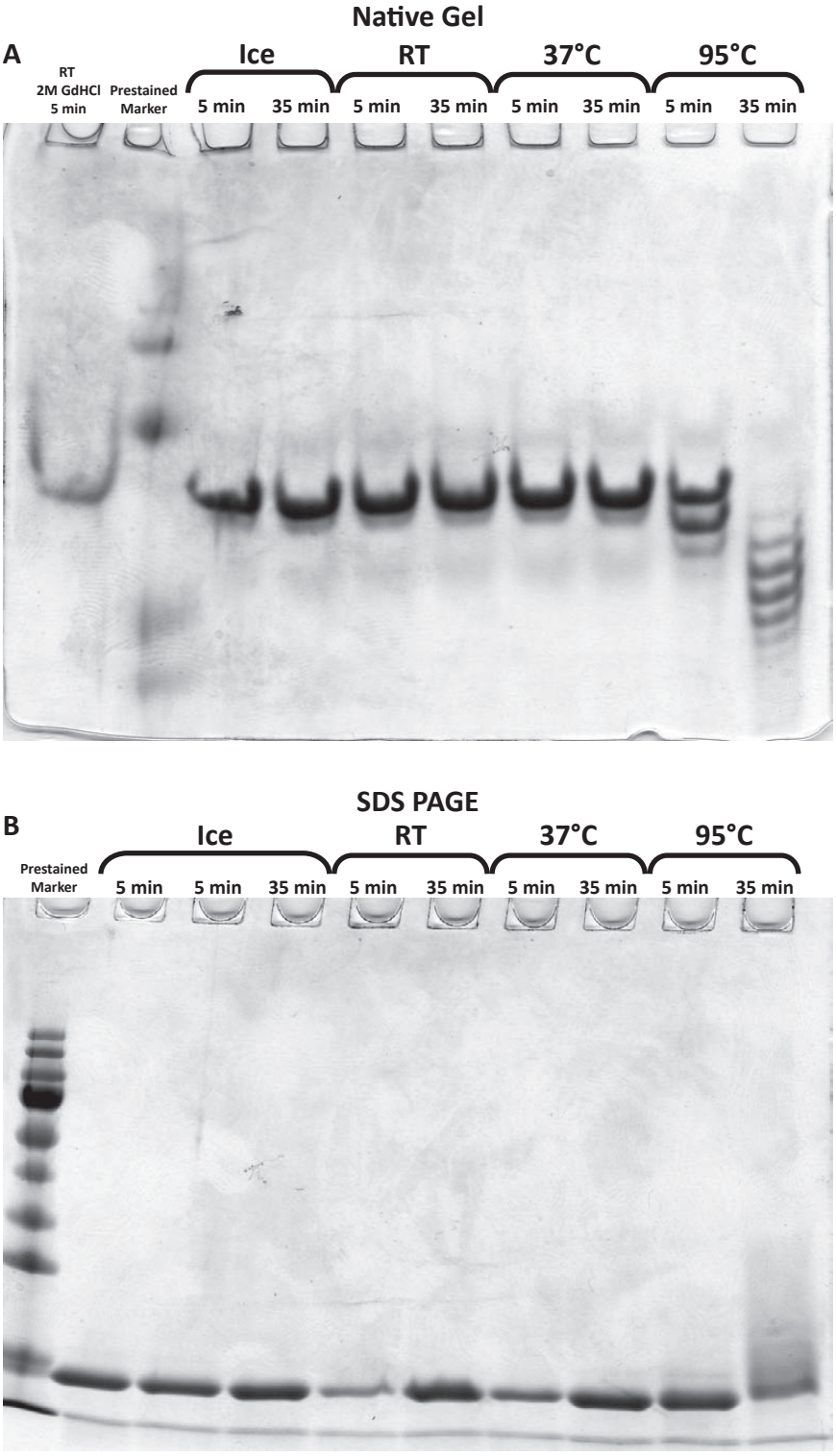


Figure 4.31. The effects of temperature and GdHCl on the mobility of CsgA-His in Native or SDS-PAGE gels and on A11 Oligomer antibody binding.

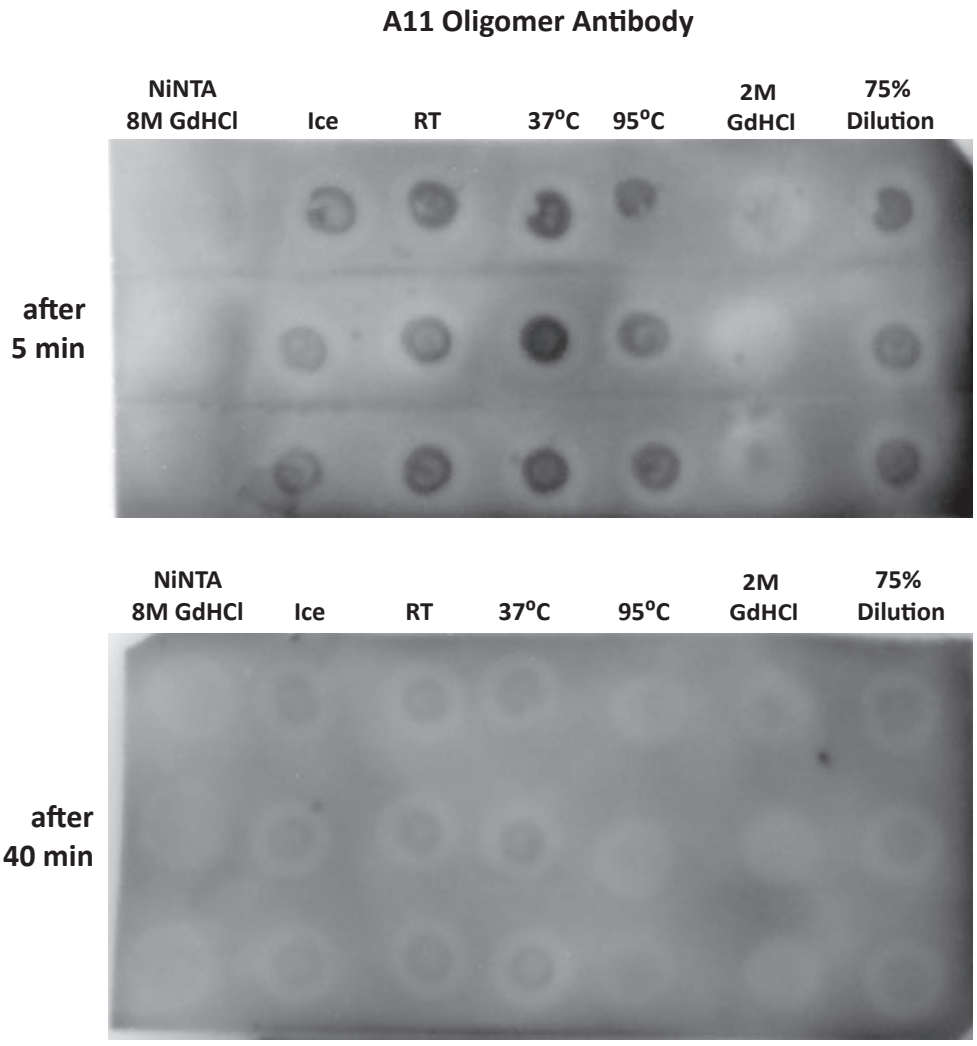


Figure 4.31 (continued). The effects of temperature and GdHCl on the mobility of CsgA-His in Native or SDS-PAGE gels and on A11 Oligomer antibody binding.

References

1. **Alteri, C. J., J. Xicohtencatl-Cortes, S. Hess, G. Caballero-Olin, J. A. Giron, and R. L. Friedman.** 2007. Mycobacterium tuberculosis produces pili during human infection. Proc Natl Acad Sci U S A **104:**5145-50.
2. **Austin, J. W., G. Sanders, W. W. Kay, and S. K. Collinson.** 1998. Thin aggregative fimbriae enhance Salmonella enteritidis biofilm formation. FEMS Microbiol Lett **162:**295-301.
3. **Badtke, M. P., N. D. Hammer, and M. R. Chapman.** 2009. Functional amyloids signal their arrival. Sci Signal **2:**pe43.
4. **Bagriantsev, S. N., V. V. Kushnirov, and S. W. Liebman.** 2006. Analysis of amyloid aggregates using agarose gel electrophoresis. Methods Enzymol **412:**33-48.
5. **Barnhart, M. M., and M. R. Chapman.** 2006. Curli biogenesis and function. Annu Rev Microbiol **60:**131-47.
6. **Barnhart, M. M., J. Lynem, and M. R. Chapman.** 2006. GlcNAc-6P levels modulate the expression of Curli fibers by Escherichia coli. J Bacteriol **188:**5212-9.
7. **Barrow, C. J., A. Yasuda, P. T. Kenny, and M. G. Zagorski.** 1992. Solution conformations and aggregational properties of synthetic amyloid beta-peptides of Alzheimer's disease. Analysis of circular dichroism spectra. J Mol Biol **225:**1075-93.
8. **Barrow, C. J., and M. G. Zagorski.** 1991. Solution structures of beta peptide and its constituent fragments: relation to amyloid deposition. Science **253:**179-82.
9. **Bhattacharyya, A. M., A. K. Thakur, and R. Wetzel.** 2005. polyglutamine aggregation nucleation: thermodynamics of a highly unfavorable protein folding reaction. Proc Natl Acad Sci U S A **102:**15400-5.
10. **Bian, Z., A. Brauner, Y. Li, and S. Normark.** 2000. Expression of and cytokine activation by Escherichia coli curli fibers in human sepsis. J Infect Dis **181:**602-12.
11. **Bian, Z., Z. Q. Yan, G. K. Hansson, P. Thoren, and S. Normark.** 2001. Activation of inducible nitric oxide synthase/nitric oxide by curli fibers leads to a fall in blood pressure during systemic Escherichia coli infection in mice. J Infect Dis **183:**612-9.
12. **Bieler, S., L. Estrada, R. Lagos, M. Baeza, J. Castilla, and C. Soto.** 2005. Amyloid formation modulates the biological activity of a bacterial protein. J Biol Chem **280:**26880-5.
13. **Bokranz, W., X. Wang, H. Tschape, and U. Romling.** 2005. Expression of cellulose and curli fimbriae by Escherichia coli isolated from the gastrointestinal tract. J Med Microbiol **54:**1171-82.
14. **Booth, D. R., M. Sunde, V. Bellotti, C. V. Robinson, W. L. Hutchinson, P. E. Fraser, P. N. Hawkins, C. M. Dobson, S. E. Radford, C. C. Blake, and M. B. Pepys.** 1997. Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. Nature **385:**787-93.

15. **Brender, J. R., K. Hartman, R. P. Nanga, N. Popovych, R. de la Salud Bea, S. Vivekanandan, E. N. Marsh, and A. Ramamoorthy.** 2010. Role of zinc in human islet amyloid polypeptide aggregation. J Am Chem Soc **132:**8973-83.
16. **Bucciantini, M., E. Giannoni, F. Chiti, F. Baroni, L. Formigli, J. Zurdo, N. Taddei, G. Ramponi, C. M. Dobson, and M. Stefani.** 2002. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. Nature **416:**507-11.
17. **Calamai, M., J. R. Kumita, J. Mifsud, C. Parrini, M. Ramazzotti, G. Ramponi, N. Taddei, F. Chiti, and C. M. Dobson.** 2006. Nature and significance of the interactions between amyloid fibrils and biological polyelectrolytes. Biochemistry **45:**12806-15.
18. **Calamai, M., N. Taddei, M. Stefani, G. Ramponi, and F. Chiti.** 2003. Relative influence of hydrophobicity and net charge in the aggregation of two homologous proteins. Biochemistry **42:**15078-83.
19. **Chapman, M. R., L. S. Robinson, J. S. Pinkner, R. Roth, J. Heuser, M. Hammar, S. Normark, and S. J. Hultgren.** 2002. Role of Escherichia coli curli operons in directing amyloid fiber formation. Science **295:**851-5.
20. **Chen, S., and R. Wetzel.** 2001. Solubilization and disaggregation of polyglutamine peptides. Protein Sci **10:**887-91.
21. **Chen, Y., and M. D. Barkley.** 1998. Toward understanding tryptophan fluorescence in proteins. Biochemistry **37:**9976-82.
22. **Cherny, I., L. Rockah, O. Levy-Nissenbaum, U. Gophna, E. Z. Ron, and E. Gazit.** 2005. The formation of Escherichia coli curli amyloid fibrils is mediated by prion-like peptide repeats. J Mol Biol **352:**245-52.
23. **Chiti, F., M. Bucciantini, C. Capanni, N. Taddei, C. M. Dobson, and M. Stefani.** 2001. Solution conditions can promote formation of either amyloid protofilaments or mature fibrils from the HypF N-terminal domain. Protein Sci **10:**2541-7.
24. **Chiti, F., M. Calamai, N. Taddei, M. Stefani, G. Ramponi, and C. M. Dobson.** 2002. Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. Proc Natl Acad Sci U S A **99 Suppl 4:**16419-26.
25. **Chiti, F., and C. M. Dobson.** 2006. Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem **75:**333-66.
26. **Claessen, D., W. de Jong, L. Dijkhuizen, and H. A. Wosten.** 2006. Regulation of Streptomyces development: reach for the sky! Trends Microbiol **14:**313-9.
27. **Claessen, D., R. Rink, W. de Jong, J. Siebring, P. de Vreugd, F. G. Boersma, L. Dijkhuizen, and H. A. Wosten.** 2003. A novel class of secreted hydrophobic proteins is involved in aerial hyphae formation in Streptomyces coelicolor by forming amyloid-like fibrils. Genes Dev **17:**1714-26.
28. **Collinson, S. K., J. M. Parker, R. S. Hodges, and W. W. Kay.** 1999. Structural predictions of AgfA, the insoluble fimbrial subunit of Salmonella thin aggregative fimbriae. J Mol Biol **290:**741-56.

29. **Coustou-Linares, V., M. L. Maddelein, J. Begueret, and S. J. Saupe.** 2001. In vivo aggregation of the HET-s prion protein of the fungus Podospora anserina. Mol Microbiol **42:**1325-35.
30. **Coustou, V., C. Deleu, S. Saupe, and J. Begueret.** 1997. The protein product of the het-s heterokaryon incompatibility gene of the fungus Podospora anserina behaves as a prion analog. Proc Natl Acad Sci U S A **94:**9773-8.
31. **Coustou, V., C. Deleu, S. J. Saupe, and J. Begueret.** 1999. Mutational analysis of the [Het-s] prion analog of Podospora anserina. A short N-terminal peptide allows prion propagation. Genetics **153:**1629-40.
32. **Demuro, A., E. Mina, R. Kayed, S. C. Milton, I. Parker, and C. G. Glabe.** 2005. Calcium dysregulation and membrane disruption as a ubiquitous neurotoxic mechanism of soluble amyloid oligomers. J Biol Chem **280:**17294-300.
33. **Elliot, M. A., N. Karoonuthaisiri, J. Huang, M. J. Bibb, S. N. Cohen, C. M. Kao, and M. J. Buttner.** 2003. The chaplins: a family of hydrophobic cell-surface proteins involved in aerial mycelium formation in Streptomyces coelicolor. Genes Dev **17:**1727-40.
34. **Epstein, E. A., and M. R. Chapman.** 2008. Polymerizing the fibre between bacteria and host cells: the biogenesis of functional amyloid fibres. Cell Microbiol **10:**1413-20.
35. **Fancy, D. A., and T. Kodadek.** 1999. Chemistry for the analysis of protein-protein interactions: rapid and efficient cross-linking triggered by long wavelength light. Proc Natl Acad Sci U S A **96:**6020-4.
36. **Fandrich, M., M. A. Fletcher, and C. M. Dobson.** 2001. Amyloid fibrils from muscle myoglobin. Nature **410:**165-6.
37. **Fowler, D. M., A. V. Koulov, C. Alory-Jost, M. S. Marks, W. E. Balch, and J. W. Kelly.** 2006. Functional amyloid formation within mammalian tissue. PLoS Biol **4:**e6.
38. **Fowler, D. M., A. V. Koulov, W. E. Balch, and J. W. Kelly.** 2007. Functional amyloid--from bacteria to humans. Trends Biochem Sci **32:**217-24.
39. **Galzitskaya, O. V., and S. O. Garbuzynskiy.** 2006. Entropy capacity determines protein folding. Proteins **63:**144-54.
40. **Galzitskaya, O. V., S. O. Garbuzynskiy, and M. Y. Lobanov.** 2006. FoldUnfold: web server for the prediction of disordered regions in protein chain. Bioinformatics **22:**2948-9.
41. **Gebbink, M. F., D. Claessen, B. Bouma, L. Dijkhuizen, and H. A. Wosten.** 2005. Amyloids--a functional coat for microorganisms. Nat Rev Microbiol **3:**333-41.
42. **Glabe, C. G., and R. Kayed.** 2006. Common structure and toxic function of amyloid oligomers implies a common mechanism of pathogenesis. Neurology **66:**S74-8.
43. **Glover, J. R., A. S. Kowal, E. C. Schirmer, M. M. Patino, J. J. Liu, and S. Lindquist.** 1997. Self-seeded fibers formed by Sup35, the protein determinant of [PSI+], a heritable prion-like factor of S. cerevisiae. Cell **89:**811-9.

44. **Gophna, U., M. Barlev, R. Seijffers, T. A. Oelschlager, J. Hacker, and E. Z. Ron.** 2001. Curli fibers mediate internalization of Escherichia coli by eukaryotic cells. Infect Immun **69:**2659-65.
45. **Gophna, U., T. A. Oelschlaeger, J. Hacker, and E. Z. Ron.** 2002. Role of fibronectin in curli-mediated internalization. FEMS Microbiol Lett **212:**55-8.
46. **Gualdi, L., L. Tagliabue, S. Bertagnoli, T. Ierano, C. De Castro, and P. Landini.** 2008. Cellulose modulates biofilm formation by counteracting curli-mediated colonization of solid surfaces in Escherichia coli. Microbiology **154:**2017-24.
47. **Guijarro, J. I., M. Sunde, J. A. Jones, I. D. Campbell, and C. M. Dobson.** 1998. Amyloid fibril formation by an SH3 domain. Proc Natl Acad Sci U S A **95:**4224-8.
48. **Hammar, M., Z. Bian, and S. Normark.** 1996. Nucleator-dependent intercellular assembly of adhesive curli organelles in Escherichia coli. Proc Natl Acad Sci U S A **93:**6562-6.
49. **Hammer, N. D., J. C. Schmidt, and M. R. Chapman.** 2007. The curli nucleator protein, CsgB, contains an amyloidogenic domain that directs CsgA polymerization. Proc Natl Acad Sci U S A **104:**12494-9.
50. **Hammer, N. D., X. Wang, B. A. McGuffie, and M. R. Chapman.** 2008. Amyloids: friend or foe? J Alzheimers Dis **13:**407-19.
51. **Hardy, J., and D. J. Selkoe.** 2002. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science **297:**353-6.
52. **Hurshman, A. R., J. T. White, E. T. Powers, and J. W. Kelly.** 2004. Transthyretin aggregation under partially denaturing conditions is a downhill polymerization. Biochemistry **43:**7365-81.
53. **Johansson, C., T. Nilsson, A. Olsen, and M. J. Wick.** 2001. The influence of curli, a MHC-I-binding bacterial surface structure, on macrophage-T cell interactions. FEMS Immunol Med Microbiol **30:**21-9.
54. **Jonas, K., H. Tomenius, A. Kader, S. Normark, U. Romling, L. M. Belova, and O. Melefors.** 2007. Roles of curli, cellulose and BapA in Salmonella biofilm morphology studied by atomic force microscopy. BMC Microbiol **7:**70.
55. **Jones, A. F., and J. Lunec.** 1987. Protein fluorescence and its relationship to free radical activity. Br J Cancer Suppl **8:**60-5.
56. **Kameoka, D., E. Masuzaki, T. Ueda, and T. Imoto.** 2007. Effect of buffer species on the unfolding and the aggregation of humanized IgG. J Biochem **142:**383-91.
57. **Kayed, R., E. Head, J. L. Thompson, T. M. McIntire, S. C. Milton, C. W. Cotman, and C. G. Glabe.** 2003. Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. Science **300:**486-9.
58. **Khurana, R., C. Coleman, C. Ionescu-Zanetti, S. A. Carter, V. Krishna, R. K. Grover, R. Roy, and S. Singh.** 2005. Mechanism of thioflavin T binding to amyloid fibrils. J Struct Biol **151:**229-38.
59. **Kisilevsky, R.** 2000. Review: amyloidogenesis-unquestioned answers and unanswered questions. J Struct Biol **130:**99-108.
60. **Lansbury, P. T., Jr.** 1999. Evolution of amyloid: what normal protein folding may tell us about fibrillogenesis and disease. Proc Natl Acad Sci U S A **96:**3342-4.

61. **Lashuel, H. A., D. Hartley, B. M. Petre, T. Walz, and P. T. Lansbury, Jr.** 2002. Neurodegenerative disease: amyloid pores from pathogenic mutations. Nature **418:**291.
62. **Leonil, J., G. Henry, D. Jouanneau, M. M. Delage, V. Forge, and J. L. Putaux.** 2008. Kinetics of fibril formation of bovine kappa-casein indicate a conformational rearrangement as a critical step in the process. J Mol Biol **381:**1267-80.
63. **Lesne, S., M. T. Koh, L. Kotilinek, R. Kayed, C. G. Glabe, A. Yang, M. Gallagher, and K. H. Ashe.** 2006. A specific amyloid-beta protein assembly in the brain impairs memory. Nature **440:**352-7.
64. **LeVine, H., 3rd.** 1999. Quantification of beta-sheet amyloid fibril structures with thioflavin T. Methods Enzymol **309:**274-84.
65. **LeVine, H., 3rd.** 1993. Thioflavine T interaction with synthetic Alzheimer's disease beta-amyloid peptides: detection of amyloid aggregation in solution. Protein Sci **2:**404-10.
66. **Linding, R., L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell.** 2003. Protein disorder prediction: implications for structural proteomics. Structure **11:**1453-9.
67. **Linding, R., R. B. Russell, V. Neduva, and T. J. Gibson.** 2003. GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res **31:**3701-8.
68. **Linding, R., J. Schymkowitz, F. Rousseau, F. Diella, and L. Serrano.** 2004. A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. J Mol Biol **342:**345-53.
69. **Liu, J. J., and S. Lindquist.** 1999. Oligopeptide-repeat expansions modulate 'protein-only' inheritance in yeast. Nature **400:**573-6.
70. **Lomakin, A., D. S. Chung, G. B. Benedek, D. A. Kirschner, and D. B. Teplow.** 1996. On the nucleation and growth of amyloid beta-protein fibrils: detection of nuclei and quantitation of rate constants. Proc Natl Acad Sci U S A **93:**1125-9.
71. **Lomakin, A., D. B. Teplow, D. A. Kirschner, and G. B. Benedek.** 1997. Kinetic theory of fibrillogenesis of amyloid beta-protein. Proc Natl Acad Sci U S A **94:**7942-7.
72. **Lundmark, K., G. T. Westermark, A. Olsen, and P. Westermark.** 2005. Protein fibrils in nature can enhance amyloid protein A amyloidosis in mice: Cross-seeding as a disease mechanism. Proc Natl Acad Sci U S A **102:**6098-102.
73. **Maji, S. K., M. H. Perrin, M. R. Sawaya, S. Jessberger, K. Vadodaria, R. A. Rissman, P. S. Singru, K. P. Nilsson, R. Simon, D. Schubert, D. Eisenberg, J. Rivier, P. Sawchenko, W. Vale, and R. Riek.** 2009. Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. Science **325:**328-32.
74. **Maji, S. K., D. Schubert, C. Rivier, S. Lee, J. E. Rivier, and R. Riek.** 2008. Amyloid as a depot for the formulation of long-acting drugs. PLoS Biol **6:**e17.

75. **Malisauskas, M., A. Darinskas, V. V. Zamotin, A. Gharibyan, I. A. Kostanyan, and L. A. Morozova-Roche.** 2006. Intermediate amyloid oligomers of lysozyme: Is their cytotoxicity a particular case or general rule for amyloid? Biochemistry (Mosc) **71:**505-12.
76. **Martsev, S. P., A. P. Dubnovitsky, A. P. Vlasov, M. Hoshino, K. Hasegawa, H. Naiki, and Y. Goto.** 2002. Amyloid fibril formation of the mouse V(L) domain at acidic pH. Biochemistry **41:**3389-95.
77. **McGuffin, L. J., K. Bryson, and D. T. Jones.** 2000. The PSIPRED protein structure prediction server. Bioinformatics **16:**404-5.
78. **Munishkina, L. A., J. Henriques, V. N. Uversky, and A. L. Fink.** 2004. Role of protein-water interactions and electrostatics in alpha-synuclein fibril formation. Biochemistry **43:**3289-300.
79. **Nelson, R., and D. Eisenberg.** 2006. Recent atomic models of amyloid fibril structure. Curr Opin Struct Biol **16:**260-5.
80. **Nelson, R., M. R. Sawaya, M. Balbirnie, A. O. Madsen, C. Riekel, R. Grothe, and D. Eisenberg.** 2005. Structure of the cross-beta spine of amyloid-like fibrils. Nature **435:**773-8.
81. **Oh, J., J. G. Kim, E. Jeon, C. H. Yoo, J. S. Moon, S. Rhee, and I. Hwang.** 2007. Amyloidogenesis of type III-dependent harpins from plant pathogenic bacteria. J Biol Chem **282:**13601-9.
82. **Olsen, A., H. Herwald, M. Wikstrom, K. Persson, E. Mattsson, and L. Bjorck.** 2002. Identification of two protein-binding and functional regions of curli, a surface organelle and virulence determinant of Escherichia coli. J Biol Chem **277:**34568-72.
83. **Otoo, H. N., K. G. Lee, W. Qiu, and P. N. Lipke.** 2008. Candida albicans Als adhesins have conserved amyloid-forming sequences. Eukaryot Cell **7:**776-82.
84. **Padrick, S. B., and A. D. Miranker.** 2002. Islet amyloid: phase partitioning and secondary nucleation are central to the mechanism of fibrillogenesis. Biochemistry **41:**4694-703.
85. **Pepinsky, R. B.** 1991. Selective precipitation of proteins from guanidine hydrochloride-containing solutions with ethanol. Anal Biochem **195:**177-81.
86. **Powers, E. T., and D. L. Powers.** 2006. The kinetics of nucleated polymerizations at high concentrations: amyloid fibril formation near and above the "supercritical concentration". Biophys J **91:**122-32.
87. **Prilusky, J., C. E. Felder, T. Zeev-Ben-Mordehai, E. H. Rydberg, O. Man, J. S. Beckmann, I. Silman, and J. L. Sussman.** 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics **21:**3435-8.
88. **Proteine.** http://www.iut-arles.up.univ-mrs.fr/w3bb/d_abim/compo-p.html.
89. **ProtParam.** http://ca.expasy.org/cgi-bin/protparam.
90. **Prusiner, S. B.** 1998. The prion diseases. Brain Pathol **8:**499-513.

91. **Quist, A., I. Doudevski, H. Lin, R. Azimova, D. Ng, B. Frangione, B. Kagan, J. Ghiso, and R. Lal.** 2005. Amyloid ion channels: a common structural link for protein-misfolding disease. Proc Natl Acad Sci U S A **102:**10427-32.
92. **Raman, B., E. Chatani, M. Kihara, T. Ban, M. Sakai, K. Hasegawa, H. Naiki, M. Rao Ch, and Y. Goto.** 2005. Critical balance of electrostatic and hydrophobic interactions is required for beta 2-microglobulin amyloid fibril growth and stability. Biochemistry **44:**1288-99.
93. **Rhoades, E., J. Agarwal, and A. Gafni.** 2000. Aggregation of an amyloidogenic fragment of human islet amyloid polypeptide. Biochim Biophys Acta **1476:**230-8.
94. **Rhoades, E., and A. Gafni.** 2003. Micelle formation by a fragment of human islet amyloid polypeptide. Biophys J **84:**3480-7.
95. **Romero, D., C. Aguilar, R. Losick, and R. Kolter.** Amyloid fibers provide structural integrity to Bacillus subtilis biofilms. Proc Natl Acad Sci U S A **107:**2230-4.
96. **Ronga, L., P. Palladino, B. Tizzano, D. Marasco, E. Benedetti, R. Ragone, and F. Rossi.** 2006. Effect of salts on the structural behavior of hPrP alpha2-helix-derived analogues: the counterion perspective. J Pept Sci **12:**790-5.
97. **Ross, E. D., A. Minton, and R. B. Wickner.** 2005. Prion domains: sequences, structures and interactions. Nat Cell Biol **7:**1039-44.
98. **Sabate, R., and J. Estelrich.** 2005. Evidence of the existence of micelles in the fibrillogenesis of beta-amyloid peptide. J Phys Chem B **109:**11027-32.
99. **Saldana, Z., J. Xicohtencatl-Cortes, F. Avelino, A. D. Phillips, J. B. Kaper, J. L. Puente, and J. A. Giron.** 2009. Synergistic role of curli and cellulose in cell adherence and biofilm formation of attaching and effacing Escherichia coli and identification of Fis as a negative regulator of curli. Environ Microbiol **11:**992-1006.
100. **Sasaki, N., R. Fukatsu, K. Tsuzuki, Y. Hayashi, T. Yoshida, N. Fujii, T. Koike, I. Wakayama, R. Yanagihara, R. Garruto, N. Amano, and Z. Makita.** 1998. Advanced glycation end products in Alzheimer's disease and other neurodegenerative diseases. Am J Pathol **153:**1149-55.
101. **Scheibel, T., J. Bloom, and S. L. Lindquist.** 2004. The elongation of yeast prion fibers involves separable steps of association and conversion. Proc Natl Acad Sci U S A **101:**2287-92.
102. **Schmittschmitt, J. P., and J. M. Scholtz.** 2003. The role of protein stability, solubility, and net charge in amyloid fibril formation. Protein Sci **12:**2374-8.
103. **Serio, T. R., A. G. Cashikar, A. S. Kowal, G. J. Sawicki, J. J. Moslehi, L. Serpell, M. F. Arnsdorf, and S. L. Lindquist.** 2000. Nucleated conformational conversion and the replication of conformational information by a prion determinant. Science **289:**1317-21.
104. **Shorter, J., and S. Lindquist.** 2004. Hsp104 catalyzes formation and elimination of self-replicating Sup35 prion conformers. Science **304:**1793-7.
105. **Shorter, J., and S. Lindquist.** 2005. Prions as adaptive conduits of memory and inheritance. Nat Rev Genet **6:**435-50.

106. **Si, K., S. Lindquist, and E. R. Kandel.** 2003. A neuronal isoform of the aplysia CPEB has prion-like properties. Cell **115:**879-91.
107. **Sikkink, L. A., and M. Ramirez-Alvarado.** 2008. Salts enhance both protein stability and amyloid formation of an immunoglobulin light chain. Biophys Chem **135:**25-31.
108. **Solomon, E. B., B. A. Niemira, G. M. Sapers, and B. A. Annous.** 2005. Biofilm formation, cellulose production, and curli biosynthesis by Salmonella originating from produce, animal, and clinical sources. J Food Prot **68:**906-12.
109. **Soreghan, B., J. Kosmoski, and C. Glabe.** 1994. Surfactant properties of Alzheimer's A beta peptides and the mechanism of amyloid aggregation. J Biol Chem **269:**28551-4.
110. **Srinivasan, R., E. M. Jones, K. Liu, J. Ghiso, R. E. Marchant, and M. G. Zagorski.** 2003. pH-dependent amyloid and protofibril formation by the ABri peptide of familial British dementia. J Mol Biol **333:**1003-23.
111. **Tanaka, M., P. Chien, N. Naber, R. Cooke, and J. S. Weissman.** 2004. Conformational variations in an infectious protein determine prion strain differences. Nature **428:**323-8.
112. **Topilina, N. I., V. Sikirzhytsky, S. Higashiya, V. V. Ermolenkov, I. K. Lednev, and J. T. Welch.** Charge Distribution and Amyloid Fibril Formation: Insights from Genetically Engineered Model Systems. Biomacromolecules.
113. **True, H. L., and S. L. Lindquist.** 2000. A yeast prion provides a mechanism for genetic variation and phenotypic diversity. Nature **407:**477-83.
114. **Tukel, C., J. H. Nishimori, R. P. Wilson, M. G. Winter, A. M. Keestra, J. P. van Putten, and A. J. Baumler.** 2010. Toll-like receptors 1 and 2 cooperatively mediate immune responses to curli, a common amyloid from enterobacterial biofilms. Cell Microbiol.
115. **Tukel, C., M. Raffatellu, A. D. Humphries, R. P. Wilson, H. L. Andrews-Polymenis, T. Gull, J. F. Figueiredo, M. H. Wong, K. S. Michelsen, M. Akcelik, L. G. Adams, and A. J. Baumler.** 2005. CsgA is a pathogen-associated molecular pattern of Salmonella enterica serotype Typhimurium that is recognized by Toll-like receptor 2. Mol Microbiol **58:**289-304.
116. **Uversky, V. N., and A. L. Fink.** 2004. Conformational constraints for amyloid fibrillation: the importance of being unfolded. Biochim Biophys Acta **1698:**131-53.
117. **Uversky, V. N., J. R. Gillespie, and A. L. Fink.** 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins **41:**415-27.
118. **Vitek, M. P., K. Bhattacharya, J. M. Glendening, E. Stopa, H. Vlassara, R. Bucala, K. Manogue, and A. Cerami.** 1994. Advanced glycation end products contribute to amyloidosis in Alzheimer disease. Proc Natl Acad Sci U S A **91:**4766-70.
119. **Vitrenko, Y. A., E. O. Gracheva, J. E. Richmond, and S. W. Liebman.** 2007. Visualization of aggregation of the Rnq1 prion domain and cross-seeding interactions with Sup35NM. J Biol Chem **282:**1779-87.

120. **Wang, X., and M. R. Chapman.** 2008. Curli provide the template for understanding controlled amyloid propagation. Prion **2:**57-60.
121. **Wang, X., and M. R. Chapman.** 2008. Sequence determinants of bacterial amyloid formation. J Mol Biol **380:**570-80.
122. **Wang, X., N. D. Hammer, and M. R. Chapman.** 2008. The molecular basis of functional bacterial amyloid polymerization and nucleation. J Biol Chem.
123. **Wang, X., D. R. Smith, J. W. Jones, and M. R. Chapman.** 2007. In Vitro Polymerization of a Functional Escherichia coli Amyloid Protein. J Biol Chem **282:**3713-9.
124. **Wang, X., Y. Zhou, J. J. Ren, N. D. Hammer, and M. R. Chapman.** 2010. Gatekeeper residues in the major curlin subunit modulate bacterial amyloid fiber biogenesis. Proc Natl Acad Sci U S A **107:**163-8.
125. **Wang, Y., S. Petty, A. Trojanowski, K. Knee, D. Goulet, I. Mukerji, and J. King.** 2009. Formation of amyloid fibrils in vitro from partially unfolded intermediates of human gammaC-crystallin. Invest Ophthalmol Vis Sci **51:**672-8.
126. **Ward, J. J., J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones.** 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol **337:**635-45.
127. **White, A. P., and M. G. Surette.** 2006. Comparative genetics of the rdar morphotype in Salmonella. J Bacteriol **188:**8395-406.
128. **Wickner, R. B., H. K. Edskes, B. T. Roberts, U. Baxa, M. M. Pierce, E. D. Ross, and A. Brachmann.** 2004. Prions: proteins as genes and infectious entities. Genes Dev **18:**470-85.
129. **Wickner, R. B., H. K. Edskes, and F. Shewmaker.** 2006. How to find a prion: [URE3], [PSI+] and [beta]. Methods **39:**3-8.
130. **Wickner, R. B., K. L. Taylor, H. K. Edskes, M. L. Maddelein, H. Moriyama, and B. T. Roberts.** 2000. Prions of yeast as heritable amyloidoses. J Struct Biol **130:**310-22.
131. **Wright, C. F., S. A. Teichmann, J. Clarke, and C. M. Dobson.** 2005. The importance of sequence diversity in the aggregation and evolution of proteins. Nature **438:**878-81.
132. **Yeh, V., J. M. Broering, A. Romanyuk, B. Chen, Y. O. Chernoff, and A. S. Bommarius.** 2009. The Hofmeister effect on amyloid formation using yeast prion protein. Protein Sci **19:**47-56.
133. **Zerovnik, E., M. Skarabot, K. Skerget, S. Giannini, V. Stoka, S. Jenko-Kokalj, and R. A. Staniforth.** 2007. Amyloid fibril formation by human stefin B: influence of pH and TFE on fibril growth and morphology. Amyloid **14:**237-47.
134. **Zhu, L., X. J. Zhang, L. Y. Wang, J. M. Zhou, and S. Perrett.** 2003. Relationship between stability of folding intermediates and amyloid formation for the yeast prion Ure2p: a quantitative analysis of the effects of pH and buffer system. J Mol Biol **328:**235-54.

135. **Zogaj, X., W. Bokranz, M. Nimtz, and U. Romling.** 2003. Production of cellulose and curli fimbriae by members of the family Enterobacteriaceae isolated from the human gastrointestinal tract. Infect Immun **71:**4151-8.
136. **Zogaj, X., M. Nimtz, M. Rohde, W. Bokranz, and U. Romling.** 2001. The multicellular morphotypes of Salmonella typhimurium and Escherichia coli produce cellulose as the second component of the extracellular matrix. Mol Microbiol **39:**1452-63.

Chapter V

Future Directions and Synopsis

Future directions

How does a gcvA mutant affect curli production?

The Congo Red (CR) defect of the *gcvA* mutant (Fig. 1.6B) led me in several directions including developing programs to predict functional amyloids (See Appendix A) and studying the relative economy of extracellular proteins (Chapter III). GcvA both activates and represses the glycine cleavage system, which is responsible for degrading glycine to ammonia and carbon dioxide (36, 43, 51, 98-101, 124-127). Originally, I believed the *gcvA* strain was defective for curli production because it lacked sufficient glycine due to derepression of the glycine cleavage system. Intriguingly, CsgD increase the levels of GlyA (23, 24), which interconverts glycine and serine (Fig. 1.6C) (53). Consequently, I tested a *glyA* mutant strain for CR binding. However, the *glyA* from the Keio collection was not defective for CR binding even on YESCA plates with $1/10^{th}$ the regular amount of casamino acids (Fig 5.1A). As this was unexpected, I verified the *glyA* mutant by testing its growth in minimal media (52, 53); without addition of glycine there was no appreciable growth (Fig. 5.2). Intriguingly, addition of serine and glycine inhibited growth in multiple strains. Accumulation of serine has been shown to starve cells of S-adenosylmethionine and C1 unit resulting in defects in cell wall biosynthesis and division (129, 130). The Congo Red binding of *gcvA, gcvR, gcvTHP* and other glycine or serine metabolism mutants were also tested (Fig 5.1). Only the *gcvA* mutant was defective for curli production (Fig. 5.1C). Supplementing CR YESCA plates with glycine or serine was unable to rescue the *gcvA* mutant (Fig. 5.1C) even when 1% glycine was added (Fig. 5.1D).

Because replacement of glycine did not restore the *gcvA* strains CR phenotype and the other glycine metabolism mutants produced wild type levels of curli, I looked at what else GcvA regulates. GcvA is required to activate the transcription of the small RNA *gcvB* (66, 101, 114). Expression of *gcvB in trans* restored the *gcvA* strain's CR binding (Fig. 5.3). I also obtained plasmids containing *gcvA* and *gcvB* from George Stauffer, each of which partially rescued the CR phenotype *gcvA* strain (Fig. 5.4). Since expression of *gcvB* over expression restored the *gcvA* strain, I also tested the phenotype of a *gcvB* mutation. The *gcvB* strain was more defective than a *gcvA* strain for CR binding. Collectively, these results suggest the curli defect in *gcvA* is due to lower amounts of *gcvB* transcript.

Several amino acid transport systems including the periplasmic dipeptide (Dpp) and oligopeptide (Opp) transport systems are negatively regulated by *gcvB* (66, 81-84, 101, 114). DppA and OppA, the periplasmic binding proteins of these systems, are up regulated in several *csg* mutants. Periplasmic extracts from these *csg* mutants have been shown to inhibit CsgA-His polymerization (Fei Li and Margery Evans, unpublished observations). One possible reason for the CR defect in the *gcvA* and *gcvB* strains is the increased production of DppA and OppA due to the lack of repression by *gcvB*. To test this hypothesis I generated double mutants of *dppA* and *oppA* in *gcvA* and *gcvB* mutant backgrounds. In both cases deletion of the periplasmic component of the peptide transport system resulted in restored CR binding (Fig. 5.5).

Collectively, these data suggest the *gcvA* mutant produces less curli because it has lower transcription of *gcvB* which typically down regulates the *dppABCDF* and *oppABCDF* operons. Without this down regulation, higher levels of DppA and OppA accumulate in the periplasm. One possible reason for the curli defect is that CsgA or another curli specific protein interacts with DppA or OppA. Both DppA and OppA have been shown to have chaperone activity (64, 65, 85, 87). Alternatively, the deletion of either periplasmic binding protein prevents transport of small peptides that may have a regulatory effect on curli production. I am currently exploring these possibilities. I have cloned DppA-His for purification and have made double mutants in *gcvA* or *gcvB* and

other parts of the *dpp* and *opp* system such as the inner membrane components. Hopefully, I will be able to tease apart why higher levels of DppA or OppA result in less curli production.

Do less expensive extracellular proteins give the cell a metabolic leg up?

Ample evidence was presented in Chapter III regarding the economy of extracellular proteins; however, a clear *in vivo* example of the competitive fitness of having less expensive extracellular proteins is lacking. Excessive production of extracellular proteins has been shown to result in a competitive disadvantage (29, 57, 67, 77). Most of these examples involve mutants in which overproduction of flagella results in reduced cellular growth rates or competitive fitness, particularly in cases where there is no clear functional advantage for extraneous flagella. For example, constitutive flagellar mutants such as *flgM* and *fliD* grow slower due to excess FliC production (57), and flagellar mutants quickly overtake wild type strains (67). Additionally, *flgG* mutants, but not *motAB* mutants, outcompete wild type bacteria on plates (29), which suggests flagella synthesis itself is what taxes the cell when function does not provide an advantage. Lack of FlgG inhibits late flagellar development and thus secretion of extracellular flagellar proteins because secretion of the anti sigma factor FlgM requires hook assembly (22); however, a *motAB* mutant does not affect flagella assembly (63). A better example of the competitive costs of excessive extracellular protein production involves toxin production in the yeast *Pichia kluyveri.* Toxin-producing and toxin-deficient strains of *P. kluyveri* were grown in the presence of either toxin-sensitive *Pichia angusta* or toxin-resistant *Metschnikowia pulcherrima*. When grown with toxin-sensitive *P. angusta*, the toxin-producing strain of *P. kluyveri* was able to outcompete the toxin-deficient strain; however, the opposite result occurred when these *P. kluyveri* strains were grown with toxin-resistant *M. pulcherrima* (77). Thus it appears either overproduction of extracellular proteins or production in cases where there is not a functional advantage taxes the cell and decreases its growth rate or competitive fitness.

However, the examples presented above do not directly compare the cost of producing a more expensive extracellular protein with a less expensive one. To establish such a comparison one needs secretion of a functionally irrelevant protein in which mutations are made to increase or decrease the relative cost of its production. I considered several secretions systems to identify the best system for such an experiment. The Type I secretion system (T1SS) appears to be the best candidate. Unlike Type II, V, and VII secretion systems, T1SS secretes the protein from the cytoplasm to the extracellular milieu in a single step; the secreted protein does not enter the periplasm where it may interact with chaperones or proteases and possibly be degraded (3, 105). Additionally, the T1SS does not require contact with host membranes for secretion unlike many Type III, IV, and VI secretion systems (2, 105).

The T1SS also has the advantage of a known tag that can be added to another protein to allow for its secretion. Angkawidjaja et al 2006 (3) added the C-terminal tag of *Pseudomonas sp.* MIS38 lipase (PMLC) to *Escherichia coli* alkaline phosphatase (AP) to make AP-PMLC. This fusion protein is under the control of an IPTG inducible promoter and was efficiently secreted into the extracellular milieu while the untagged AP remained in the cytoplasm. Such a system would be excellent for testing the cost associated with more expensive extracellular proteins. Theoretically, one could mutate residues between the C-terminal tag and AP domain in the fusion protein to increase or decrease the metabolic cost of producing it. For example, addition of more tyrosine, tryptophan, and phenylalanine residues would increase the relative cost of the fusion protein, while addition of glycine, alanine, and serine residues would decrease the relative cost (1, 27, 44).

Using a library of such fusion proteins one could test the growth rate and competitive fitness of strains secreting proteins with different metabolic cost. To test the competitive fitness between strains one could mutate the AP domain of a particular strain to be completely inactive. AP active strains are easily differentiated; they will produce a black precipitate on plates amended with BCIP (5-bromo-4-chloro-3;-indoylphosphate p-Toluidine Salt) and NBT (Nitro-Blue Tetrazolium Chloride). If a

disadvantage for increased metabolic cost is found, increasing the levels of two proteins should exacerbate differences in growth or competitive fitness. By making similar sets of mutations in a cytoplasmically localized AP one could test whether the growth or competitive fitness defects are greater for more expensive proteins when they are secreted. Additionally, by adding amino acid media supplements, especially in *E. coli* single amino acid auxotrophs or strains lacking an amino acid transport systems, one could modulate and study the relative costs of secreting different proteins.

Experiments involving cellular costs and competitive fitness are not without difficulty and would require careful controls. The most important consideration in comparing strains expressing different fusion proteins is making sure they produce and secrete the same amount of protein. As long as mutations did not affect antibody binding, this could be accomplished by western blotting for AP-PMLC in the culture medium. Alternatively, one could test the enzymatic activity of the secreted proteins. However, the AP-PMLC used in Angkawidjaja et al 2006 did not fold particularly well unless it was purified in a glutathione redox buffer (3). This is likely due to the lack of disulfide bonds which are generally added to AP by the DsbA and DsbB system when AP is in the periplasm (97). There is a cold adapted alkaline phosphatase in *Vibrio sp*. (6) that does not require a disulfide bond for activity; however, it is not particularly stable at room temperature. Another enzyme particularly one with a colorimetric assay that can be adapted to plates could overcome these problems.

Are there expensive extracellular proteins?

The examination of protein economy began out of a desire to quantify the relative richness of glycine in CsgA. Once the amino acid compositions of all the proteins in *E. coli* were calculated, only a small step was needed to calculate their relative synthetic costs. Using this data, I noted another intriguing aspect of CsgA, a lower cost relative to other proteins. In Chapter III, I found that most if not all of the extracellular proteins in other microbes were economical; however, there were outliers amongst

extracellular proteins. Two of these worth noting are alpha factor in yeast and the major subunit of the conjugative pilus in *E. coli*.

Initially, Matt and I enjoyed entertaining the idea this discrepancy reflected that sex was more expensive. Many higher animals such a birds incur a fitness cost through extravagant displays to attract a mate (80). However, such an anthropomorphic thought belied a simpler explanation. Tzvi Tzfira approached us concerning the economy of transformation by *Agrobacterium tumefaciens,* in which a system homologous to the conjugative pilus is used to transfer DNA into plants (2, 32, 109, 110). When I examined the protein cost in *A. tumefaciens,* I found similar results as in other microbes. Extracellular proteins were significantly less expensive (Fig. 5.7). When I focused on the expense of Type IV system proteins, I noticed that many of the trends were similar to Type III and Type I secretion systems. Membrane and regulatory elements were average or relatively expensive while secreted proteins were economical.

Intriguingly, the Type IV pilus subunit VirB2 was more expensive than an average protein when examined for protein cost (red circle, Fig. 5.7). However, when using amino acid mass instead of synthetic cost, VirB2 is one of the most economical in the cell (green circle, Fig. 5.7). Accordingly, VirB2 has a relatively low carbon and nitrogen content. Collectively, the economic rank for VirB2 is 2822, 55, 227 and 399 out of 5402 proteins using synthetic costs, mass, carbon, and nitrogen content to calculate relative amino acid cost, respectively. Such a strange result made me reexamine why there was such a discrepancy. As noted in Chapter I, each of the different ways to cost biases for certain amino acids over others. Consequently, I examined the amino acid composition of VirB2 and noticed it had a very low percent of acidic residues; less than 1.7% of the residues in VirB2 were acidic.

I examined the composition of other proteins to see if any had low acidic residue contents. A histogram of the percentage of Glu and Asp residues results in a bimodal distribution (Fig. 5.8). A similar distribution is seen for basic residues (data not shown) of which VirB2 is also depleted. Intriguingly, this bimodal distribution has a biological basis. Many inner membrane proteins are among the proteins with the least amount of

charged amino acids. These proteins include aquaporin and transporters of sodium, benzoate, ammonia, and other charged molecules (Fig. 5.8B). Collectively, these results suggest VirB2 is under the same strong selection to reduce charged residues in its composition as proteins that transport charged molecules. Accordingly, the role of VirB2 in the cell is to transport negatively charged DNA (2, 109, 110). Additionally, I found that many of the type IV effectors (2) including VirD2, VirD5, VirE2, VirE3, and VirF were quite rich in charged amino acids, all of which had greater than 26.5% EDRK. Thus it appears that the composition of VirB2 is the result of a delicate balancing act between economy of carbon atoms and a functional reduction of charged amino acids.

Synopsis

Curli are thin proteinaceous fibers produced by many Enterobacteriaceae (11, 16, 21, 42, 89, 117, 118, 120, 121, 131), primarily as a structural component of biofilms (26, 54, 93, 96, 116). The formation of curli is incredibly complex and includes regulation of and by CsgD (4, 5, 11, 17, 19, 31, 33-35, 40, 46-48, 69, 70, 74, 78, 90, 91, 115, 116, 128); the accessory proteins CsgE, CsgF, and CsgG (11, 21, 30, 37, 62, 68, 88); and the major and minor fiber subunits CsgA and CsgB (21, 41, 42, 117, 118, 120, 121). Curli have been implicated in a variety of other biological functions including environmental persistence (92, 111, 119, 122, 123); pathogenesis including host cell adherence, internalization, and invasion (38, 39, 50, 56, 58, 59, 61, 94, 102, 112, 113, 119), binding to host factors (13, 25, 45, 71-74, 76, 95), and inducing the host inflammatory response (14, 15, 72, 106-108, 119); and adherence to plant (9, 10, 18, 49, 55, 60, 104) and manmade surfaces (7, 12, 16, 20, 26, 75, 79, 103, 111).

In this work I focused on the development, evolution and amyloidogenesis of curli using genetic, *in silico,* and biochemical tools. To study the development of curli formation in *Escherichia coli* K-12, I screened the Keio collection, a near complete collection of deletion mutants. I found more than 300 genes that modulated curli production. For example, the regulator of glycine metabolism GcvA, the sodium antiporter NhaA, NADH dehydrogenase, and many other proteins involved in

fundamental cellular processes affect curli production. Many of these genes likely regulate curli either directly or indirectly through either two component systems or the alternative sigma factors RpoS and RpoE. Despite the many regulatory and metabolic genes found to affect curli, there is little overlap between genes affecting curli and motility suggesting curli production is a distinct cellular lifestyle.

To study the evolution of curli, I examined the composition and cost of amino acids in CsgA and CsgB. Both proteins, as well as many other extracellular proteins in a wide variety of microbes, appear to have been selected for increased economy. In curli producing cells, curli production is one of the most highly expressed proteins made, yet because of the lack of import systems many of the cost associated with building the biofilm matrix are not recovered. Thus selection for less expensive amino acids in curli saves the cell needed resources during stationary phase, a time where carbon and other resources are imitated.

Finally, to study the amyloidogenesis of curli, I examined the conditions that promote the *in vitro* polymerization of purified CsgA-His. Like disease-associated amyloids, CsgA-His bound Thioflavin T upon polymerizing into fibers, reacted with the amyloid specific A11 antibody, self seeded, and displayed other aspects of amyloid fiber formation. However, CsgA also displayed some unique properties. The quick adoption of the species recognized by the A11 antibody suggests that the many repeating units in CsgA may limit some of the toxic intermediates in amyloidogenesis. Additionally, as an evolved amyloid with a separate nucleator, I find the complementary pI's of CsgA and CsgB quite intriguing.

As a model system, many aspects of curli formation have been explored in the literature including the complex regulation, secretion, assembly, biological function, and amyloidogenesis. Collectively, this investigation sheds new light on the biology of the functional amyloid curli and the evolutionary pressures on extracellular proteins. My study of curli has led me down branched paths too numerous to mention. There remain many questions, both unanswered and unasked. Hopefully, this work will beget exciting new directions of inquiry.

Figure Legends

Fig. 5.1. Addition of glycine, serine, or both has little effect on the CR phenotype of several stains lacking glycine and serine metabolism genes. (A) Strains were grown on CR YESCA with 1/10th the normal amount of Casamino acids per liter. (B) Strains were grown on CR YESCA with half the normal amount of Casamino acids per liter. (C) Strains were grown on regular CR YESCA. (D) Even addition of 1% glycine did not rescue the CR binding of the *gcvA* mutant. In each case the strains were grown for 2 days at 26^{o}C.

Fig. 5.2. The *glyA* strain cannot grow in minimal media without the addition of glycine. W minimal media (pH 7.4) was amended 0.4% Glucose, 0.2% Glutamine, and 0.1ug/mL Thiamine. The final concentration of 1X glycine or serine if present was 0.01%. (A) After 36 hours, there is no appreciable growth in *glyA* strain in the absence of glycine. Glycine or glycine and serine addition rescue the growth defect of *glyA.* (B). Absorbance readings of the BW25113, *gcvA,* and *glyA* strains at the times indicated. A *glyA* strain should requires glycine to grow in minimal media (52, 53); the Keio collection *glyA* strain does require glycine addition for growth.

Fig. 5.3. A plasmid expressing *gcvB* rescues the CR binding defect of a *gcvA* mutant. The RNA *gcvB* was cloned into pQE-70 which is a IPTG inducible high copy number plasmid to create pGcvB. Transformation of pGcvB into *gcvA* strains partially rescued the strains CR binding defect, especially when IPTG was added. The different patches for each strain represent independent isolates. Strains were grown on CR YESCA plates for 2 days at 26^{o}C.

Fig. 5.4. Plasmids expressing *gcvB* or *gcvA* rescue the CR binding defect of the *gcvA* mutant. The pGS plasmids were obtained from George Stauffer and were transformed into *gcvA* strains. Plasmids containing *gcvA* or *gcvB* were able to partially complement the CR defect of the *gcvA* strain. Strains were grown on CR YESCA plates for 2 days at 26^{o}C.

Fig. 5.5. A *gcvB* deletion binds less CR than the *gcvA* mutant. The *gcvB* strain bound less Congo Red than the *gcvA* strain. A *gcvB* deletion strain was made by replacing the *gcvB* gene with a kanamyacin cassette using red swap mutagenesis (28). The different patches for each strain represent independent isolates. The *gcvA dppA* strain was created by flipping out the Kan cassette from the Keio *gcvA* strain and moving the *dppA::Kan* region using P1 transduction (8, 28). Strains were grown on CR YESCA plates for 2 days at 26^{o}C.

Fig. 5.6. Deletion of *dppA* or *oppA* rescues the CR defect of the *gcvA* and *gcvB* strains. Although *dppA* and *oppA* mutants were wild type for curli production, double mutants of either with *gcvA* or *gcvB* resulted in increased CR binding. Double deletion mutants were made by flipping out the Kan cassette with pCP20 (8, 28). Strains were grown on CR YESCA plates for 2 days at 26^{o}C.

Fig. 5.7. Average synthetic cost and mass of proteins in different locations from *Agrobacterium tumefaciens*. Average synthetic cost was calculated using values for chemoheterotrophic bacteria (1, 44). Protein locations were from PSORTDB version 3.0 (86); a few type IV effectors were amended to extracellular (2). VirD5 was not changed but is economical. (A) Average synthetic cost in ATPs of proteins in various locations (B) Average mass of amino acids of proteins in various locations.

Fig. 5.8. Distribution of amino acid composition in proteins of *Agrobacterium tumefaciens*. (A) Histogram of the percent of glutamic acid and aspartic acid. (B) Histogram of percent of basic and acidic residues. Green (All proteins), Red (Inner Membrane proteins), and Purple (Transporters). Inner membrane proteins were predicted using PSORTDB version 3.0. Transporters include aquaporin, SecG, CCOI and inner membrane proteins that were annotated as ABC transport membrane spanning, antiporter, or symporter.

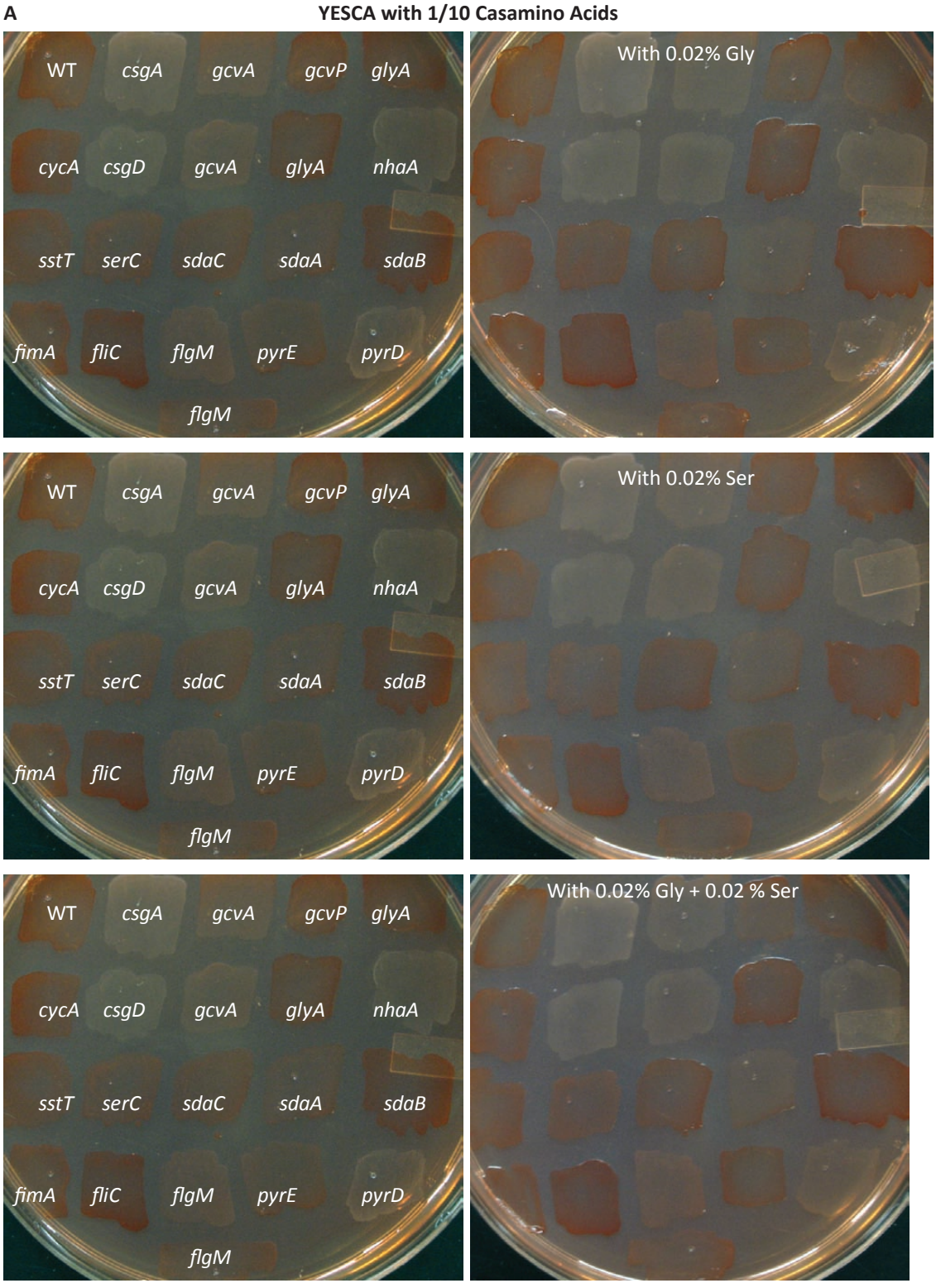


Figure 5.1. Addition of glycine, serine, or both has little effect on the CR phenotype of several stains lacking glycine and serine metabolism genes.

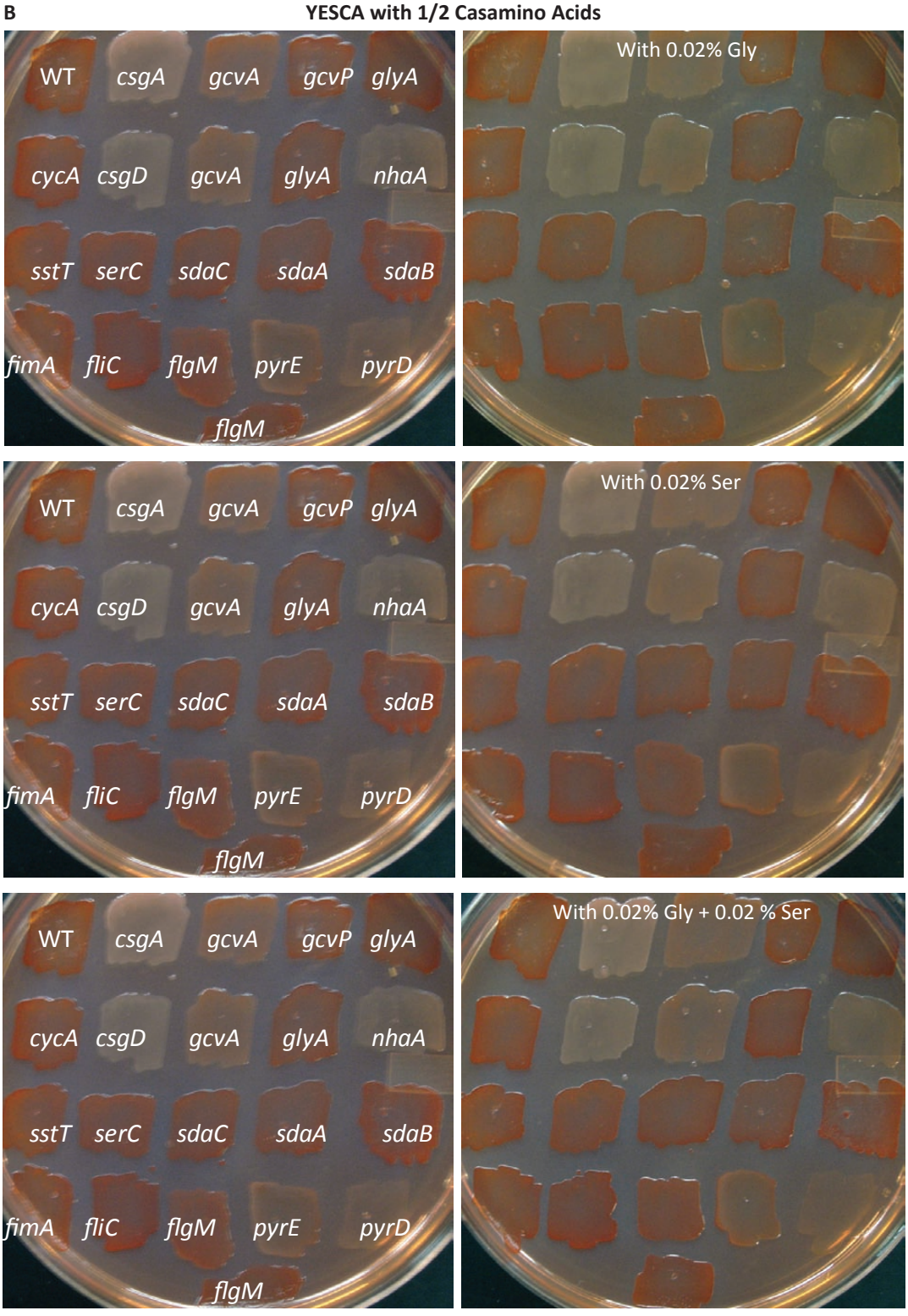


Figure 5.1 (continued). Addition of glycine, serine, or both has little effect on the CR phenotype of several stains lacking glycine and serine metabolism genes.

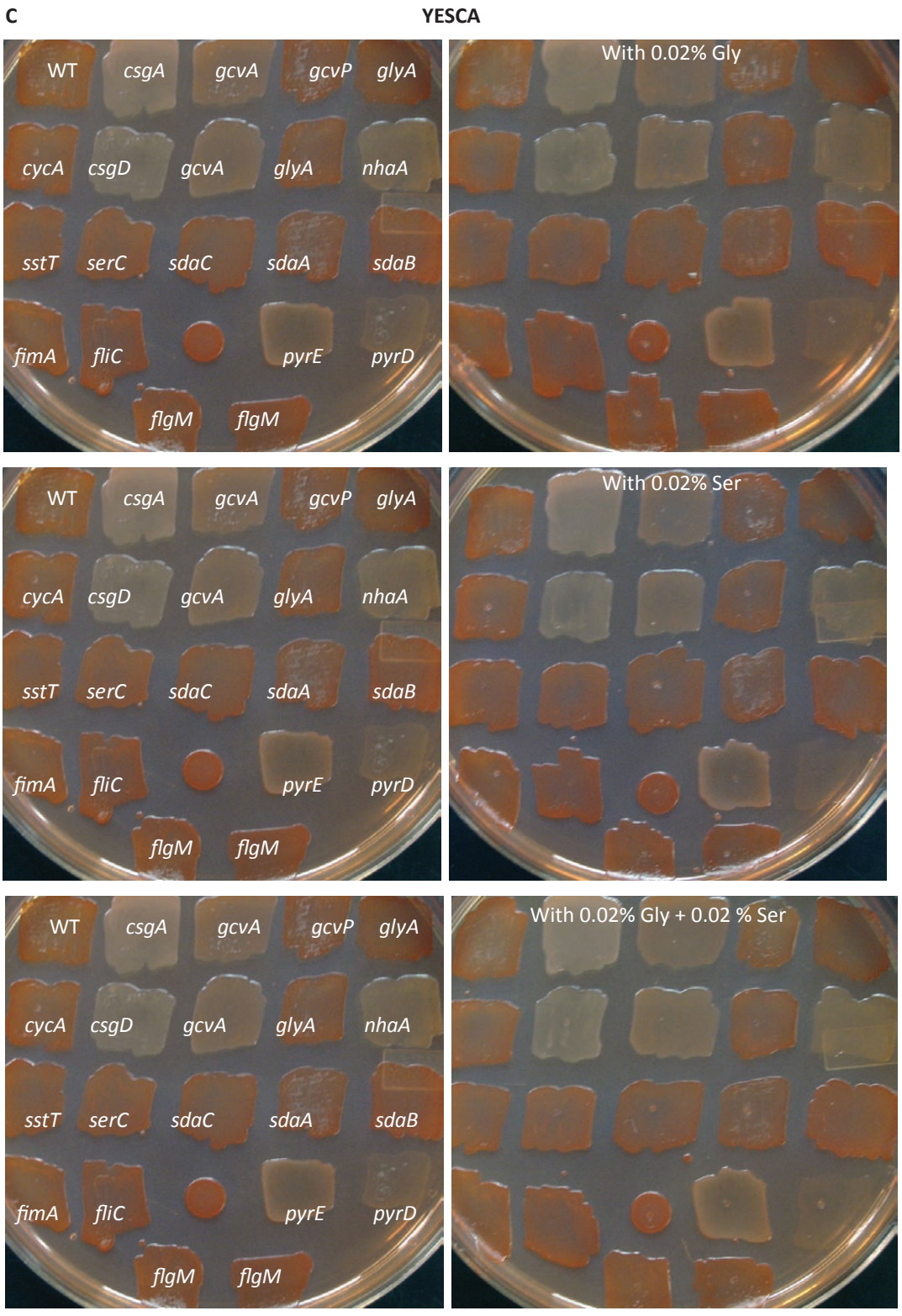


Figure 5.1 (continued). Addition of glycine, serine, or both has little effect on the CR phenotype of several stains lacking glycine and serine metabolism genes.

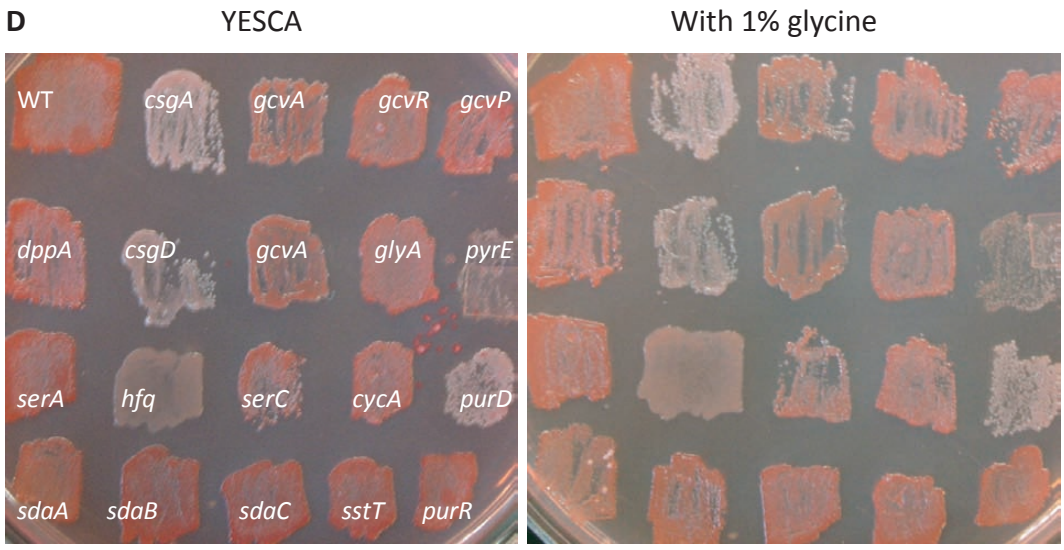


Figure 5.1 (continued). Addition of glycine, serine, or both has little effect on the CR phenotype of several stains lacking glycine and serine metabolism genes.

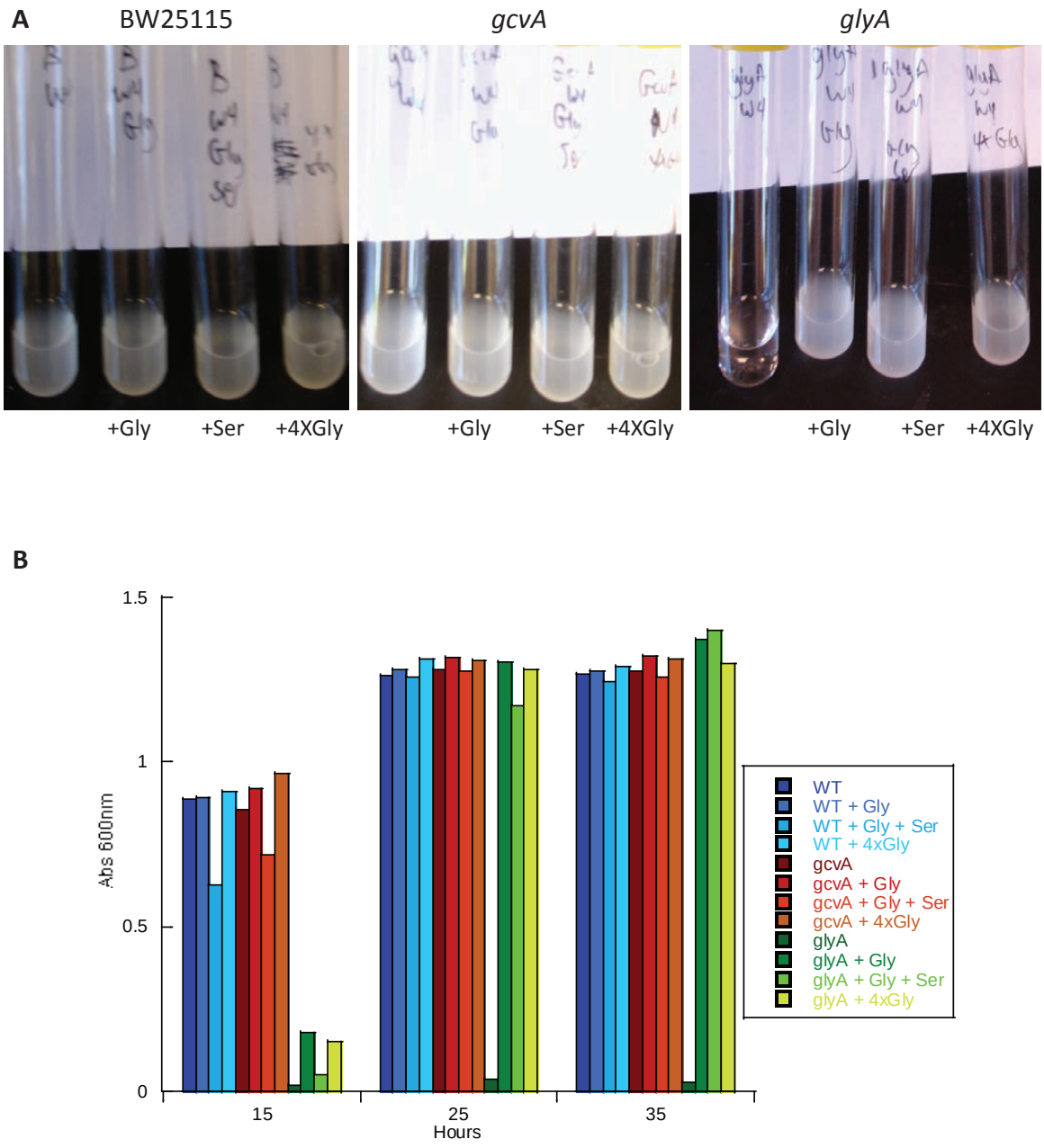


Figure 5.2. The *glyA* strain cannot grow in minimal media without the addition of glycine.

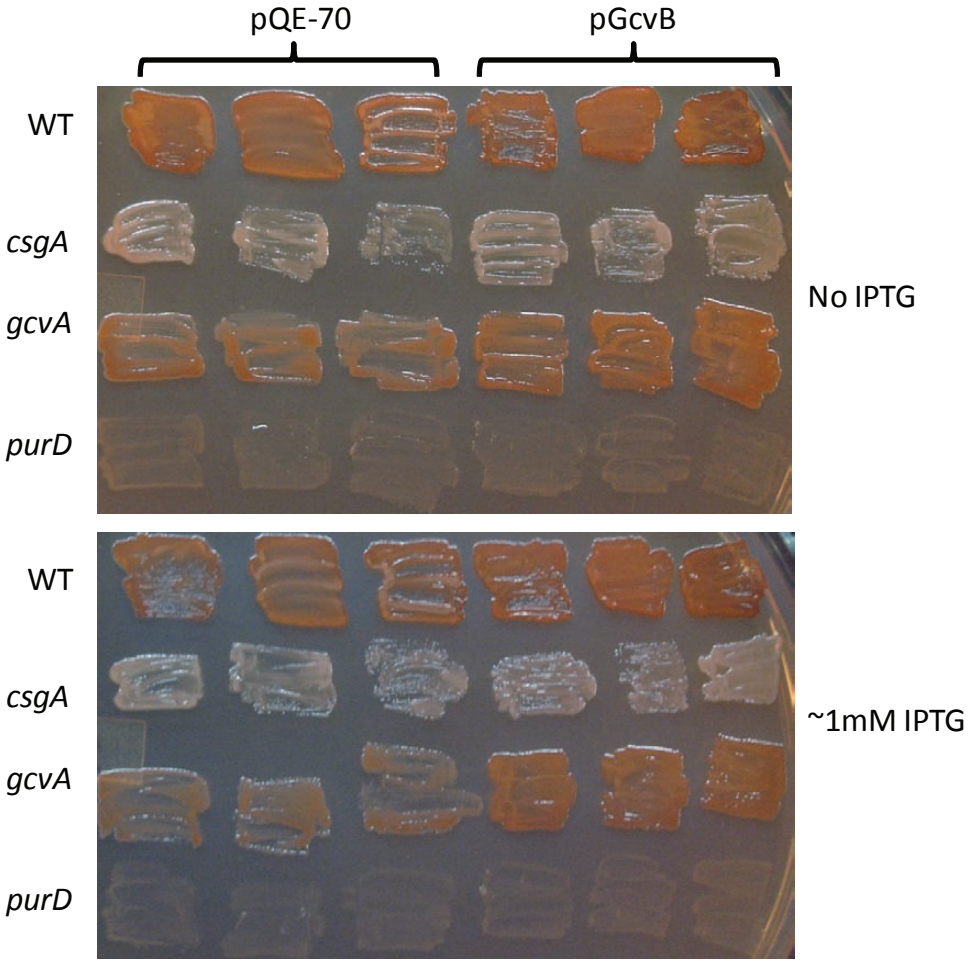
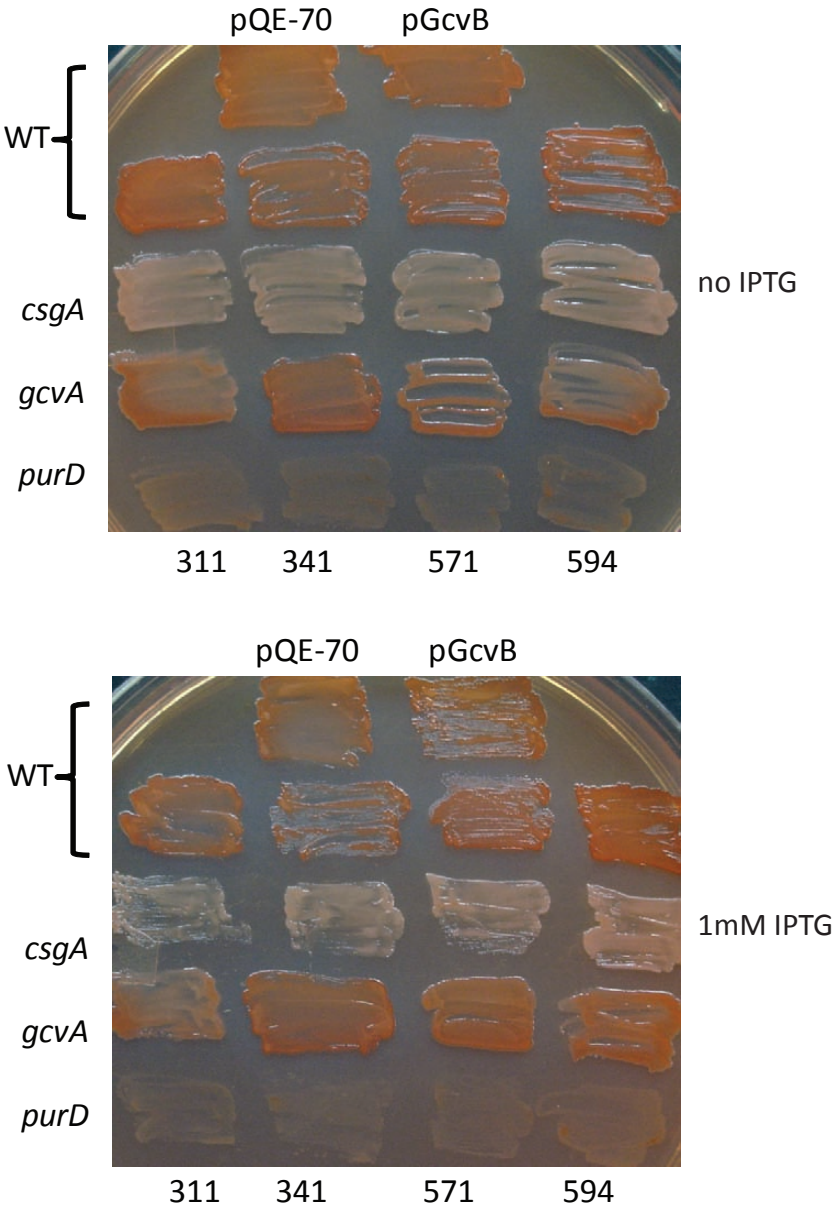


Figure. 5.3. A plasmid expressing *gcvB* rescues the CR binding defect of a *gcvA* mutant.



pGS311 – low copy vector control
pGS341 – *gcvA* on pGS311
pGS571 – *gcvB* on high copy plasmid
pGS594 – *gcvB* on pGS311

Figure. 5.4. Plasmids expressing *gcvB* or *gcvA* rescue the CR binding defect of the *gcvA* mutant.

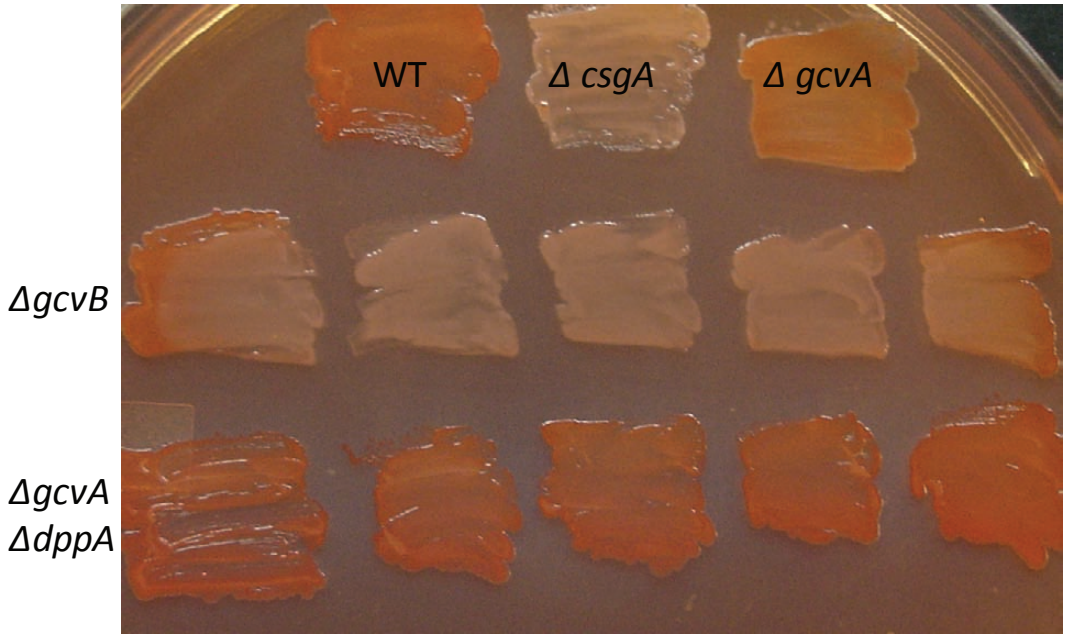


Figure. 5.5. A *gcvB* deletion binds less CR than the *gcvA* mutant.

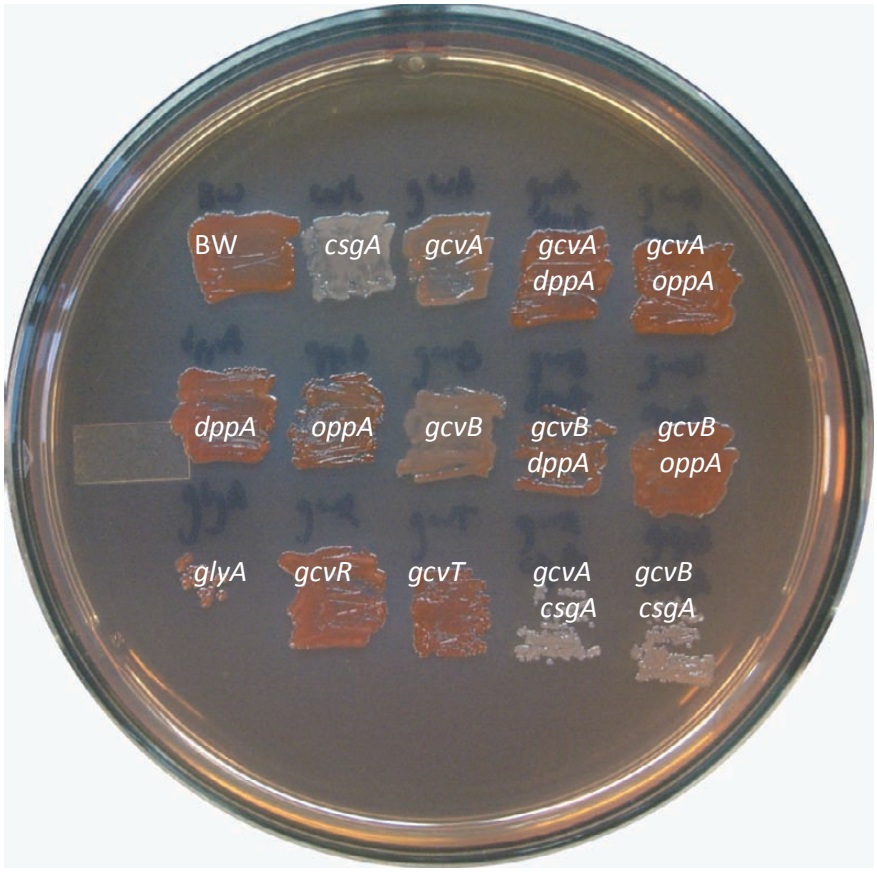


Figure. 5.6. Deletion of *dppA* or *oppA* rescues the CR defect of the *gcvA* and *gcvB* strains.

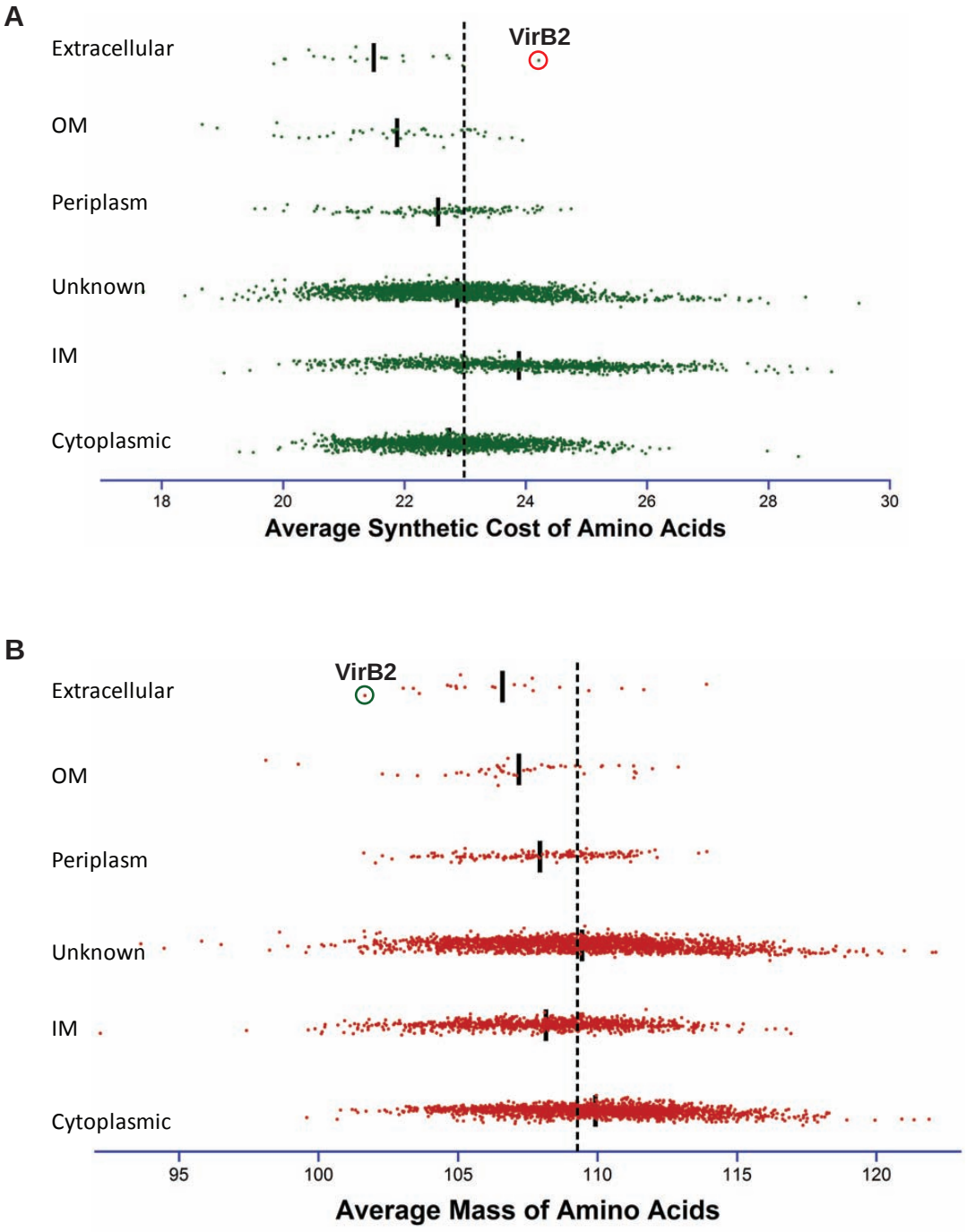


Figure 5.7. Average synthetic cost and mass of proteins in different locations from *Agrobacterium tumefaciens*

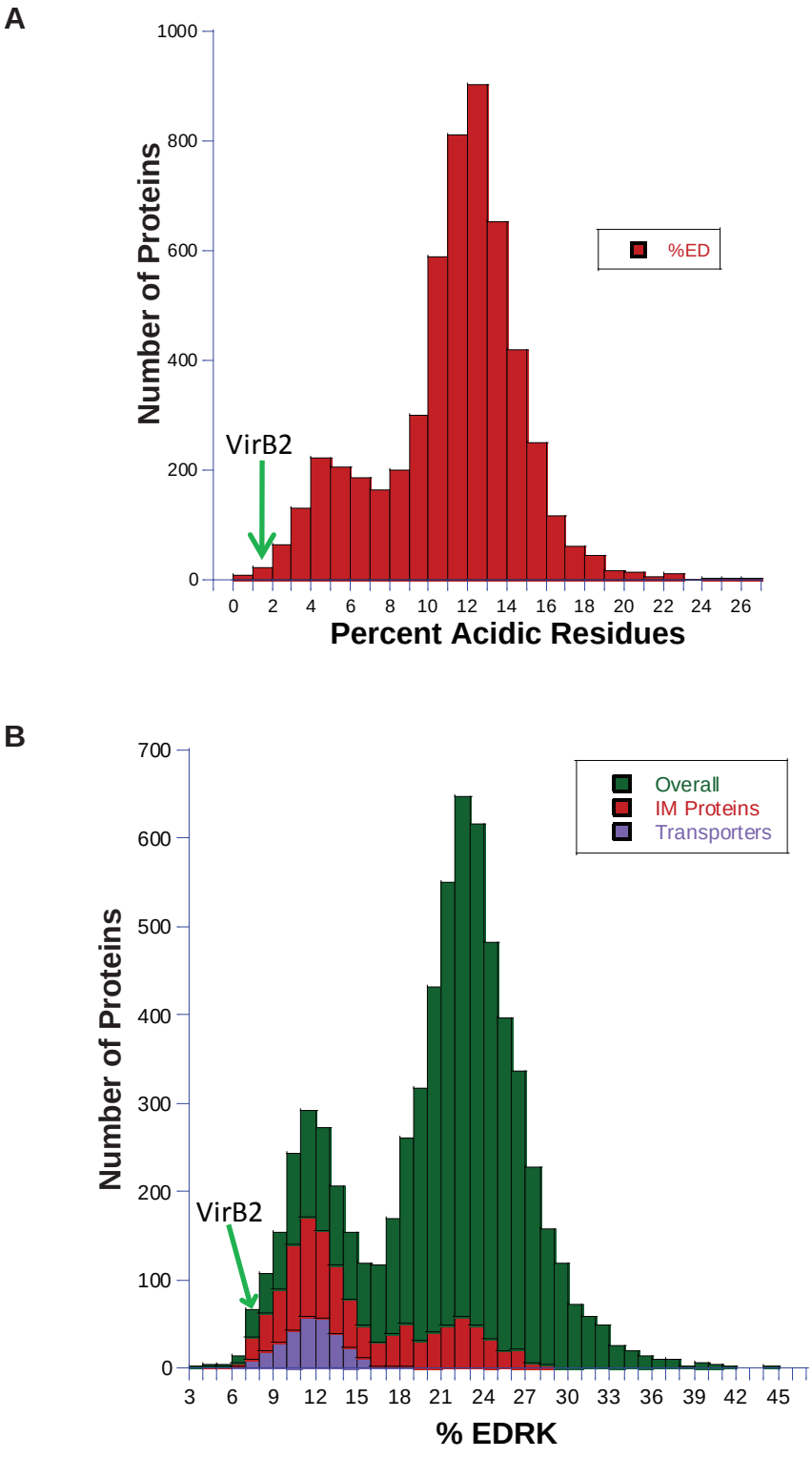


Figure 5.8 Distribution of amino acid composition in proteins of *Agrobacterium tumefaciens.*

References

1. **Akashi, H., and T. Gojobori.** 2002. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc Natl Acad Sci U S A **99:**3695-700.
2. **Alvarez-Martinez, C. E., and P. J. Christie.** 2009. Biological diversity of prokaryotic type IV secretion systems. Microbiol Mol Biol Rev **73:**775-808.
3. **Angkawidjaja, C., K. Kuwahara, K. Omori, Y. Koga, K. Takano, and S. Kanaya.** 2006. Extracellular secretion of Escherichia coli alkaline phosphatase with a C-terminal tag by type I secretion system: purification and biochemical characterization. Protein Eng Des Sel **19:**337-43.
4. **Arnqvist, A., A. Olsen, and S. Normark.** 1994. Sigma S-dependent growth-phase induction of the csgBA promoter in Escherichia coli can be achieved in vivo by sigma 70 in the absence of the nucleoid-associated protein H-NS. Mol Microbiol **13:**1021-32.
5. **Arnqvist, A., A. Olsen, J. Pfeifer, D. G. Russell, and S. Normark.** 1992. The Crl protein activates cryptic genes for curli formation and fibronectin binding in Escherichia coli HB101. Mol Microbiol **6:**2443-52.
6. **Asgeirsson, B., B. V. Adalbjornsson, and G. A. Gylfason.** 2007. Engineered disulfide bonds increase active-site local stability and reduce catalytic activity of a cold-adapted alkaline phosphatase. Biochim Biophys Acta **1774:**679-87.
7. **Austin, J. W., G. Sanders, W. W. Kay, and S. K. Collinson.** 1998. Thin aggregative fimbriae enhance Salmonella enteritidis biofilm formation. FEMS Microbiol Lett **162:**295-301.
8. **Baba, T., T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori.** 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol **2:**2006 0008.
9. **Barak, J. D., L. Gorski, P. Naraghi-Arani, and A. O. Charkowski.** 2005. Salmonella enterica virulence genes are required for bacterial attachment to plant tissue. Appl Environ Microbiol **71:**5685-91.
10. **Barak, J. D., L. C. Whitehand, and A. O. Charkowski.** 2002. Differences in attachment of Salmonella enterica serovars and Escherichia coli O157:H7 to alfalfa sprouts. Appl Environ Microbiol **68:**4758-63.
11. **Barnhart, M. M., and M. R. Chapman.** 2006. Curli biogenesis and function. Annu Rev Microbiol **60:**131-47.
12. **Beloin, C., A. Houry, M. Froment, J. M. Ghigo, and N. Henry.** 2008. A Short-Time Scale Colloidal System Reveals Early Bacterial Adhesion Dynamics. PLoS Biol **6:**e167.
13. **Ben Nasr, A., A. Olsen, U. Sjobring, W. Muller-Esterl, and L. Bjorck.** 1996. Assembly of human contact phase proteins and release of bradykinin at the surface of curli-expressing Escherichia coli. Mol Microbiol **20:**927-35.
14. **Bian, Z., A. Brauner, Y. Li, and S. Normark.** 2000. Expression of and cytokine activation by Escherichia coli curli fibers in human sepsis. J Infect Dis **181:**602-12.

15. **Bian, Z., Z. Q. Yan, G. K. Hansson, P. Thoren, and S. Normark.** 2001. Activation of inducible nitric oxide synthase/nitric oxide by curli fibers leads to a fall in blood pressure during systemic Escherichia coli infection in mice. J Infect Dis **183:**612-9.
16. **Bokranz, W., X. Wang, H. Tschape, and U. Romling.** 2005. Expression of cellulose and curli fimbriae by Escherichia coli isolated from the gastrointestinal tract. J Med Microbiol **54:**1171-82.
17. **Bougdour, A., C. Lelong, and J. Geiselmann.** 2004. Crl, a low temperature-induced protein in Escherichia coli that binds directly to the stationary phase sigma subunit of RNA polymerase. J Biol Chem **279:**19540-50.
18. **Boyer, R. R., S. S. Sumner, R. C. Williams, M. D. Pierson, D. L. Popham, and K. E. Kniel.** 2007. Influence of curli expression by Escherichia coli 0157:H7 on the cell's overall hydrophobicity, charge, and ability to attach to lettuce. J Food Prot **70:**1339-45.
19. **Brown, P. K., C. M. Dozois, C. A. Nickerson, A. Zuppardo, J. Terlonge, and R. Curtiss, 3rd.** 2001. MlrA, a novel regulator of curli (AgF) and extracellular matrix synthesis by Escherichia coli and Salmonella enterica serovar Typhimurium. Mol Microbiol **41:**349-63.
20. **Castonguay, M. H., S. van der Schaaf, W. Koester, J. Krooneman, W. van der Meer, H. Harmsen, and P. Landini.** 2006. Biofilm formation by Escherichia coli is stimulated by synergistic interactions and co-adhesion mechanisms with adherence-proficient bacteria. Res Microbiol **157:**471-8.
21. **Chapman, M. R., L. S. Robinson, J. S. Pinkner, R. Roth, J. Heuser, M. Hammar, S. Normark, and S. J. Hultgren.** 2002. Role of Escherichia coli curli operons in directing amyloid fiber formation. Science **295:**851-5.
22. **Chilcott, G. S., and K. T. Hughes.** 2000. Coupling of flagellar gene expression to flagellar assembly in Salmonella enterica serovar typhimurium and Escherichia coli. Microbiol Mol Biol Rev **64:**694-708.
23. **Chirwa, N. T., and M. B. Herrington.** 2003. CsgD, a regulator of curli and cellulose synthesis, also regulates serine hydroxymethyltransferase synthesis in Escherichia coli K-12. Microbiology **149:**525-35.
24. **Chirwa, N. T., and M. B. Herrington.** 2004. Role of MetR and PurR in the activation of glyA by CsgD in Escherichia coli K-12. Can J Microbiol **50:**683-90.
25. **Collinson, S. K., P. C. Doig, J. L. Doran, S. Clouthier, T. J. Trust, and W. W. Kay.** 1993. Thin, aggregative fimbriae mediate binding of Salmonella enteritidis to fibronectin. J Bacteriol **175:**12-8.
26. **Cookson, A. L., W. A. Cooley, and M. J. Woodward.** 2002. The role of type 1 and curli fimbriae of Shiga toxin-producing Escherichia coli in adherence to abiotic surfaces. Int J Med Microbiol **292:**195-205.
27. **Craig, C. L., and R. S. Weber.** 1998. Selection costs of amino acid substitutions in ColE1 and Colla gene clusters harbored by Escherichia coli. Mol Biol Evol **15:**774-6.

28. **Datsenko, K. A., and B. L. Wanner.** 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A **97:**6640-5.
29. **Easom, C. A., and D. J. Clarke.** 2008. Motility is required for the competitive fitness of entomopathogenic Photorhabdus luminescens during insect infection. BMC Microbiol **8:**168.
30. **Epstein, E. A., and M. R. Chapman.** 2008. Polymerizing the fibre between bacteria and host cells: the biogenesis of functional amyloid fibres. Cell Microbiol **10:**1413-20.
31. **Ferrieres, L., and D. J. Clarke.** 2003. The RcsC sensor kinase is required for normal biofilm formation in Escherichia coli K-12 and controls the expression of a regulon in response to growth on a solid surface. Mol Microbiol **50:**1665-82.
32. **Fronzes, R., P. J. Christie, and G. Waksman.** 2009. The structural biology of type IV secretion systems. Nat Rev Microbiol **7:**703-14.
33. **Gerstel, U., C. Park, and U. Romling.** 2003. Complex regulation of csgD promoter activity by global regulatory proteins. Mol Microbiol **49:**639-54.
34. **Gerstel, U., and U. Romling.** 2003. The csgD promoter, a control unit for biofilm formation in Salmonella typhimurium. Res Microbiol **154:**659-67.
35. **Gerstel, U., and U. Romling.** 2001. Oxygen tension and nutrient starvation are major signals that regulate agfD promoter activity and expression of the multicellular morphotype in Salmonella typhimurium. Environ Microbiol **3:**638-48.
36. **Ghrist, A. C., G. Heil, and G. V. Stauffer.** 2001. GcvR interacts with GcvA to inhibit activation of the Escherichia coli glycine cleavage operon. Microbiology **147:**2215-21.
37. **Gibson, D. L., A. P. White, C. M. Rajotte, and W. W. Kay.** 2007. AgfC and AgfE facilitate extracellular thin aggregative fimbriae synthesis in Salmonella enteritidis. Microbiology **153:**1131-40.
38. **Gophna, U., M. Barlev, R. Seijffers, T. A. Oelschlager, J. Hacker, and E. Z. Ron.** 2001. Curli fibers mediate internalization of Escherichia coli by eukaryotic cells. Infect Immun **69:**2659-65.
39. **Gophna, U., T. A. Oelschlaeger, J. Hacker, and E. Z. Ron.** 2002. Role of fibronectin in curli-mediated internalization. FEMS Microbiol Lett **212:**55-8.
40. **Hammar, M., A. Arnqvist, Z. Bian, A. Olsen, and S. Normark.** 1995. Expression of two csg operons is required for production of fibronectin- and congo red-binding curli polymers in Escherichia coli K-12. Mol Microbiol **18:**661-70.
41. **Hammar, M., Z. Bian, and S. Normark.** 1996. Nucleator-dependent intercellular assembly of adhesive curli organelles in Escherichia coli. Proc Natl Acad Sci U S A **93:**6562-6.
42. **Hammer, N. D., J. C. Schmidt, and M. R. Chapman.** 2007. The curli nucleator protein, CsgB, contains an amyloidogenic domain that directs CsgA polymerization. Proc Natl Acad Sci U S A **104:**12494-9.

43. **Heil, G., L. T. Stauffer, and G. V. Stauffer.** 2002. Glycine binds the transcriptional accessory protein GcvR to disrupt a GcvA/GcvR interaction and allow GcvA-mediated activation of the Escherichia coli gcvTHP operon. Microbiology **148:**2203-14.
44. **Heizer, E. M., Jr., D. W. Raiford, M. L. Raymer, T. E. Doom, R. V. Miller, and D. E. Krane.** 2006. Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. Mol Biol Evol **23:**1670-80.
45. **Herwald, H., M. Morgelin, A. Olsen, M. Rhen, B. Dahlback, W. Muller-Esterl, and L. Bjorck.** 1998. Activation of the contact-phase system on bacterial surfaces--a clue to serious complications in infectious diseases. Nat Med **4:**298-302.
46. **Holmqvist, E., J. Reimegard, M. Sterk, N. Grantcharova, U. Romling, and E. G. Wagner.** 2010. Two antisense RNAs target the transcriptional regulator CsgD to inhibit curli synthesis. Embo J **29:**1840-50.
47. **Hung, D. L., T. L. Raivio, C. H. Jones, T. J. Silhavy, and S. J. Hultgren.** 2001. Cpx signaling pathway monitors biogenesis and affects assembly and expression of P pili. Embo J **20:**1508-18.
48. **Ishihama, A.** 2010. Prokaryotic genome regulation: multifactor promoters, multitarget regulators and hierarchic networks. FEMS Microbiol Rev.
49. **Jeter, C., and A. G. Matthysse.** 2005. Characterization of the binding of diarrheagenic strains of E. coli to plant surfaces and the role of curli in the interaction of the bacteria with alfalfa sprouts. Mol Plant Microbe Interact **18:**1235-42.
50. **Johansson, C., T. Nilsson, A. Olsen, and M. J. Wick.** 2001. The influence of curli, a MHC-I-binding bacterial surface structure, on macrophage-T cell interactions. FEMS Immunol Med Microbiol **30:**21-9.
51. **Jourdan, A. D., and G. V. Stauffer.** 1999. GcvA-mediated activation of gcvT-lacZ expression involves the carboxy-terminal domain of the alpha subunit of RNA polymerase. FEMS Microbiol Lett **181:**307-12.
52. **Joyce, A. R., J. L. Reed, A. White, R. Edwards, A. Osterman, T. Baba, H. Mori, S. A. Lesely, B. O. Palsson, and S. Agarwalla.** 2006. Experimental and computational assessment of conditionally essential genes in Escherichia coli. J Bacteriol **188:**8259-71.
53. **Karp, P. D., I. M. Keseler, A. Shearer, M. Latendresse, M. Krummenacker, S. M. Paley, I. Paulsen, J. Collado-Vides, S. Gama-Castro, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penaloza-Spinola, C. Bonavides-Martinez, and J. Ingraham.** 2007. Multidimensional annotation of the Escherichia coli K-12 genome. Nucleic Acids Res **35:**7577-90.
54. **Kikuchi, T., Y. Mizunoe, A. Takade, S. Naito, and S. Yoshida.** 2005. Curli fibers are required for development of biofilm architecture in Escherichia coli K-12 and enhance bacterial adherence to human uroepithelial cells. Microbiol Immunol **49:**875-84.

55. **Kim, J. K., and M. A. Harrison.** 2009. Surrogate selection for Escherichia coli O157:H7 based on cryotolerance and attachment to romaine lettuce. J Food Prot **72:**1385-91.
56. **Kim, S. H., and Y. H. Kim.** 2004. Escherichia coli O157:H7 adherence to HEp-2 cells is implicated with curli expression and outer membrane integrity. J Vet Sci **5:**119-24.
57. **Kutsukake, K., and T. Iino.** 1994. Role of the FliA-FlgM regulatory system on the transcriptional control of the flagellar regulon and flagellar formation in Salmonella typhimurium. J Bacteriol **176:**3598-605.
58. **La Ragione, R. M., W. A. Cooley, and M. J. Woodward.** 2000. The role of fimbriae and flagella in the adherence of avian strains of Escherichia coli O78:K80 to tissue culture cells and tracheal and gut explants. J Med Microbiol **49:**327-38.
59. **La Ragione, R. M., I. M. McLaren, G. Foster, W. A. Cooley, and M. J. Woodward.** 2002. Phenotypic and genotypic characterization of avian Escherichia coli O86:K61 isolates possessing a gamma-like intimin. Appl Environ Microbiol **68:**4932-42.
60. **Lapidot, A., and S. Yaron.** 2009. Transfer of Salmonella enterica serovar Typhimurium from contaminated irrigation water to parsley is dependent on curli and cellulose, the biofilm matrix components. J Food Prot **72:**618-23.
61. **Lawley, T. D., K. Chan, L. J. Thompson, C. C. Kim, G. R. Govoni, and D. M. Monack.** 2006. Genome-wide screen for Salmonella genes required for long-term systemic infection of the mouse. PLoS Pathog **2:**e11.
62. **Loferer, H., M. Hammar, and S. Normark.** 1997. Availability of the fibre subunit CsgA and the nucleator protein CsgB during assembly of fibronectin-binding curli is limited by the intracellular concentration of the novel lipoprotein CsgG. Mol Microbiol **26:**11-23.
63. **Macnab, R. M.** 2003. How bacteria assemble flagella. Annu Rev Microbiol **57:**77-100.
64. **Matsuzaki, M., Y. Kiso, I. Yamamoto, and T. Satoh.** 2000. Gene disruption analysis of DppA isolated as a periplasmic molecular chaperone-like protein for folding of dimethyl sulfoxide reductase in Rhodobacter sphaeroides f. sp. denitrificans. FEMS Microbiol Lett **193:**223-9.
65. **Matsuzaki, M., Y. Kiso, I. Yamamoto, and T. Satoh.** 1998. Isolation of a periplasmic molecular chaperone-like protein of Rhodobacter sphaeroides f. sp. denitrificans that is homologous to the dipeptide transport protein DppA of Escherichia coli. J Bacteriol **180:**2718-22.
66. **McArthur, S. D., S. C. Pulvermacher, and G. V. Stauffer.** 2006. The Yersinia pestis gcvB gene encodes two small regulatory RNA molecules. BMC Microbiol **6:**52.
67. **Neidhardt, F. C., R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low Jr., B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger.** 1996. Escherichia coli and Salmonella: Cellular and Molecular Biology, 2nd ed, vol. ASM Press, Washington, D.C.

68. **Nenninger, A. A., L. S. Robinson, and S. J. Hultgren.** 2009. Localized and efficient curli nucleation requires the chaperone-like amyloid assembly protein CsgF. Proc Natl Acad Sci U S A **106:**900-5.
69. **Ogasawara, H., A. Hasegawa, E. Kanda, T. Miki, K. Yamamoto, and A. Ishihama.** 2007. Genomic SELEX search for target promoters under the control of the PhoQP-RstBA signal relay cascade. J Bacteriol **189:**4791-9.
70. **Ogasawara, H., K. Yamada, A. Kori, K. Yamamoto, and A. Ishihama.** 2010. Regulation of the E. coli csgD Promoter: Interplay between Five Transcription Factors. Microbiology.
71. **Olsen, A., A. Arnqvist, M. Hammar, S. Sukupolvi, and S. Normark.** 1993. The RpoS sigma factor relieves H-NS-mediated transcriptional repression of csgA, the subunit gene of fibronectin-binding curli in Escherichia coli. Mol Microbiol **7:**523-36.
72. **Olsen, A., H. Herwald, M. Wikstrom, K. Persson, E. Mattsson, and L. Bjorck.** 2002. Identification of two protein-binding and functional regions of curli, a surface organelle and virulence determinant of Escherichia coli. J Biol Chem **277:**34568-72.
73. **Olsen, A., A. Jonsson, and S. Normark.** 1989. Fibronectin binding mediated by a novel class of surface organelles on Escherichia coli. Nature **338:**652-5.
74. **Olsen, A., M. J. Wick, M. Morgelin, and L. Bjorck.** 1998. Curli, fibrous surface proteins of Escherichia coli, interact with major histocompatibility complex class I molecules. Infect Immun **66:**944-9.
75. **Pawar, D. M., M. L. Rossman, and J. Chen.** 2005. Role of curli fimbriae in mediating the cells of enterohaemorrhagic Escherichia coli to attach to abiotic surfaces. J Appl Microbiol **99:**418-25.
76. **Persson, K., W. Russell, M. Morgelin, and H. Herwald.** 2003. The conversion of fibrinogen to fibrin at the surface of curliated Escherichia coli bacteria leads to the generation of proinflammatory fibrinopeptides. J Biol Chem **278:**31884-90.
77. **Pintar, J., and W. T. Starmer.** 2003. The costs and benefits of killer toxin production by the yeast Pichia kluyveri. Antonie Van Leeuwenhoek **83:**89-97.
78. **Pratt, L. A., and T. J. Silhavy.** 1998. Crl stimulates RpoS activity during stationary phase. Mol Microbiol **29:**1225-36.
79. **Prigent-Combaret, C., G. Prensier, T. T. Le Thi, O. Vidal, P. Lejeune, and C. Dorel.** 2000. Developmental pathway for biofilm formation in curli-producing Escherichia coli strains: role of flagella, curli and colanic acid. Environ Microbiol **2:**450-64.
80. **Pryke, S. R., S. Andersson, and M. J. Lawes.** 2001. Sexual selection of multiple handicaps in the red-collared widowbird: female choice of tail length but not carotenoid display. Evolution **55:**1452-63.
81. **Pulvermacher, S. C., L. T. Stauffer, and G. V. Stauffer.** 2009. Role of the Escherichia coli Hfq protein in GcvB regulation of oppA and dppA mRNAs. Microbiology **155:**115-23.

82. **Pulvermacher, S. C., L. T. Stauffer, and G. V. Stauffer.** 2008. The role of the small regulatory RNA GcvB in GcvB/mRNA posttranscriptional regulation of oppA and dppA in Escherichia coli. FEMS Microbiol Lett **281:**42-50.
83. **Pulvermacher, S. C., L. T. Stauffer, and G. V. Stauffer.** 2009. Role of the sRNA GcvB in regulation of cycA in Escherichia coli. Microbiology **155:**106-14.
84. **Pulvermacher, S. C., L. T. Stauffer, and G. V. Stauffer.** 2009. The small RNA GcvB regulates sstT mRNA expression in Escherichia coli. J Bacteriol **191:**238-48.
85. **Remaut, H., C. Bompard-Gilles, C. Goffin, J. M. Frere, and J. Van Beeumen.** 2001. Structure of the Bacillus subtilis D-aminopeptidase DppA reveals a novel self-compartmentalizing protease. Nat Struct Biol **8:**674-8.
86. **Rey, S., M. Acab, J. L. Gardy, M. R. Laird, K. deFays, C. Lambert, and F. S. Brinkman.** 2005. PSORTdb: a protein subcellular localization database for bacteria. Nucleic Acids Res **33:**D164-8.
87. **Richarme, G., and T. D. Caldas.** 1997. Chaperone properties of the bacterial periplasmic substrate-binding proteins. J Biol Chem **272:**15607-12.
88. **Robinson, L. S., E. M. Ashman, S. J. Hultgren, and M. R. Chapman.** 2006. Secretion of curli fibre subunits is mediated by the outer membrane-localized CsgG protein. Mol Microbiol **59:**870-81.
89. **Romling, U.** 2005. Characterization of the rdar morphotype, a multicellular behaviour in Enterobacteriaceae. Cell Mol Life Sci **62:**1234-46.
90. **Romling, U., Z. Bian, M. Hammar, W. D. Sierralta, and S. Normark.** 1998. Curli fibers are highly conserved between Salmonella typhimurium and Escherichia coli with respect to operon structure and regulation. J Bacteriol **180:**722-31.
91. **Romling, U., W. D. Sierralta, K. Eriksson, and S. Normark.** 1998. Multicellular and aggregative behaviour of Salmonella typhimurium strains is controlled by mutations in the agfD promoter. Mol Microbiol **28:**249-64.
92. **Ryu, J. H., and L. R. Beuchat.** 2005. Biofilm formation by Escherichia coli O157:H7 on stainless steel: effect of exopolysaccharide and Curli production on its resistance to chlorine. Appl Environ Microbiol **71:**247-54.
93. **Ryu, J. H., H. Kim, J. F. Frank, and L. R. Beuchat.** 2004. Attachment and biofilm formation on stainless steel by Escherichia coli O157:H7 as affected by curli production. Lett Appl Microbiol **39:**359-62.
94. **Saldana, Z., J. Xicohtencatl-Cortes, F. Avelino, A. D. Phillips, J. B. Kaper, J. L. Puente, and J. A. Giron.** 2009. Synergistic role of curli and cellulose in cell adherence and biofilm formation of attaching and effacing Escherichia coli and identification of Fis as a negative regulator of curli. Environ Microbiol **11:**992-1006.
95. **Sjobring, U., G. Pohl, and A. Olsen.** 1994. Plasminogen, absorbed by Escherichia coli expressing curli or by Salmonella enteritidis expressing thin aggregative fimbriae, can be activated by simultaneously captured tissue-type plasminogen activator (t-PA). Mol Microbiol **14:**443-52.
96. **Solomon, E. B., B. A. Niemira, G. M. Sapers, and B. A. Annous.** 2005. Biofilm formation, cellulose production, and curli biosynthesis by Salmonella originating from produce, animal, and clinical sources. J Food Prot **68:**906-12.

97. **Sone, M., S. Kishigami, T. Yoshihisa, and K. Ito.** 1997. Roles of disulfide bonds in bacterial alkaline phosphatase. J Biol Chem **272:**6174-8.
98. **Stauffer, G. V., L. T. Stauffer, and M. D. Plamann.** 1989. The Salmonella typhimurium glycine cleavage enzyme system. Mol Gen Genet **220:**154-6.
99. **Stauffer, L. T., S. J. Fogarty, and G. V. Stauffer.** 1994. Characterization of the Escherichia coli gcv operon. Gene **142:**17-22.
100. **Stauffer, L. T., M. D. Plamann, and G. V. Stauffer.** 1986. Cloning and characterization of the glycine-cleavage enzyme system of Escherichia coli. Gene **44:**219-26.
101. **Stauffer, L. T., and G. V. Stauffer.** 2005. GcvA interacts with both the alpha and sigma subunits of RNA polymerase to activate the Escherichia coli gcvB gene and the gcvTHP operon. FEMS Microbiol Lett **242:**333-8.
102. **Sukupolvi, S., R. G. Lorenz, J. I. Gordon, Z. Bian, J. D. Pfeifer, S. J. Normark, and M. Rhen.** 1997. Expression of thin aggregative fimbriae promotes interaction of Salmonella typhimurium SR-11 with mouse small intestinal epithelial cells. Infect Immun **65:**5320-5.
103. **Szabo, E., A. Skedsmo, A. Sonnevend, K. Al-Dhaheri, L. Emody, A. Usmani, and T. Pal.** 2005. Curli expression of enterotoxigenic Escherichia coli. Folia Microbiol (Praha) **50:**40-6.
104. **Torres, A. G., C. Jeter, W. Langley, and A. G. Matthysse.** 2005. Differential binding of Escherichia coli O157:H7 to alfalfa, human epithelial cells, and plastic is mediated by a variety of surface structures. Appl Environ Microbiol **71:**8008-15.
105. **Tseng, T. T., B. M. Tyler, and J. C. Setubal.** 2009. Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. BMC Microbiol **9 Suppl 1:**S2.
106. **Tukel, C., J. H. Nishimori, R. P. Wilson, M. G. Winter, A. M. Keestra, J. P. van Putten, and A. J. Baumler.** 2010. Toll-like receptors 1 and 2 cooperatively mediate immune responses to curli, a common amyloid from enterobacterial biofilms. Cell Microbiol.
107. **Tukel, C., M. Raffatellu, A. D. Humphries, R. P. Wilson, H. L. Andrews-Polymenis, T. Gull, J. F. Figueiredo, M. H. Wong, K. S. Michelsen, M. Akcelik, L. G. Adams, and A. J. Baumler.** 2005. CsgA is a pathogen-associated molecular pattern of Salmonella enterica serotype Typhimurium that is recognized by Toll-like receptor 2. Mol Microbiol **58:**289-304.
108. **Tukel, C., R. P. Wilson, J. H. Nishimori, M. Pezeshki, B. A. Chromy, and A. J. Baumler.** 2009. Responses to amyloids of microbial and host origin are mediated through toll-like receptor 2. Cell Host Microbe **6:**45-53.
109. **Tzfira, T., and V. Citovsky.** 2006. Agrobacterium-mediated genetic transformation of plants: biology and biotechnology. Curr Opin Biotechnol **17:**147-54.
110. **Tzfira, T., and V. Citovsky.** 2002. Partners-in-infection: host proteins involved in the transformation of plant cells by Agrobacterium. Trends Cell Biol **12:**121-9.

111. **Uhlich, G. A., P. H. Cooke, and E. B. Solomon.** 2006. Analyses of the red-dry-rough phenotype of an Escherichia coli O157:H7 strain and its role in biofilm formation and resistance to antibacterial agents. Appl Environ Microbiol **72:**2564-72.
112. **Uhlich, G. A., N. W. t. Gunther, D. O. Bayles, and D. A. Mosier.** 2009. The CsgA and Lpp proteins of an Escherichia coli O157:H7 strain affect HEp-2 cell invasion, motility, and biofilm formation. Infect Immun **77:**1543-52.
113. **Uhlich, G. A., J. E. Keen, and R. O. Elder.** 2002. Variations in the csgD promoter of Escherichia coli O157:H7 associated with increased virulence in mice and increased invasion of HEp-2 cells. Infect Immun **70:**395-9.
114. **Urbanowski, M. L., L. T. Stauffer, and G. V. Stauffer.** 2000. The gcvB gene encodes a small untranslated RNA involved in expression of the dipeptide and oligopeptide transport systems in Escherichia coli. Mol Microbiol **37:**856-68.
115. **Vianney, A., G. Jubelin, S. Renault, C. Dorel, P. Lejeune, and J. C. Lazzaroni.** 2005. Escherichia coli tol and rcs genes participate in the complex network affecting curli synthesis. Microbiology **151:**2487-97.
116. **Vidal, O., R. Longin, C. Prigent-Combaret, C. Dorel, M. Hooreman, and P. Lejeune.** 1998. Isolation of an Escherichia coli K-12 mutant strain able to form biofilms on inert surfaces: involvement of a new ompR allele that increases curli expression. J Bacteriol **180:**2442-9.
117. **Wang, X., and M. R. Chapman.** 2008. Curli provide the template for understanding controlled amyloid propagation. Prion **2:**57-60.
118. **Wang, X., and M. R. Chapman.** 2008. Sequence determinants of bacterial amyloid formation. J Mol Biol **380:**570-80.
119. **Wang, X., M. Rochon, A. Lamprokostopoulou, H. Lunsdorf, M. Nimtz, and U. Romling.** 2006. Impact of biofilm matrix components on interaction of commensal Escherichia coli with the gastrointestinal cell line HT-29. Cell Mol Life Sci **63:**2352-63.
120. **Wang, X., D. R. Smith, J. W. Jones, and M. R. Chapman.** 2007. In Vitro Polymerization of a Functional Escherichia coli Amyloid Protein. J Biol Chem **282:**3713-9.
121. **Wang, X., Y. Zhou, J. J. Ren, N. D. Hammer, and M. R. Chapman.** 2010. Gatekeeper residues in the major curlin subunit modulate bacterial amyloid fiber biogenesis. Proc Natl Acad Sci U S A **107:**163-8.
122. **White, A. P., D. L. Gibson, W. Kim, W. W. Kay, and M. G. Surette.** 2006. Thin aggregative fimbriae and cellulose enhance long-term survival and persistence of Salmonella. J Bacteriol **188:**3219-27.
123. **White, A. P., and M. G. Surette.** 2006. Comparative genetics of the rdar morphotype in Salmonella. J Bacteriol **188:**8395-406.
124. **Wilson, R. L., and G. V. Stauffer.** 1994. DNA sequence and characterization of GcvA, a LysR family regulatory protein for the Escherichia coli glycine cleavage enzyme system. J Bacteriol **176:**2862-8.

125. **Wilson, R. L., L. T. Stauffer, and G. V. Stauffer.** 1993. Roles of the GcvA and PurR proteins in negative regulation of the Escherichia coli glycine cleavage enzyme system. J Bacteriol **175:**5129-34.
126. **Wilson, R. L., M. L. Urbanowski, and G. V. Stauffer.** 1995. DNA binding sites of the LysR-type regulator GcvA in the gcv and gcvA control regions of Escherichia coli. J Bacteriol **177:**4940-6.
127. **Wonderling, L. D., M. L. Urbanowski, and G. V. Stauffer.** 2000. GcvA binding site 1 in the gcvTHP promoter of Escherichia coli is required for GcvA-mediated repression but not for GcvA-mediated activation. Microbiology **146 (Pt 11):**2909-18.
128. **Zakikhany, K., C. R. Harrington, M. Nimtz, J. C. Hinton, and U. Romling.** 2010. Unphosphorylated CsgD controls biofilm formation in Salmonella enterica serovar Typhimurium. Mol Microbiol.
129. **Zhang, X., Z. W. El-Hajj, and E. Newman.** 2010. DEFICIENCY IN L-SERINE DEAMINASE INTERFERES WITH ONE-CARBON METABOLISM AND CELL WALL SYNTHESIS IN E. coli K-12. J Bacteriol.
130. **Zhang, X., and E. Newman.** 2008. Deficiency in l-serine deaminase results in abnormal growth and cell division of Escherichia coli K-12. Mol Microbiol **69:**870-81.
131. **Zogaj, X., W. Bokranz, M. Nimtz, and U. Romling.** 2003. Production of cellulose and curli fimbriae by members of the family Enterobacteriaceae isolated from the human gastrointestinal tract. Infect Immun **71:**4151-8.

Appendix A: Perl programs

QNPDE5020c.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @SEQ=(<>);
my $source=$ARGV; $source=~s/\..../;
open FILE, ">$source-50aa.txt" or die "failed to clear";
print FILE "QN-PDE,Sequence,Amino Acids,Description,Location\n";
close FILE;
foreach (@SEQ) {
	my @spl = split('\:,');
	foreach ($spl[1]) {
	my @array=split '';
	my $stuff=scalar(@array);
	my $place=0;
	START:
	open FILE, ">>$source-50aa.txt" or die "failed to open";
	$stuff=scalar(@array);
	if ($stuff<=50){goto END;
	}else{
		my @try=join('',@array[0..49]);
			foreach (@try) {
				$place++;
				my $last=$place+49;
				my $q=0;
				my $n=0;
				my $p=0;
				my $d=0;
				my $e=0;
				my $qnpde=0;
				$q=(tr/Q//);
				$n=(tr/N//);
				$p=(tr/P//);
				$d=(tr/D//);
				$e=(tr/E//);
				$qnpde=$q+$n-$p-$d-$e;
				print FILE $qnpde,",",@try,",",$place," to ",$last,",",$spl[0],"\n";
				shift @array;}
		goto START;}}
END:}
close FILE;
open FILE2, ">$source-20aa.txt" or die "failed to clear";
print FILE2 "QN-PDE,Sequence,Amino Acids,Description,Location\n";
close FILE2;

foreach (@SEQ) {
	my @spl = split('\:,');
	foreach ($spl[1]) {
	my @array=split '';
	my $stuff=scalar(@array);
	my $place=0;
	START:
```

QNPDE5020c.perl (continued)

```
        open FILE2, ">>$source-20aa.txt" or die "failed to open";
        $stuff=scalar(@array);
        if ($stuff<=20){goto END;
        }else{
                my @try=join('',@array[0..19]);
                        foreach (@try) {
                                $place++;
                                my $last=$place+19;
                                my $q=0;
                                my $n=0;
                                my $p=0;
                                my $d=0;
                                my $e=0;
                                my $qnpde=0;
                                $q=(tr/Q//);
                                $n=(tr/N//);
                                $p=(tr/P//);
                                $d=(tr/D//);
                                $e=(tr/E//);
                                $qnpde=$q+$n-$p-$d-$e;
                                print FILE2 $qnpde,",",@try,",",$place," to ",$last,",",$spl[0],"\n";
                                shift @array;}
                goto START;}}
END:}
```

Rationale: Possible functional amyloid prediction. This program will take a file of multiple protein sequences and for each protein analyze the number of Q,N,P,D,E residues in 50 and 20 amino acid windows. There must be a colon (:) between descriptors and the sequence on each line. Each Q and N adds plus one, each beta breaker amino acid (PDE) is -1.

Input: Simple text file "Each line is 'description, location,:, sequence "

Output: Inputfilename-20aa.txt 20aa window
Inputfilename-50aa.txt 50aa window

Notes: Output files can be opened in Excel using comma (,) delimited data. If not using location another descriptor could be used. This program will also work on a single protein as well.

QNPDE5020.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @SEQ=(<>);
my $source=$ARGV; $source=~s/\..../;
open FILE, ">$source-50aa.txt" or die "failed to clear";
print FILE "QN-PDE,Sequence,Amino Acids,Description,Location\n";
close FILE;
foreach (@SEQ) {
	my @spl = split('\',');
	foreach ($spl[1]) {
	my @array=split '';
	my $stuff=scalar(@array);
	my $place=0;
	START:
	open FILE, ">>$source-50aa.txt" or die "failed to open";
	$stuff=scalar(@array);
	if ($stuff<=50){goto END;
	}else{
		my @try=join('',@array[0..49]);
			foreach (@try) {
				$place++;
				my $last=$place+49;
				my $q=0;
				my $n=0;
				my $p=0;
				my $d=0;
				my $e=0;
				my $qnpde=0;
				$q=(tr/Q//);
				$n=(tr/N//);
				$p=(tr/P//);
				$d=(tr/D//);
				$e=(tr/E//);
				$qnpde=$q+$n-$p-$d-$e;
				print FILE $qnpde,",",@try,",",$place," to ",$last,",",$spl[0],"\n";
				shift @array;}
		goto START;}}
END:}
close FILE;
open FILE2, ">$source-20aa.txt" or die "failed to clear";
print FILE2 "QN-PDE,Sequence,Amino Acids,Description,Location\n";
close FILE2;
foreach (@SEQ) {
	my @spl = split('\',');
	foreach ($spl[1]) {
	my @array=split '';
	my $stuff=scalar(@array);
	my $place=0;
	START:
	open FILE2, ">>$source-20aa.txt" or die "failed to open";
```

QNPDE5020.perl (continued)

```
        $stuff=scalar(@array);
        if ($stuff<=20){goto END;
        }else{
                my @try=join('',@array[0..19]);
                        foreach (@try) {
                                $place++;
                                my $last=$place+19;
                                my $q=0;
                                my $n=0;
                                my $p=0;
                                my $d=0;
                                my $e=0;
                                my $qnpde=0;
                                $q=(tr/Q//);
                                $n=(tr/N//);
                                $p=(tr/P//);
                                $d=(tr/D//);
                                $e=(tr/E//);
                                $qnpde=$q+$n-$p-$d-$e;
                                print FILE2 $qnpde,",",@try,",",$place," to ",$last,",",$spl[0],"\n";
                                shift @array;}
                goto START;}}
END:}
```

Rationale: Same usage as QPDE5020.perl except ‘ is used as a divider. However, depending on the source file the symbol (') is read incorrectly as(‘) and the split function does not work. There must be a (') dividing descriptions and sequence. Each Q and N are plus one, each beta breaker amino acid (PDE) is -1. Originally used on simplepos.txt outputs from QNPDEpos50.perl, QNPDEpos20.perl, or QNPDEposPercent.perl.

Input: Simple text file “Each line is ‘description, location, ', sequence ”

Output: Inputfilename-20aa.txt 20aa window
Inputfilename-50aa.txt 50aa window

Notes: Output files can be opened in Excel using comma (,) delimited data. If not using location another description could be used. This program will also work on a single protein as well. Preferably use QNPDE5020c.perl as (:) are rare in fasta file descriptions and (') can cause problems if (‘) is mistakenly used.

QNPDEposPercent.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my $line;
my @before=(<>);
chomp(@before);
foreach (@before) {
s/,//g;
s/->//g;
s/\>/\n\>/g;
s/\|Cytoplasmic /\| Cytoplasmic/g;
s/\|CytoplasmicMembrane/,CytoMembrane,,,/g;
s/\|Cytoplasmic/,Cytoplasmic,,,/g;
s/\|Unknown \(This protein may have multiple localization sites\.\)/,UnknowMultiple,,,/g;
s/\|Unknown /\| Unknown /g;
s/\|Unknown/,Unknown,,,/g;
s/\|Cellwall /\| Cellwall /g;
s/\|Cellwall/,Cellwall,,,/g;
s/\|Extracellular /\| Extracellular /g;
s/\|Extracellular/,Extracellular,,,/g;
}
my $file="simplepos.txt";
my @file2="scoredpos.txt";
open FILE, ">simplepos.txt" or die "unable to open $file $!";
print FILE @before;
close FILE;
open FILE, "simplepos.txt" or die "this didnt work";
my @after=<FILE>;
close FILE;
shift @after;
open FILE, ">simplepos.txt" or die "this didnt work";
print FILE @after;
close FILE;
open FILE, "<simplepos.txt" or die "this didnt work either";
my @c=<FILE>;
close FILE;
open FILE2, ">scoredpos.txt" or die "failed to open";
print FILE2 "Description,Location,Number Q,Number N,Number P,Number D,Number E,Total AA,Percent
QN-P,Percent QN-PD,Percent QN-PE,Percent QN-PDE,Sequence\n";
foreach (@c) {
open FILE2, ">>scoredpos.txt" or die "failed on open again";
my @spl = split(',,');
print FILE2 $spl[0];
foreach ($spl[1]) {
my $q=(tr/Q//);
my $n=(tr/N//);
my $p=(tr/P//);
my $d=(tr/D//);
my $e=(tr/E//);
my $s=(s/(.)/$1/sg);
$s=$s-2;
```

QNPDEposPercent.perl (continued)

```
my $qnp=(100*($q+$n-$p)/$s);
my $qnpd=(100*($q+$n-$p-$d)/$s);
my $qnpe=(100*($q+$n-$p-$e)/$s);
my $qnpde=(100*($q+$n-$p-$d-$e)/$s);
print FILE2 ",",$q,",",$n,",",$p,",",$d,",",$e,",",$s,",",$qnp,",",$qnpd,",",$qnpe,",",$qnpde,$spl[1];}
}
```

Rationale: An older way of predicting functional amyloids. This program will take a Gram-positive organismal.faa file from pSORTdb v3 and then output the %QN-P, %QN-PD, and % QN-PDE. Several functional amyloids rank highly when adding the percent of Gln and Asn and subtracting the percent of beta breaker residues such as Pro, Asp, and Glu .

Input: An organism.faa file from pSORTdb v3 Not tested on new pSORTdb v4

Output: simplepos.txt comma delimited text file of description,location,sequence
scorepos.txt simplepos.txt plus % of single amino acids & combinations

Notes: Must be a Gram-positive organism as the first part of the program converts the organism.faa file to a comma delimited text file. The output file is readable using Excel and is comma delimited.

QNPDEpos50.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @before=(<>);
chomp(@before);
foreach (@before) {
	s/,//g;
	s/->//g;
	s/\>/\n\>/g;
	s/\|Cytoplasmic /\| Cytoplasmic/g;
	s/\|CytoplasmicMembrane/,CytoMembrane,'\,/g;
	s/\|Cytoplasmic/,Cytoplasmic,\',/g;
	s/\|Unknown \(This protein may have multiple localization sites\.\)/,UnknowMultiple,\',/g;
	s/\|Unknown /\| Unknown /g;
	s/\|Unknown/,Unknown,\',/g;
	s/\|Cellwall /\| Cellwall /g;
	s/\|Cellwall/,Cellwall,\',/g;
	s/\|Extracellular /\| Extracellular /g;
	s/\|Extracellular/,Extracellular,\',/g;
}
my $file="simpleseq.txt";
my @file2="rollingpos.txt";
open FILE, ">simpleseq.txt" or die "unable to open $file $!";
print FILE @before;
close FILE;
open FILE, "simpleseq.txt" or die "this didnt work";
my @after=<FILE>;
close FILE;
shift @after;
open FILE, ">simpleseq.txt" or die "this didnt work";
print FILE @after;
close FILE;
open FILE, "<simpleseq.txt" or die "this didnt work either";
my @in=<FILE>;
close FILE;
open FILE2, ">rollingpos.txt" or die "failed to clear";
print FILE2 "QN-PDE,Sequence,Amino Acids,Description,Location\n";
close FILE2;
foreach (@in) {
	my @spl = split('\',');
	foreach ($spl[1]) {
	my @array=split '';
	my $stuff=scalar(@array);
	my $place=0;
	START:
	open FILE2, ">>rollingpos.txt" or die "failed to open";
	$stuff=scalar(@array);
	if ($stuff<=50){goto END;
	}else{
		my @try=join('',@array[0..49]);
			foreach (@try) {
```

QNPDEpos50.perl (continued)

```
                                        $place++;
                                        my $last=$place+49;
                                        my $q=0;
                                        my $n=0;
                                        my $p=0;
                                        my $d=0;
                                        my $e=0;
                                        my $qnpde=0;
                                        $q=(tr/Q//);
                                        $n=(tr/N//);
                                        $p=(tr/P//);
                                        $d=(tr/D//);
                                        $e=(tr/E//);
                                        $qnpde=$q+$n-$p-$d-$e;
                                        print FILE2 $qnpde,",",@try,",",$place," to ",$last,",",$spl[0],"\n";
                                        shift @array;}
                        goto START;}}
END:}
```

Rationale: Same usage as QPDE5020c.perl but only outputs 50 aa window and must uses a Gram-positive organism.faa file from pSORTdb v3. Each Q and N are plus one, each beta breaker amino acid (PDE) is -1.

Input: An organism.faa file from pSORTdb v3 Not tested on new pSORTdb v4

Output: simplepos.txt comma delimited text file of description,location,sequence
rollingpos.txt Score for 50 aa window

Notes: Must be a Gram-positive organism as the first part of the program converts the organism.faa file to a comma delimited text file. The output file is readable using Excel and is comma delimited.

QNPDEpos20.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @before=(<>);
chomp(@before);
foreach (@before) {
	s/,//g;
	s/->//g;
	s/\>/\n\>/g;
	s/\|Cytoplasmic /\| Cytoplasmic/g;
	s/\|CytoplasmicMembrane/,CytoMembrane,'\,/g;
	s/\|Cytoplasmic/,Cytoplasmic,\',/g;
	s/\|Unknown \(This protein may have multiple localization sites\.\)/,UnknowMultiple,\',/g;
	s/\|Unknown /\| Unknown /g;
	s/\|Unknown/,Unknown,\',/g;
	s/\|Cellwall /\| Cellwall /g;
	s/\|Cellwall/,Cellwall,\',/g;
	s/\|Extracellular /\| Extracellular /g;
	s/\|Extracellular/,Extracellular,\',/g;
}
my $file="simpleseq.txt";
my @file2="rollingpos.txt";
open FILE, ">simpleseq.txt" or die "unable to open $file $!";
print FILE @before;
close FILE;
open FILE, "simpleseq.txt" or die "this didnt work";
my @after=<FILE>;
close FILE;
shift @after;
open FILE, ">simpleseq.txt" or die "this didnt work";
print FILE @after;
close FILE;
open FILE, "<simpleseq.txt" or die "this didnt work either";
my @in=<FILE>;
close FILE;
open FILE2, ">rollingpos.txt" or die "failed to clear";
print FILE2 "QN-PDE,Sequence,Amino Acids,Description,Location\n";
close FILE2;
foreach (@in) {
	my @spl = split('\',');
	foreach ($spl[1]) {
	my @array=split '';
	my $stuff=scalar(@array);
	my $place=0;
	START:
	open FILE2, ">>rollingpos.txt" or die "failed to open";
	$stuff=scalar(@array);
	if ($stuff<=20){goto END;
	}else{
		my @try=join('',@array[0..19]);
			foreach (@try) {
```

QNPDEpos20.perl (continued)

```
                                        $place++;
                                        my $last=$place+19;
                                        my $q=0;
                                        my $n=0;
                                        my $p=0;
                                        my $d=0;
                                        my $e=0;
                                        my $qnpde=0;
                                        $q=(tr/Q//);
                                        $n=(tr/N//);
                                        $p=(tr/P//);
                                        $d=(tr/D//);
                                        $e=(tr/E//);
                                        $qnpde=$q+$n-$p-$d-$e;
                                        print FILE2 $qnpde,",",@try,",",$place," to ",$last,",",$spl[0],"\n";
                                        shift @array;}
                        goto START;}}
END:}
```

Rationale: Same usage as QPDE5020c.perl but only outputs 20 aa window and must uses a Gram-positive organism.faa file from pSORTdb v3. Each Q and N are plus one, each beta breaker amino acid (PDE) is -1.

Input: An organism.faa file from pSORTdb v3 Not tested on new pSORTdb v4

Output: simplepos.txt comma delimited text file of description,location,sequence
rollingpos.txt Score for 20 aa window

Notes: Must be a Gram-positive organism as the first part of the program converts the organism.faa file to a comma delimited text file. The output file is readable using Excel and is comma delimited.

QNPDEone.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @before=(<>);
my $source=$ARGV; $source=~s/\..../;
open FILE, ">$source-QNPDE.txt" or die "failed to clear";
print FILE "QN-PDE,Sequence,Amino Acids,Description,Location\n";
close FILE;
foreach (@before) {
	my @spl = split('\',');
	foreach ($spl[1]) {
	my @array=split '';
	my $stuff=scalar(@array);
	my $place=0;
	START:
	open FILE, ">>$source-QNPDE.txt" or die "failed to open";
	$stuff=scalar(@array);
	if ($stuff<=20){goto END;
	}else{
		my @try=join('',@array[0..19]);
			foreach (@try) {
				$place++;
				my $last=$place+19;
				my $q=0;
				my $n=0;
				my $p=0;
				my $d=0;
				my $e=0;
				my $qnpde=0;
				$q=(tr/Q//);
				$n=(tr/N//);
				$p=(tr/P//);
				$d=(tr/D//);
				$e=(tr/E//);
				$qnpde=$q+$n-$p-$d-$e;
				print FILE $qnpde,",",@try,",",$place," to ",$last,",",$spl[0],"\n";
				shift @array;}
		goto START;}}
END:}
```

Rationale: Possible functional amyloid prediction. Takes a single protein and counts the number of Q,N,P,D, and E residues in a 20 amino acid window. Each Q and N are plus one, each beta breaker amino acid (PDE) is -1.

Input: Simple text file "Each line is 'description, location, ', sequence "

Output: Inputfilename-QNPDE.txt

Notes: Must divide description and location with a (') not a (').

QNPDEnegPercent.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my $line;
my @before=(<>);
chomp(@before);
foreach (@before) {
s/,//g;
s/->//g;
s/\>/\n\>/g;
s/\|Cytoplasmic /\| Cytoplasmic/g;
s/\|CytoplasmicMembrane/,CytoMembrane,,,/g;
s/\|Cytoplasmic/,Cytoplasmic,,,/g;
s/\|Unknown \(This protein may have multiple localization sites\.\)/,UnknowMultiple,,,/g;
s/\|Unknown /\| Unknown /g;
s/\|Unknown/,Unknown,,,/g;
s/\|OuterMembrane/,OuterMembrane,,,/g;
s/\|Periplasmic /\| Periplasmic /g;
s/\|Periplasmic/,Periplasmic,,,/g;
s/\|Extracellular /\| Extracellular /g;
s/\|Extracellular/,Extracellular,,,/g;
}
my $file="simpleneg.txt";
my @file2="scoredneg.txt";
open FILE, ">simpleneg.txt" or die "unable to open $file $!";
print FILE @before;
close FILE;
open FILE, "simpleneg.txt" or die "this didnt work";
my @after=<FILE>;
close FILE;
shift @after;
open FILE, ">simpleneg.txt" or die "this didnt work";
print FILE @after;
close FILE;
open FILE, "<simpleneg.txt" or die "this didnt work either";
my @c=<FILE>;
close FILE;
open FILE2, ">scoredneg.txt" or die "failed to open";
print FILE2 "Description,Location,Number Q,Number N,Number P,Number D,Number E,Total AA,Percent
QN-P,Percent QN-PD,Percent QN-PE,Percent QN-PDE,Sequence\n";
foreach (@c) {
open FILE2, ">>scoredneg.txt" or die "failed on open again";
my @spl = split(',,');
print FILE2 $spl[0];
foreach ($spl[1]) {
my $q=(tr/Q//);
my $n=(tr/N//);
my $p=(tr/P//);
my $d=(tr/D//);
my $e=(tr/E//);
my $s=(s/(.)/$1/sg);
```

QNPDEnegPercent.perl (continued)

```
$s=$s-2;
my $qnp=(100*($q+$n-$p)/$s);
my $qnpd=(100*($q+$n-$p-$d)/$s);
my $qnpe=(100*($q+$n-$p-$e)/$s);
my $qnpde=(100*($q+$n-$p-$d-$e)/$s);
print FILE2 ",",$q,",",$n,",",$p,",",$d,",",$e,",",$s,",",$qnp,",",$qnpd,",",$qnpe,",",$qnpde,$spl[1];}
}
```

Rationale: An older way of predicting functional amyloids. This program will take a Gram-negative organismal.faa file from pSORTdb v3 and then output the %QN-P, %QN-PD, % QN-PDE. Several functional amyloids rank highly when adding the percent of Gln and Asn and subtracting the percent of beta breaker residues such as Pro, Asp, and Glu .

Input: An organism.faa file from pSORTdb v3 Not tested on new pSORTdb v4

Output: simpleneg.txt comma delimited text file of description,location,sequence
scoreneg.txt simpleneg.txt plus % of single amino acids & combinations

Notes: Must be a Gram-negative organism as the first part of the program converts the organism.faa file to a comma delimited text file. The output file is readable using Excel and is comma delimited.

QNPDE50.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @SEQ=(<>);
my $source=$ARGV; $source=~s/\..../;
open FILE, ">$source-50aa.txt" or die "failed to clear";
print FILE "QN-PDE,Sequence,Amino Acids,Description,Location\n";
close FILE;
foreach (@SEQ) {
        my @spl = split('\',');
        foreach ($spl[1]) {
        my @array=split '';
        my $stuff=scalar(@array);
        my $place=0;
        START:
        open FILE, ">>$source-50aa.txt" or die "failed to open";
        $stuff=scalar(@array);
        if ($stuff<=50){goto END;
        }else{
                my @try=join('',@array[0..49]);
                        foreach (@try) {
                                $place++;
                                my $last=$place+49;
                                my $q=0;
                                my $n=0;
                                my $p=0;
                                my $d=0;
                                my $e=0;
                                my $qnpde=0;
                                $q=(tr/Q//);
                                $n=(tr/N//);
                                $p=(tr/P//);
                                $d=(tr/D//);
                                $e=(tr/E//);
                                $qnpde=$q+$n-$p-$d-$e;
                                print FILE $qnpde,",",@try,",",$place," to ",$last,",",$spl[0],"\n";
                                shift @array;}
                goto START;}}
END:}
```

Rationale: Same usage as QPDE5020.perl except ‘ is used as a divider and only 50 aa windows are used.

Input: Simple text file “Each line is ‘description, location, ', sequence ”

Output: Inputfilename-50aa.txt 50aa window

Notes: Output files can be opened in Excel using comma (,) delimited data. If not using location another description could be used. Must have a (') not a (‘).

Sorter.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @sorting=(<>);
my @file="sortpos.txt";
my @sorted=shift @sorting;
@sorted=sort @sorting;
open FILE, ">sortpos.txt";
print FILE @sorting
```

Rationale: Used to sort large output files from the QNPDE amino acid window programs which can result in several hundred thousand to over a million lines. Program should also work on Gram-negative organisms too

Input: output file from 20 or 50 amino acid amyloid prediction program

Output: sortpos.txt

Findreplace.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @file=(<>);
foreach (@file) {
s/Staphylococcus aureus subsp\. aureus str\. Newman complete genome/Newman/g;}
my $replace="short".$ARGV;
open(FILE2,">$replace");
print FILE2 @file;
```

Rationale: Used to remove extraneous description from large output files from programs that calculate the number of QNPDE over a 20 or 50 amino acid window. File reduction allows easier manipulation in Excel or Word.

Input: Simple text file

Output: short.inputfilename

Notes: Much quicker than using Word or other search and replace programs and should not crash as easily. Easily modified to search and replace other text items. Change bold text to what needs to be found and the italic text to the replacing text. If either text uses punctuation you must add a \ before. For example “subsp. aureus” would be “subsp\. aureus”.

Disopred.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @one=0;
my @result=0;
my @before=(<>);
my $file="sequence.txt";
my $file2="EB.txt";
my $file3="all.txt";
foreach (@before) {
my @one=split/\t/;
open FILE, ">sequence.txt" or die "Failed to open sequence.txt";
print FILE $one[0],"\n",$one[1]; close FILE;
system("./rundisopred sequence.txt 2") == 0 or die "Failed on running DISOPRED";
open FILE, "sequence.horiz_d";
@result=<FILE>; close FILE;
open FILE2, "$one[0].txt" or die "Failed to open EB file";
print FILE2 $one[0],"\n",@result; close FILE2;
open FILE3, ">>Alldisopred.txt" or die "Failed to open Alldisopred.txt";
print FILE3 $one[0],"\n",@result;    }
```

Rationale: Program to send multiple protein sequence files to Disopred2.

Input: Simple text file. Each line is ">description**[TAB]**sequence".

Output: An individual text file of each protein named after the protein description.
Alldisopred.txt - combines all the individual Disopred2 results.

Notes: I set up Disopred2 to run in Ubuntu. The **2** above sets the false positive rate. See Fixdisopred.perl. Alldisopred.txt needs to be fixed.

An alternative file is called Disopred2.perl which uses a second non redundant database and allows two large files to be worked on at once. The nciblast which is part of the disopred prediction program only runs on one processor. Utilizing a second program should roughly double output. The outputs are the same for both programs; however, output files are named differently so that no file overwriting or combining occurs.

Fixdisopred.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @before=(<>);
my $source=$ARGV; $source=~s/\..../;
my $lines=join('',@before);
open FILE, ">$source.old";
print FILE @before; close FILE;
$lines=~s/(conf: \npred: .{60}\n  AA: .{60}\n.{66})//g;
$lines=~s/(conf: \t\npred: .{60}\n  AA: .{60}\n.{66})//g;
open FILE2, ">$source.new";
print FILE2 $lines; close FILE2;
```

Rationale: I found a bug in Disopred2 vs 2.4. Protein sequences that are a multiples of 60 (60,120,180,240,etc.) amino acids have their last 60 amino acids duplicated in the Disopred2 results. This program strips out the offending lines so they do not affect DIS.perl, DIScost.perl, and further analysis of multiple protein Disopred results. I emailed Daniel Buchan (d.buchan@cs.ucl.ac.uk) about the bug which was corrected in version 2.42.

Input:	Alldisopred.txt	From Disopred.perl
Output:	Alldisopred.old	Alldisopred.txt file backup
	Alldisopred.new	Fixed Alldisopred.txt file suitable for use.

Fixdis.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @before=(<>); my $count=0;
my $source=$ARGV; $source=~s/all\.txt//;
foreach (@before) {
	if($_=~/conf: \n/){
		$_ = "bad"; }
	if($_=~/conf: \t\n/){
		$_ = "bad"; }	}
open FILE, ">temp.txt";	print FILE @before;	close FILE;
open FILE, "<temp.txt";
my @after = <FILE>;	close FILE;
foreach (@after) {
	if($_=~/badpred: /){
		$_ = "bad"; }	}
open FILE, ">temp2.txt"; print FILE @after;	close FILE;
open FILE, "<temp2.txt";
my @final = <FILE>;	close FILE;
foreach (@final) {
	if($_=~/bad  AA: /){
		$count++;
		$_ = ""; }}
open FILE, ">$source.txt";
print FILE @final; close FILE;
unlink "temp.txt";
unlink "temp2.txt";
print "Number of proteins with doubled last lines is ",$count;
```

Rationale: I found a bug in Disopred2 vs 2.4. Protein sequences that are a multiples of 60 (60,120,180,240,etc.) amino acids have their last 60 amino acids duplicated in the Disopred2 results. This program found the number of bad sequences allowing me to figure out what was wrong with Disopred and allowed me to check Fixdisospred.perl to make sure it was correcting the problem.

Input: Alldisopred.txt From Disopred.perl

Output: Prints number of proteins with a duplicated last line.

DIS.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @before=(<>);
my @ss; my @seq;
my @file; my @seqonly; my @final;
my $source=$ARGV; $source=~s/\..../;
foreach(@before) {
        if ($_=~/\>/) {push(@ss,$_,");}
        if ($_=~s/pred\: //) {push(@ss,$_);}}
foreach(@before) {
        if ($_=~/\>/) {push(@seq,$_,");}
        if ($_=~s/  AA\: //) {push(@seq,$_);}          }
chomp (@ss); chomp (@seq);
foreach (@ss) {s/\>/\n\>/g;}
foreach (@seq) {s/\>/\n\>/g;}
open FILE, ">$source-SEQ.txt" or die "unable to save Sequence file";
print FILE @seq; close FILE;
open FILE, ">$source-DS.txt" or die "unable to save Structure (DS) file";
print FILE @ss; close FILE;
open FILE, "<$source-SEQ.txt"; my @seqfile = <FILE>; close FILE;
open FILE, "<$source-DS.txt"; my @ssonly = <FILE>; close FILE;
foreach (@seqfile) {
        if ($_=~s/\>EB\d\d\d\d//) {
                push(@seqonly,$_);} #Modify this pattern if not >EB1245
        if ($_=~s/\>gi\d\d\d\d\d\d\d\d//) {
                push(@seqonly,$_);} #Or add your pattern as another if statement
}
foreach (@seqonly) {
        push(@final,$ssonly[1],$_);
        shift @ssonly;}
chomp @final;
foreach (@final) {s/\>/\n\>/g;}
open FILE, ">$source-final.txt" or die "unable to open Final file";
print FILE @final; close FILE;
```

Rationale: Strips out sequences and Disopred2 results from multiple results in Alldisopred.txt result file and converts them into a comma delimited text file suitable for Excel and DIScost.perl.

Input: Name.txt (Fixed Disopred.perl file using Fixdisopred.perl)

Output: Name-DS.txt Each line is “Description,Disopred2 result”.
Name-SEQ.txt Each line is “Description,sequence”.
Name-final.txt Each line is “Description,Disopred2,sequence”.

Notes: See Fixdisopred.perl. Alldisopred.txt needs to be fixed if using Disopred2 version 2.4.

DisCOST.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @before=(<>);
my @all; my @smp; my @splA; my @splB;
my @A; my @B; my @arrA; my @arrB;
my @final; my $line;
my $source=$ARGV; $source=~s/\-final\..../;
foreach (@before) {
	s/\>EB\d\d\d\d\,//; #Modify this pattern if not >EB1245
	s/\>gi\d\d\d\d\d\d\d\,//; # or add your pattern as another substituion
	@all=@before; }
shift @all;
chomp @all;
foreach (@all) {
	@smp=split ',';
	push (@A,$smp[0]); #@A appears chomped
	push (@B,$smp[1]);	}
foreach (@A) {@splA=split '';
	foreach (@splA) {push (@arrA,$_);}}
foreach (@B) {@splB=split '';
	foreach (@splB) {push (@arrB,$_);}}
foreach (@arrA) {
	$line=join('',$_,$arrB[0]);
	shift @arrB;
	push (@final,$line,","); }
open FILE, ">$source-DSaaRAW.txt" or die "failed to open DSaaRAW file";
print FILE @final; close FILE;
my $DSA=my $DSC=my $DSD=my $DSE=my $DSF=my $DSG=my $DSH=my $DSI=my $DSK=my
$DSL=my $DSM=my $DSN=my $DSP=my $DSQ=my $DSR=my $DSS=my $DST=my $DSV=my $DSW=my
$DSY=0;
my $ORA=my $ORC=my $ORD=my $ORE=my $ORF=my $ORG=my $ORH=my $ORI=my $ORK=my
$ORL=my $ORM=my $ORN=my $ORP=my $ORQ=my $ORR=my $ORS=my $ORT=my $ORV=my $ORW=my
$ORY=0;
foreach (@final) {
	if ($_=~/\*A/) {$DSA++;}	if ($_=~/\*C/) {$DSC++;}	if ($_=~/\*D/) {$DSD++;}
	if ($_=~/\*E/) {$DSE++;}	if ($_=~/\*F/) {$DSF++;}	if ($_=~/\*G/) {$DSG++;}
	if ($_=~/\*H/) {$DSH++;}	if ($_=~/\*I/) {$DSI++;}	if ($_=~/\*K/) {$DSK++;}
	if ($_=~/\*L/) {$DSL++;}	if ($_=~/\*M/) {$DSM++;}	if ($_=~/\*N/) {$DSN++;}
	if ($_=~/\*P/) {$DSP++;}	if ($_=~/\*Q/) {$DSQ++;}	if ($_=~/\*R/) {$DSR++;}
	if ($_=~/\*S/) {$DSS++;}	if ($_=~/\*T/) {$DST++;}	if ($_=~/\*V/) {$DSV++;}
	if ($_=~/\*W/) {$DSW++;}	if ($_=~/\*Y/) {$DSY++;}	if ($_=~/\.A/) {$ORA++;}
	if ($_=~/\.C/) {$ORC++;}	if ($_=~/\.D/) {$ORD++;}	if ($_=~/\.E/) {$ORE++;}
	if ($_=~/\.F/) {$ORF++;}	if ($_=~/\.G/) {$ORG++;}	if ($_=~/\.H/) {$ORH++;}
	if ($_=~/\.I/) {$ORI++;}	if ($_=~/\.K/) {$ORK++;}	if ($_=~/\.L/) {$ORL++;}
	if ($_=~/\.M/) {$ORM++;}	if ($_=~/\.N/) {$ORN++;}	if ($_=~/\.P/) {$ORP++;}
	if ($_=~/\.Q/) {$ORQ++;}	if ($_=~/\.R/) {$ORR++;}	if ($_=~/\.S/) {$ORS++;}
	if ($_=~/\.T/) {$ORT++;}	if ($_=~/\.V/) {$ORV++;}	if ($_=~/\.W/) {$ORW++;}
	if ($_=~/\.Y/) {$ORY++;}	}
open FILE2, ">$source-DSaa.txt" or die "failed to open DSaa result file";
print FILE2 "Amino Acid,A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\n",
```

DisCOST.perl (continued)

```
	"Disordered",",",$DSA,",",$DSC,",",$DSD,",",$DSE,",",$DSF,",",$DSG,",",$DSH,",",$DSI,",",$DSL,","
,$DSK,",",$DSM,",",$DSN,",",$DSP,",",$DSQ,",",$DSR,",",$DSS,",",$DST,",",$DSV,",",$DSW,",",$DSY,"\n",
	"Ordered",",",$ORA,",",$ORC,",",$ORD,",",$ORE,",",$ORF,",",$ORG,",",$ORH,",",$ORI,",",$ORL,","
,$ORK,",",$ORM,",",$ORN,",",$ORP,",",$ORQ,",",$ORR,",",$ORS,",",$ORT,",",$ORV,",",$ORW,",",$ORY,"\n"
;
	close FILE2;
```

Rationale: Allows one to count the number of each amino acid that is either disordered or ordered out of a large set of proteins analyzed by Disospred2. Using this data one can see the bias of amino acid or their relative cost in disordered and ordered regions as predicted by Disopred. First you must use Disopred.perl, Fixdisopred.perl (if using the Disopred 2.4 or lower), and Dis.perl.

Input: Name-final.txt from Dis.perl

Output:	Inputfilename-DSaaRAW.txt	Each aa and its predition divided by a (,)
	Inpufilename-DSaa.txt	Comma delimited results (see below)

Notes: The DSaa file is a comma delimited file suitable for use by Excel. The output list the number of each amino acid out of all the protein sequences in the input file that was predicted to be either disordered or ordered.

Shuffle.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @one=0;
my @result=0;
my @before=(<>);
my $file="sequence.txt";
my $file2="EB.txt";
my $file3="all.txt";
foreach (@before) {
my @one=split/\t/;
open FILE, ">sequence.txt" or die "Failed to open sequence.txt";
print FILE $one[0],"\n",$one[1];
close FILE;
system("./runpsiprednr sequence.txt") == 0 or die "Failed on running PSIPRED";
open FILE, "sequence.horiz";
@result=<FILE>;
close FILE;
open FILE2, "$one[0].txt" or die "Failed to open EB file";
print FILE2 $one[0],"\n",@result;
close FILE2;
open FILE3, ">>all.txt" or die "Failed to open all.txt";
print FILE3 $one[0],"\n",@result;
}
```

Rationale: Sends multiple sequences to Psipred - similar to Disospred.perl except used Psipred predicts secondary structure as either helix, strand, or coil.

Input: Simple text file. Each line is “>description**[TAB]**sequence”.

Output: An individual text file of each protein named after the protein description. All.txt - combines all the individual Psipred results.

Notes: There are several other perl files that do similar things and include Shufflehome.perl, Shufflesingle.perl, and Shuffleunfil.perl. Shufflehome.perl uses a second non redundant database to increase the number of sequences that can be processed at once. The nciblast which is part of the psipred prediction program only runs on one processor. Using a second program instance should double the output. Shufflesingle.perl uses just neural networks (no blasting) and is much faster but not as accurate as when sequences are first blasted and a consensus sequence of sorts is used. Shuffleunfil.perl uses an unfiltered database for blasting and gives slightly different results from a filtered database. The output are quite similar for each of the other perl files; however, output files are named differently so that no file overwriting or combining occurs. Care should be used in using two program instances as the individual protein files may be overwritten if using similar protein datasets (inputs).

Psipred.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @before=(<>);
my @ss; my @seq;
my @file; my @seqonly; my @final;
my $source=$ARGV; $source=~s/\.txt//;
foreach(@before) {
        if ($_=~/\>/) {push(@ss,$_,");}
        if ($_=~s/Pred\: //) {push(@ss,$_);}}
foreach(@before) {
        if ($_=~/\>/) {push(@seq,$_,");}
        if ($_=~s/  AA\: //) {push(@seq,$_);}          }
chomp (@ss); chomp (@seq);
foreach (@ss) {s/\>/\n\>/g;}
foreach (@seq) {s/\>/\n\>/g;}
open FILE, ">$source-SEQ.txt" or die "unable to save Sequence file";
print FILE @seq; close FILE;
open FILE, ">$source-HEC.txt" or die "unable to save Structure (HEC) file";
print FILE @ss; close FILE;
open FILE, "<$source-SEQ.txt"; my @seqfile = <FILE>; close FILE;
open FILE, "<$source-HEC.txt"; my @ssonly = <FILE>; close FILE;
foreach (@seqfile) {
        if ($_=~s/\>EB\d\d\d\d//) {
                push(@seqonly,$_);} #Modify this pattern if not >EB1245
        if ($_=~s/\>gi\d\d\d\d\d\d\d\d//) {
                push(@seqonly,$_);} #Or add your pattern as another if statement
}
foreach (@seqonly) {
        push(@final,$ssonly[1],$_);
        shift @ssonly;}
chomp @final;
foreach (@final) {s/\>/\n\>/g;}
open FILE, ">$source-final.txt" or die "unable to open Final file";
print FILE @final; close FILE
```

Rationale: Similar to Dis.perl except used for Psipred results. Strips out sequences and Psipred from multiple results in the All.txt result file and converts them into a comma delimited text file suitable for Excel and HECcost.perl.

Input:	Name.txt	Output file from Shuffle.perl or similar protein
Output:	Name-HEC.txt	Each line is “Description,Psipred result”.
	Name-SEQ.txt	Each line is “Description,sequence”.
	Name-final.txt	Each line is “Description,Psipred result,sequence”.

HECcost.perl

```
#!/usr/bin/perl
use warnings;
use strict;
my @before=(<>);
my @all; my @smp; my @splA; my @splB;
my @A; my @B; my @arrA; my @arrB;
my @final; my $line;
my $source=$ARGV; $source=~s/\-final\.txt//;
foreach (@before) {
        s/\>EB\d\d\d\d\,//; #Modify this pattern if not >EB1245
        s/\>gi\d\d\d\d\d\d\d\d\,//; # or add your pattern as another substituion
        @all=@before; }
shift @all;
chomp @all;
foreach (@all) {
        @smp=split ',';
        push (@A,$smp[0]); #@A appears chomped
        push (@B,$smp[1]);          }
foreach (@A) {@splA=split '';
        foreach (@splA) {push (@arrA,$_);}}
foreach (@B) {@splB=split '';
        foreach (@splB) {push (@arrB,$_);}}
foreach (@arrA) {
        $line=join('',$_,$arrB[0]);
        shift @arrB;
        push (@final,$line,","); }
open FILE, ">$source-S2aaRAW.txt" or die "failed to open SSaaRAW file";
print FILE @final; close FILE;
my $HA=my $HC=my $HD=my $HE=my $HF=my $HG=my $HH=my $HI=my $HK=my $HL=my $HM=my
$HN=my $HP=my $HQ=my $HR=my $HS=my $HT=my $HV=my $HW=my $HY=0;
my $CA=my $CC=my $CD=my $CE=my $CF=my $CG=my $CH=my $CI=my $CK=my $CL=my $CM=my
$CN=my $CP=my $CQ=my $CR=my $CS=my $CT=my $CV=my $CW=my $CY=0;
my $EA=my $EC=my $ED=my $EE=my $EF=my $EG=my $EH=my $EI=my $EK=my $EL=my $EM=my
$EN=my $EP=my $EQ=my $ER=my $ES=my $ET=my $EV=my $EW=my $EY=0;
foreach (@final) {
        if ($_=~/HA/) {$HA++;}      if ($_=~/HC/) {$HC++;}      if ($_=~/HD/) {$HD++;}
        if ($_=~/HE/) {$HE++;}      if ($_=~/HF/) {$HF++;}      if ($_=~/HG/) {$HG++;}
        if ($_=~/HH/) {$HH++;}      if ($_=~/HI/) {$HI++;}      if ($_=~/HK/) {$HK++;}
        if ($_=~/HL/) {$HL++;}      if ($_=~/HM/) {$HM++;}  if ($_=~/HN/) {$HN++;}
        if ($_=~/HP/) {$HP++;}      if ($_=~/HQ/) {$HQ++;}      if ($_=~/HR/) {$HR++;}
        if ($_=~/HS/) {$HS++;}      if ($_=~/HT/) {$HT++;}      if ($_=~/HV/) {$HV++;}
        if ($_=~/HW/) {$HW++;}  if ($_=~/HY/) {$HY++;}      if ($_=~/CA/) {$CA++;}
        if ($_=~/CC/) {$CC++;}      if ($_=~/CD/) {$CD++;}      if ($_=~/CE/) {$CE++;}
        if ($_=~/CF/) {$CF++;}      if ($_=~/CG/) {$CG++;}      if ($_=~/CH/) {$CH++;}
        if ($_=~/CI/) {$CI++;}      if ($_=~/CK/) {$CK++;}      if ($_=~/CL/) {$CL++;}
        if ($_=~/CM/) {$CM++;}  if ($_=~/CN/) {$CN++;}      if ($_=~/CP/) {$CP++;}
        if ($_=~/CQ/) {$CQ++;}      if ($_=~/CR/) {$CR++;}      if ($_=~/CS/) {$CS++;}
        if ($_=~/CT/) {$CT++;}      if ($_=~/CV/) {$CV++;}      if ($_=~/CW/) {$CW++;}
        if ($_=~/CY/) {$CY++;}      if ($_=~/EA/) {$EA++;}      if ($_=~/EC/) {$EC++;}
        if ($_=~/ED/) {$ED++;}      if ($_=~/EE/) {$EE++;}      if ($_=~/EF/) {$EF++;}
```

HECcost.perl (continued)

```
        if ($_=~/EG/) {$EG++;}      if ($_=~/EH/) {$EH++;}      if ($_=~/EI/) {$EI++;}
        if ($_=~/EK/) {$EK++;}      if ($_=~/EL/) {$EL++;}      if ($_=~/EM/) {$EM++;}
        if ($_=~/EN/) {$EN++;}      if ($_=~/EP/) {$EP++;}      if ($_=~/EQ/) {$EQ++;}
        if ($_=~/ER/) {$ER++;}      if ($_=~/ES/) {$ES++;}      if ($_=~/ET/) {$ET++;}
        if ($_=~/EV/) {$EV++;}      if ($_=~/EW/) {$EW++;}    if ($_=~/EY/) {$EY++;}      }
open FILE, ">$source-S2aa.txt" or die "failed to open S2aa result file";
print FILE "Amino Acid,A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\n",
"Helix",",",$HA,",",$HC,",",$HD,",",$HE,",",$HF,",",$HG,",",$HH,",",$HI,",",$HL,",",$HK,",",$HM,",",$HN,",",
$HP,",",$HQ,",",$HR,",",$HS,",",$HT,",",$HV,",",$HW,",",$HY,"\n",
"Coil",",",$CA,",",$CC,",",$CD,",",$CE,",",$CF,",",$CG,",",$CH,",",$CI,",",$CL,",",$CK,",",$CM,",",$CN,",",$CP
,",",$CQ,",",$CR,",",$CS,",",$CT,",",$CV,",",$CW,",",$CY,"\n",
"Strand",",",$EA,",",$EC,",",$ED,",",$EE,",",$EF,",",$EG,",",$EH,",",$EI,",",$EL,",",$EK,",",$EM,",",$EN,",",$E
P,",",$EQ,",",$ER,",",$ES,",",$ET,",",$EV,",",$EW,",",$EY,"\n";
close FILE;
```

Rationale: Similar to DisCost.perl but for Psipred results. Allows one to count the number of

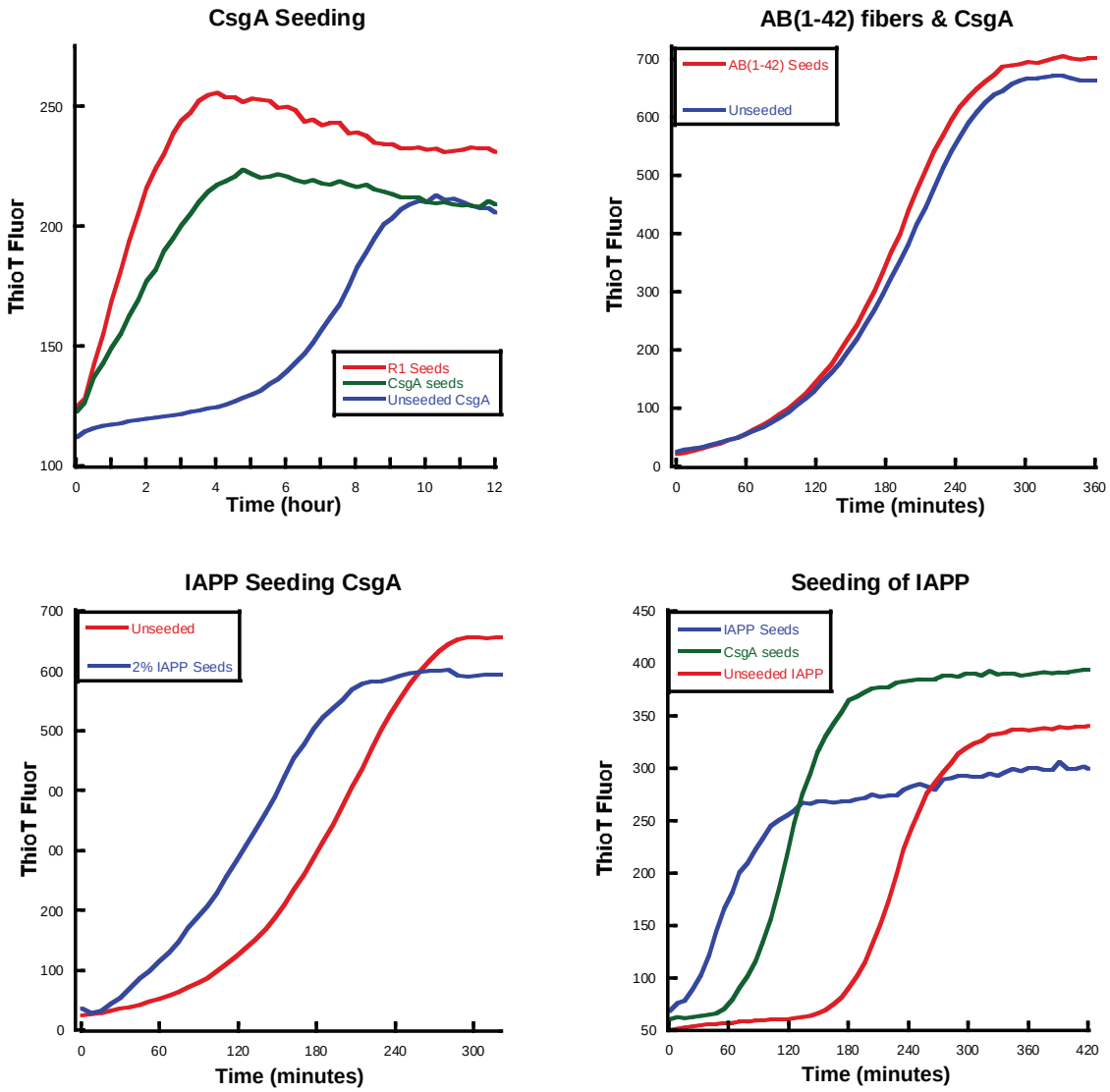
each amino acid that is either in a helix (H), strand (E), or coil (C) from a large set of proteins analyzed by Psipred using Shuffle.perl and Psipred.perl. Using this data one can see the bias of amino acid or their relative cost in each secondary structure.

Input: Name-final.txt from Psipred.perl

Output: Inputfilename-SSaaRAW.txt Each aa and its prediction divided by a (,)
Inpufilename-SSaa.txt Comma delimited results (see below)

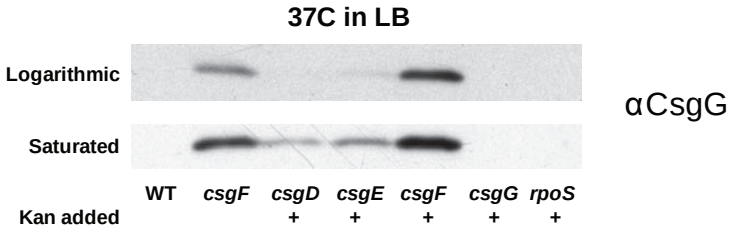
Notes: The SSaa file is a comma delimited file suitable for use by Excel. The output list the number of each amino acid out of all the protein sequences in the input file that was predicted to be either helix, strand, or coil.

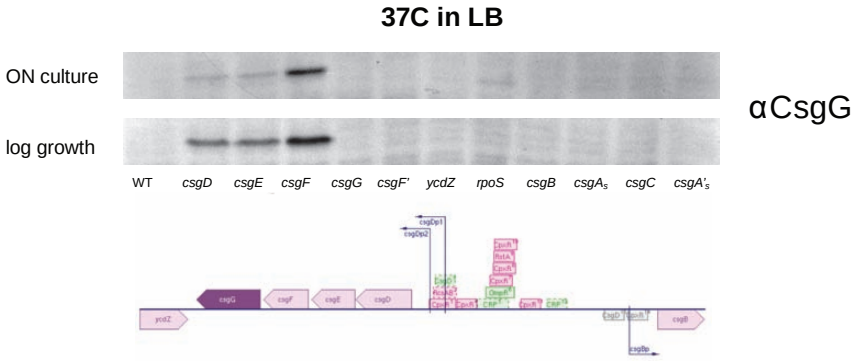
Appendix B: Seeding and Cross seeding with CsgA-His

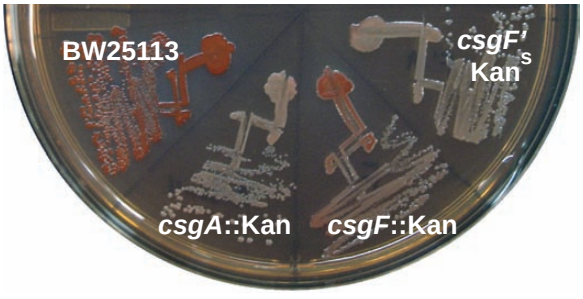


IAPP used in the seeding of CsgA-His was from Andrew Miranker. See Koo and Miranker (2005) for this method. When IAPP was made by the Chen and Wetzel (2001) method by Jenz no seeding occured; see Hammer et al PNAS (2007). However, preformed fibers of CsgA-His were able to seed unpolymerized IAPP prepared by the method used in Hammer et al PNAS (2007).

Appendix C: Downstream promoting and polar effects when using Keio collection strains

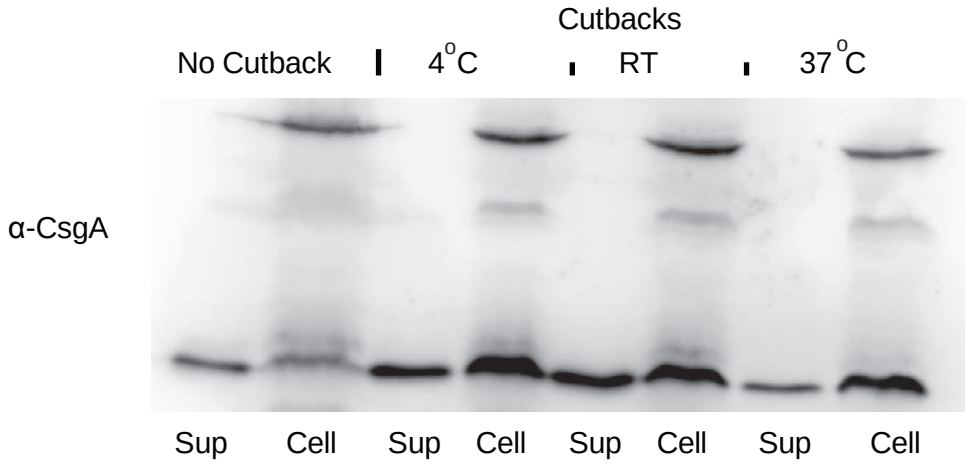






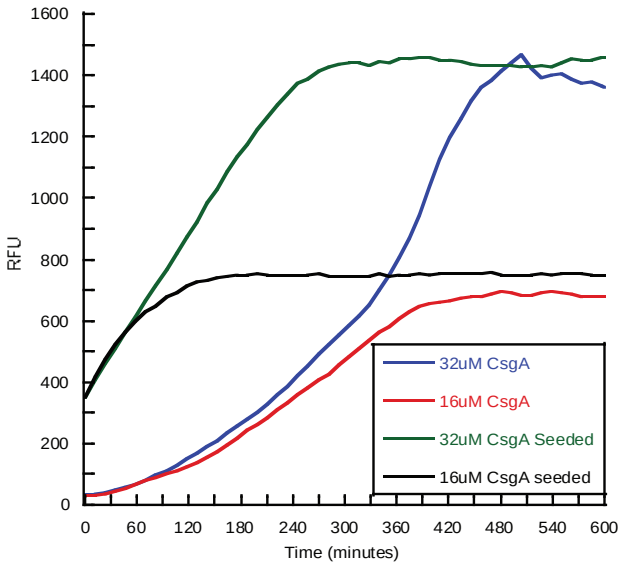
The Kan resistance cassettes in Keio collection strains promote downstream genes that are usually off. CsgG is present in cells grown in LB overnight or logarithmically only in *csgD*, *csgE*, and *csgF* strains. Other Keio mutants had no effect. When the Kan cassette was removed using pCP20, no CsgG is seen in LB; however this mutant (*csgF'*) may be polar on *csgG*. The *csgF'* strain does not have polymerized CsgA in the agar of CR plates.

Appendix D: Effects of cold shock on CsgA-His yields



When purifying CsgA-His, addition of room temperature (RT) or colder LB to mid-log (OD600 0.4 to 0.8) cells increases the final yield on a per cell basis (cells are induced at OD600 0.8 to 1.2). This effect appears to be due to a cold shock since LB of the same temp (37°C) has not effect on the amounts of CsgA in the supernatant fraction. Yields are increased two to four fold using 50% volume additions. The RT LB addtion lowered the culture temperature to 32°C; the 4°C LB addition lowered the temperature to 27°C. By the time of induction temperatures had returned to 37°C.

Appendix E: Method to purify and store CsgA-His for later use



Protocol for storage

1. Elute CsgA-His in 8M GuHCl using NiNTA.
2. Buffer exchange into H20 to remove all salts and filter though 0.02µm Filter
3. Imediately precipitate using Chloroform Methanol to prevent amyloid formation
4. Pellet, Decant, and Evaporate
5. Store at -20°C or -80°C

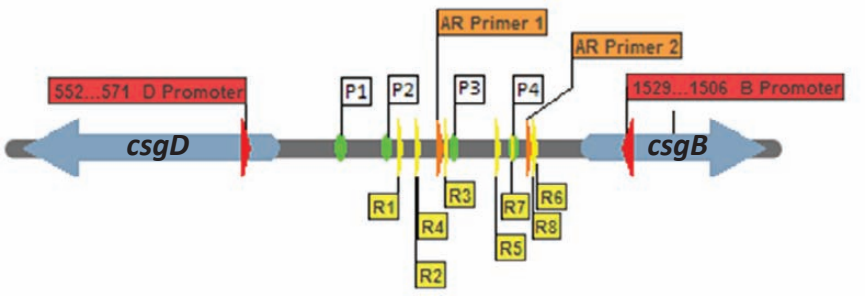
Resuspension (approximately 1 to 2 hours)

1. Add 8M GdHCl 1% TFA to CsgA-His (50µM). Mix and bath sonicate 20 min.
2. Centrifuge 20 minutes at >10,000xG. Carefully remove and keep top 80%.
3. Fitler through 0.02µm Anotop.
4. Buffer exchange use G-25 column into 50mM KPi pH 7.2. Use immediately.

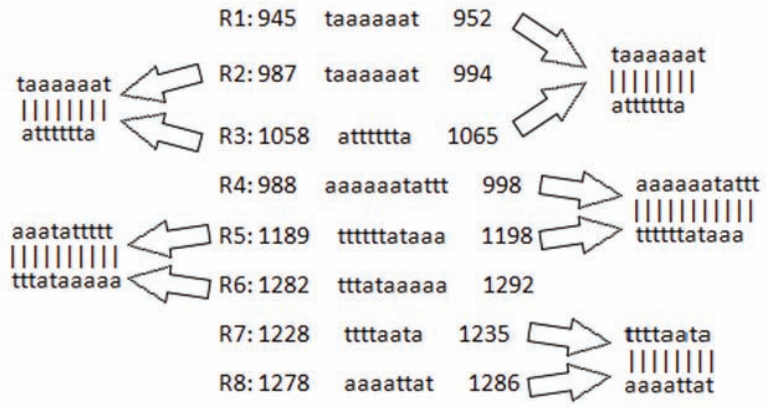
This method can be used to make a lot of CsgA protein for later use and should result in a more consistent lag phase as you can always resuspend the same amount of CsgA using the same protocol.

Other methods for precipitation and resuspension were also tried. TCA, acetone, and EtOH did not appreciably precipitate CsgA-His. Ammonium sulfate did work but was much slower than CH3Cl/MeOH which instantly precipitate CsgA-His. Peptides of AB and CsgA repeating using (R1-R5) can be denatured and resuspended using the Chen & Wetzel (2001) method. These methods denature any seeds with TFA or HFIP which is then evaporated away. The resulting peptide film is resuspened in 2mM KOH or 2mM HCl and then buffered appropriately. CsgA-his treated in this manner did not resuspend well in either 2mM HCl or 2mM KOH, in water, or in 50mM KPi. Additionally, no lag phase was observed when trying these methods.

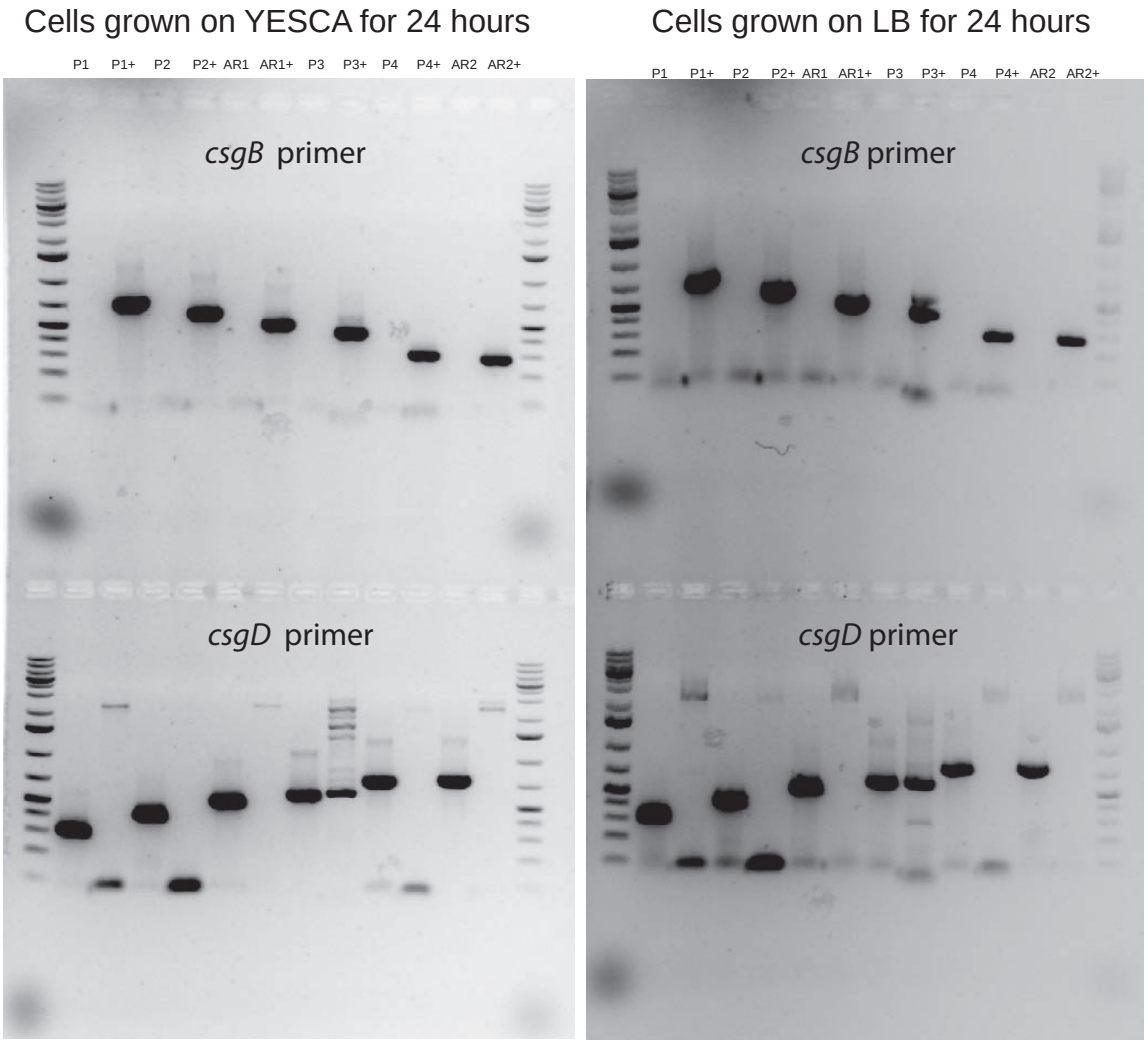
Appendix F: The intergenic region between *csgD* and *csgB* does not invert



Possible flipping combinations



There are eight palindromic regions in the promoter region between *csgD* and *csgB* which would allow for an inversion similar to the Type I pili sysytem. Six sets of internal primers were made to ensure that each segment between repeats were covered and that there were no large regions between either genes. Using PCR one can then check for inversion of DNA elements.



Primers that annealed to either *csgB* or *csgD* were mixed with one primer of a set of complementary primers. The position of these second primers are indicated on the previous page. For each PCR reaction a product will only be generated if the second primer anneals to the opposite strand of DNA from the *csgB* or *csgD* primer. For example, a product is expected if the csgD primer is used with any of the second set of primers that anneal to the negative strand. Primers that anneal to the postive strand are indicated with a (+). If the second primer set anneals in the middle of an invertable DNA element and this element is present in both the default and inverted state, then a PCR product is expected for both primers in a comlementary set. When using the *csgB* primer, none of the sets had products for both primers, indicating there are not invertible DNA elements in BW25113 under the conditions tested. The *csgD* primer generated artifacts.