AGGREGATION, DISSEMINATION, AND ANALYSIS OF HIGH-THROUGHPUT SCIENTIFIC DATA SETS IN THE FIELD OF PROTEOMICS

by

Jayson A. Falkner

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2008

Doctoral Committee

Professor Philip C. Andrews, Chair
Professor Daniel M. Burns Jr
Assistant Professor Matthew A. Young
Assistant Professor Alexey Nesvizhskii

I dedicate this to my parents and family for supporting me in pursuit of over-education to the fullest. Thankfully I've learned that academics, science, and the pursuit of intellect are of little to benefit an individual. Friends, family, and society give purpose to both work and life. One is not without the other. Thanks Mom and Dad.

Acknowledgments

Phil Andrews has been a constant source of ideas, enthusiasm, and support for my work. Pretty much every part of this thesis was inspired in one way or another by Phil, and I feel very fortunate to have had a mentor that wanted include me in most everything. Even sending me around the world for countless conferences, workshops, and meetings. I think that all mentors work in mysterious and unmeasurable ways, and I have been unbelievably lucky to have such a good mentor and friend.

Pete Ulintz, Eric Simon, Anastasia Yocum, Bryan Smith, James “Augie” Hill, and the rest of Phil's lab have been great friends and contributed in many ways to my work. Dan Burns, David States, Brian Athey, Gil Omenn, Alexey Nesvizhskii, Matt Young, Heather Carlson, David Burke, and others in the program are all to thank for advice, guidance, and teaching me about what getting a PhD is really about.

Preface

I am a tool builder and optimizer, and at times a decent theorist, if I can muster the attention. I enjoy understanding a problem, and determining efficient solutions. I honestly believe that most scientific work is a series of failures that results in something worth writing about. My favorite quote from a peer, David J States, describes complex projects, “Fail early and often, or late and spectacular.” This quote reminds me that most problems can be solved by repeated small failures, and I like to think that I'm pretty good at that. I hope to keep avoiding the spectacular failures throughout life. So far so good. Proteomics poses many complex problems. The field itself relies greatly on Bioinformatics, which itself is still quite a novel term to most and certainly a young field. An appropriate one-line summary is the following. I understand how to and can collect my own data but primarily my job is to develop algorithms and use statistics to determine what can be learned from huge amounts of proteomics data. Certainly a few great software tools exist for proteomics and almost every other year a significant advancement occurs in the instrumentation, which dramatically changes the amount, quality, and type of data generated. Early on in my career I realized this critical point. Proteomics, science in general too, is an appreciation of how much we don't yet know. The best that can be done is to narrow the scope to a tangible problem and to move it forward. Repeat the process a few times and you have a thesis.

The specific proteomics problem I elected to work on is that of inferring peptide and protein sequences from mass spectrometry data. My efforts are described in considerable detail throughout this manuscript. I feel it is important to note that I have also spent a lot of effort and time enabling others to repeat what I've done and apply my algorithms and software tools to their proteomics problems. Never did I entertain the thought of my work completely solving all the problems of proteomics. Rather, I hoped to make several significant advancements, and greatly accelerate similar research. I feel I have been successful in my efforts, and in particular I have high confidence in many components of my work, particularly Tranche and Bonanza. These latter projects are now being used by a large number of proteomics researchers. In particular, I am quite pleased that all of my work is available as both free and open-source, largely thanks to shared philosophies with Phil Andrews and public accessibility efforts by the NCI and NCRR.

Table of Contents

Dedication ii

Acknowledgments iii

Preface iv

List of Figures vii

List of Tables viii

List of Abbreviations/Acronyms ix

Chapter

I. Introduction 1

II. ProteomeCommons.org 19

III. Aurum 27

IV. Tranche 41

V. Tranche Implementation and Technical Details 51

VI. Bonanza 62

VII. Conclusion 85

Figures 91

Tables 106

Bibliography 109

List of Figures

Figure 1-1 92

Figure 1-2 93

Figure 2-1 94

Figure 3-1 95

Figure 3-2 96

Figure 4-1 97

Figure 4-2 98

Figure 5-1 99

Figure 5-2 100

Figure 5-3 101

Figure 6-1 102

Figure 6-2 103

Figure 6-3 104

Figure 6-4 105

List of Tables

Table 6-1 107

Table 6-2 108

List of Abbreviations/Acronyms

MS: Mass Spectrometry

MS/MS: Tandem Mass Spectrometer

API: Application Programmer's Interface

TOF: Time of Flight

MALDI: Matrix Assisted Laser Desorption Ionization

ESI: Electrospray Ionization

SLD: Soft Laser Desorption

TIC: Total Ion Current

MudPIT: Multidimensional Protein Identification Technology

Chapter 1

Introduction

Mass spectrometry is formally defined as an analytical technique that measures of the mass to charge ratio of ions. Typically a collection of ions are analyzed simultaneously and the mass spectrometer generates a mass spectrum (MS) that can be used to interpret the mass to charge ratio of ion species present in detectable quantities. Modern mass spectrometers are largely based off of designs by A.J. Dempster and F.W. Aston, developed in 1918 and 1919 respectively. Aston later received the Nobel Prize in Chemistry for his work in mass spectrometry in 1922. However, much more recent developments in mass spectrometry have made the technique viable for analyzing ions that were previous difficult to desorb or ionize. In 1987 both electrospray ionization (ESI)[1] and soft laser desorption (SLD)[2], developed by John B. Fenn et al. and Koichi Tanaka et al. respectively, and matrix assisted laser desorption/ionization (MALDI)[3], developed by Franz Hillenkamp et and Michael Karas, were developed. In 2002 John B. Fenn and Koichi Tanaka were awarded the Nobel Prize in Chemistry for developments in mass spectrometry[4], albeit with a lack of award to Michael Karas and Franz Hillenkamp. Both ESI and MALDI are now in widespread use and allow for the analysis of peptides and proteins and their complexes via mass spectrometry. While mass spectrometry-based protein analysis is not the only tool used in proteomics, without the ESI and MALDI ionization method the work presented in this thesis would not be possible.

The Nobel Prize awarded in 2002 to Fenn and Tanka reflects that their work enabled a key part of proteomics, the ionization of intact peptides and proteins without the use of chemical modifications to enhance volatility. Once successfully ionized, the masses of these molecules could accurately be measured, which enabled a wealth of new knowledge to be collected on biological samples. Further developments in mass spectrometers and software would enable complex protein samples, such as tissue or serum, to be analyzed in high-throughput experiments and for one of the first times it is possible to attempt to survey the state of many, and potentially all, proteins in a living organism. The full potential of this development has yet to be realized, but certainly many notable experiments have been performed. One of the most influential strategies of protein analysis via mass spectrometry was describe by Eng et al. [5]. Eng describes the process of a shotgun proteomics experiment a statistical correlation of peptide identifications to mass spectrometry data and subsequent inference of protein identifications. In short, many proteins are too large to be ionized well for mass spectrometry and many mixtures are too complex to be analyzed alone. Shotgun proteomics relies on converting a complex mixture of protein into a set of smaller peptides that are applicable for mass spectrometry, those peptides are then automatically separated based on intrinsic properties, and finally mass spectrometry is used to analyze the entire sample. Post data collection, the resulting mass spectra are compared against a library of known protein sequences. A statistical analysis is then used to infer the original set of proteins analyzed. Shotgun proteomics is often done with a similar set of conditions for creating peptides, separating, analyzing, and finally inferring protein identifications; however, each of the particular components can be altered and various forms of such alteration comprise a large portion of current proteomics literature [6, 7, 8].

Most all of the proteomics work described in this manuscript is predicated on developments and refinements of the initial shotgun proteomics strategy. Good reviews exist [9] including comprehensive terminology standards [10], but it is relevant to review this information in order to frame the rest of the manuscript coherently. Figure: 1-1 provides a conceptual overview of shotgun proteomics and also includes a brief cartoon that illustrates the protein chemistry of interest. Peptide and protein chemistry itself is a complete discipline; however, for the purposes of this manuscript, it largely suffices to present proteins as nothing more than a long string of English alphabet characters. In this context, Peptides are simply shorter strings of the same characters. The sequence (not active structure) of proteins can be reasonably well represented by such strings because each protein is primarily comprised of combinations of the standard 20 amino acids. Modifications of these 20 amino acids do occur in proteins but discussion of relevant ones will largely be left undiscussed until chapter 6. Figure 1-1 continues the proteins as strings analogy to describe mass spectrometry based proteomics.

Figure 1-1b illustrates the effect of proteolytic digestion of a protein, or splitting the string in context of the textual analogy mentioned above. Different mass spectrometers are capable of measuring a broad range of mass to charge ratios at various sensitives; however, generally, mass spectrometers target ions with a m/z between the range of 200-3,000 Da when applied to shotgun proteomics. The majority of data analyzed in this manuscript is generated by a mass spectrometer optimized to analyze m/z of 900-2,500 Da. Phrased differently, strings of approximately 6 to 25 characters. Known protein sequences range anywhere from a few characters to thousands, often with more than a few hundred characters per protein. Thus, the process of proteolysis is critical in order to convert proteins into peptides that are compatible for mass spectrometry analysis. Trypsin in particular is popular because it tends to work predictably, results in many peptides of the desired size, and finally tends to leave each peptide with a single positively charged amino acid at the C-terminus.

Figure 1-1c illustrates what will be presented in this manuscript as a primarily black-box process of mass spectrometry. Peptides are ionized and analyzed by the mass spectrometer and resulting spectra are generated. For the purpose of the work described in this manuscript, attention must be focused on the meaning of the mass spectra; however, the inner-workings of the physical mass spectrometer are relevant in the sensitivity and resolution of the mass spectra. Several excellent reviews for commonly used mass spectrometers exist, and discussion of particular relevant features will not occur until chapter 6. Figure 1-2 clarifies the precise data of interest in this manuscript, a mass spectrum, which is also represented in Figure 1-1c. Resulting spectra have several properties of interest. Depending on the mass spectrometer, spectra will represent different features – e.g. mass defect, mass limit, and mass range – have different mass resolution and mass resolving power for determination of isotopic states. Each of these aspects generally contributes significantly to the confidence of identifications inferred by software developed for protein and peptide identification. In order to infer peptides represented by a mass spectrum, each *m/z* must be compared to a set of known existing peptide masses. Such a list can easily be obtained by processing an existing list of possible protein sequences to obtain a set of all theoretical peptides. Many appropriate publicly accessible protein databases exist, including the RefSeq databases from the National Center for Biotechnology Information (NCBI) [11] and the European Bioinformatics Institute's (EBI) International Protein Index (IPI) [12]. A complete list and archive of current and previous versions of these protein databases is available from the ProteomeCommons.org/Tranche FASTA resource [13]. Assuming one has an appropriate list of theoretical peptides and a mass spectrum, inference of peptides present in the spectra can be accomplished by creating a sublist of theoretical peptides that would have the same *m/z* as ions present in the mass spectra. This list can further be reduced by filtering out peptides that should not be present according to experimental steps taken prior to peptide ionization. Generally, even the most minimal lists of theoretical peptides for a shotgun proteomics experiment can result in ambiguous matches to observed m/z in a mass spectrum. Several statistical approaches have been developed to address this [14, 15]. In short, algorithms rely on the mass accuracy of the MS instrument to reduce the potential peptides matches for any observed MS peak. This practice is then combined with a statistical estimation of how likely a random protein will have a peptide that matches an observed peak. The results can often yield a confident identification for samples with low complexity; however, tandem mass spectrometry (MS/MS) is typically employed as an orthogonal approach to identify present peptides with high confidence in samples of higher complexity. If an instrument is MS/MS capable, typically MS analysis is completely ignored in favor of the information rich MS/MS. Figure 1-2 illustrates an example MS/MS spectrum. The data are similar to that of a MS spectrum, but instead of looking at multiple peptides simultaneously, a single peptide is isolated from the MS scan, fragmented, and re-analyzed by itself. The resulting MS/MS spectrum represents a ladder of masses that can be matched back to the theoretical amino acid sequence of the source peptide. Thus, for a shotgun proteomics, a logical way to infer peptides from a complex mixture is to separate them as best as possible prior to ionization for MS, repeatedly select different MS ions for MS/MS, and finally create a software package that can efficiently process all spectra and match appropriate theoretical peptide sequences (Figure 1-1d). If the results are taken one step further, as shown in Figure 1-1e the set of identified peptide sequences can be used to identity what source proteins were likely present.

Current software trends in the field of mass spectrometry-based proteomics can be well described by the final step illustrated by Figure 1-1 (c) and (d). Tandem mass spectra represent a wealth of information that can be valuable for understanding a number of biological and physical processes. Exactly how peptides fragment and form MS/MS is not completely understood and is itself an active area of research [16-18]. Significantly different fragmentation can be observed depending on properties of the peptide, amount and frequency of applied excitation energy, and characteristics of different mass spectrometers. Due in part to the incompletely understood fragmentation mechanisms, it should be no surprise that another active area of research is that of refining software algorithms to correctly infer peptide identifications based on MS/MS data [19-30]. It is fair to state that the peptide inference problem itself is the most active area of current research. Many groups are working on logical refinements to existing search algorithms in order to improve the statistics, selectively identify particular post translational modifications, or simply speed up performance. Post MS/MS analysis represents the next, currently popular area of algorithm development. Many researchers are working on creating better software for inferring protein identifications based on sets of inferred peptides [31-33]. A naive approach of identifying all proteins that share an observed peptide will excessively identify proteins. Many proteins share the same peptides, and effort must be placed in identifying what protein most likely is represented given all observed peptide identifications. This is particularly important when dealing with homologous proteins that share large portions of sequence. Protein inference algorithms must carefully identify what homologous proteins are clearly present versus ambiguously identifiable proteins. Additionally, several areas of research distinct from the original MudPIT analysis pipeline have also emerged. Protein database independent identification of peptide sequences based solely on MS/MS data is quickly becoming a viable alternative to database techniques. This practice is often referred to as *de novo* peptide sequencing [34]. The *de novo* algorithms have had limited widespread adoption, but share a significant portion of active research interest. Also, libraries of known peptide identifications and mass spectra have started to emerge [35-37]. The practice is based around the concept that existing MS/MS based peptide identifications can be recycled. A statistically valid identification should generally hold true across data sets, and valuable time and identifications can be inferred from old data to new. Such libraries are becoming popular now because of the increasing availability of data sets, in a large part due to Tranche. Comparison to both a known library of identifications and a theoretical set of proteins is a promising standard for future MS/MS analysis. It is both a very logical direction and becoming much more practical due to open-access, large-scale data set publication. Finally, several resources have emerged to help aggregate mass spectrometry based information [38-41]. This thesis work describes one of the most prominent, Tranche, which has established a P2P network for scientific data sharing. A core set of computers supported by various groups and organizations maintains the majority of data on the network; however, individual users computers are also used to help host data, increase the availability of on-line data, and speed up downloads.

Critique of Existing Methodology

Reinvention of the Wheel – Several critiques can be made of mass spectrometry based proteomics efforts at the start of this thesis work. May of many of these critiques still apply to the current field. First and foremost is repetition of labor, or so called “reinvention of the wheel”. Sequest (1994) is widely accepted as the first statistical algorithm for shotgun proteomics-based peptide and protein identification. More than a decade later seemingly far too many research groups are still actively developing algorithms that are fundamentally similar to Sequest. The ProteomeCommons.org tools page provides a list of at least 20 different MS/MS search engine tools. All of which are remarkably small evolutions from the original Sequest algorithm. Few if any revolutionary techniques have been introduced. Even Sequest itself is still often used as a benchmark of novel developments. Conceptually is is easy to recognize that Sequest should not currently be considered a state of the art algorithm. It is the first generation of statistical scoring MS/MS software algorithms, it identifies peak lists largely in ignorance of the size or quality of the input data set, and even Sequest's author acknowledges deficiencies with the algorithm. Ten years ago Sequest was a great tool for MS/MS based proteomics. Currently Sequest alone is not a benchmark anyone should use when comparing MS/MS search engine developments, note Sequest with PeptideProphet is a special case. Current MS/MS search engines must measure up to significantly more than what Sequest had to. The open-source search engines X!Tandem [42] and OMSSA [43] provide excellent examples of statistical scoring based on the entire set of data being processed. The algorithms are not perfect, most manuscripts demonstrate significant benefits to using other commercial algorithms, but the open-source algorithms do provide both a free and explicit example of the statistical process behind MS/MS search algorithms. A refined scoring algorithm named k_score for X!Tandem is another noteworthy refinement the provides a much more sophisticated scoring metric for the X!Tandem code. It would be helpful to see any new MS/MS search engine developed demonstrate a significant improvement versus both X!Tandem and OMSSA. Ideally, not a questionable 5-10% improvement, but an improvement that can not easily be reached by modifying or altering the input parameters of the open-source, widely accessible search algorithms. If such an improvement cannot be obtained, then any effort invested in the new algorithm is of questionable benefit compared to simply using the existing algorithms. Unfortunately, the vast majority of MS/MS research papers being published seem to demonstrate a marginal improvement over existing algorithms – sometimes over nothing but the original Sequest – and this type of publication does not seem to be moving the field forward. Rather, in my personal opinion, this fuels individual groups egos and establishes an environment where significant effort is invested in figuring out how to make particular MS/MS algorithms appear superior versus a comparable algorithm. With that stated, I do not propose the publication ban of minor evolutions to statistical MS/MS search algorithms. Rather it would be nice to see a generally accepted sediment that one should not expect to impress anyone with a new MS/MS search algorithm. If a group wishes to publish an algorithm, they should bear the burden of picking several openly accessible data sets and running their algorithm versus other available algorithms. Currently this burden is problematic because no clear benchmark proteomics data sets have been adopted by the community. The Aurum data set presented in this thesis is a good candidate, but even it is not yet in widespread use, nor will it be an ideal data set until both LTQ/Orbitrap and the MALDI TOF/TOF data is published, which is planned in subsequent publications of the data set. With the critique stated, the concluding comment regards a positive side-effect from the plethora of basically similar MS/MS search algorithms. Most active proteome informatics groups have their own MS/MS search engine, and are invested in keeping it mainstream use. This is best cited by the 2007 and 2008 ABRF iPRG groups (of which the author is a 2008 member). The organizing members are tasked with comparing MS/MS data sets, which forces cross communication regarding results and the similarity thereof. This cross communication is good because I think it will accelerate the community realization that relatively few significant evolutions in statistical MS/MS search algorithms have been realized. It remains to be seen if the iPRG's participating community will benefit, but certainly the key software developers are aware of each other's work. Hopefully future efforts in the iPRG and similar projects in the proteomics community will focus on the establishment of benchmark datasets and reusable analyses results for peer-review publications that wish to claim improvements in MS/MS search algorithms. This would be an excellent cornerstone for journals to build upon and enforce that “novel” algorithms actually compare against existing data sets that the publishing lab did not produce. It would also alleviate the burden of asking individual labs to acquire and run all current MS/MS search engines including commercial products. The results of existing analyses could be made available for direct comparison against.

Unacceptable Publication Standards – Another critical critique is the lack of reproducibility of bioinformatics analyses, primarily due to lack of access to the original data sets. Mass spectrometry data sets have quickly grown in size. A single experiment can easily generate gigabytes of raw data. No proteomics journal currently requires that the full data set accompany a peer-reviewed article for publication, nor does any journal currently require that such data are published elsewhere. Several years ago this practice would not be considered unreasonable given the size and quantity of proteomics experiments; however, recent developments in proteomics, Chapters 4 and 5, have illustrated efficient methods of both storing and disseminating proteomics data sets of virtually any size. Not only are the approaches demonstrated, but they are free to use. It is no longer acceptable to claim the size of data sets as prohibitive to their complete publication. Several journals have made formal recommendations to encourage data sharing to help with reproducibility of data analyses. While not ubiquitous, the practice of full publication of data sets, parameters, and software used in bioinformatics analysis has gained traction to the extent that it may indeed soon be standard practice. Use of Tranche goes to greatly support this point, and it is possibly the most significant contribution this thesis work has provided to the proteomics community.

Outside of the size of data sets being published, three other issues are presented when scientists are tasked with publication of their data sets. First is that of protection of unidentified but valuable information in the data set. Second is that of the inability of others to process the data without access to a vendor's proprietary software. Third, is that most data is junk. It is not desirable to establish a trash heap of public accessible data – only high quality data should be published. It is my opinion that these three arguments are not appropriate for those interested in basic science research. Generally, all government funded peer-review research should mandate that raw data sets be released upon acceptance of a peer-review manuscript with the agreement that use of the raw data constitutes formal citation. Regarding the first argument, protection of unidentified data. The peer-review publication is the authors chance to present a complete analysis of the data set. It should be expected that further information might be mined from the data, yet it is not appropriate to attempt any sort of claim to subsequent use of the data. The Science Commons has described why this is not appropriate in detail, and in short the answer is that it is ridiculous to try and maintain an indefinite chain of citation, permission, and credit for data sets. It is reasonable to expect a citation if one's data set is used; however, no formal requirement should be maintained. Successful reuse of data sets, especially creative reuse, should not be burdened by the originating author or groups ego. A best effort system should exist to cite and support publishers of data, but data must be complete free for reuse if it is actually going to be helpful.

In regards to the second argument, the inability of users to process the actual data set. Yes, it is likely that most users will not have a license to the original vendor's software. However, if there is no penalty to publishing data (Tranche demonstrates that there is not), then it should be published. Some users will be able to access the raw data, and in due time the vendors format may become accessible via free to use tools. Several examples of this latter case are present in the ProteomeCommons.org IO framework presented in this work. Many vendors have published binary tools that can process proprietary data files and produce open-access file formats. Additionally, several free tools were created in collaboration with commercial vendors to expose their file format openly. Ideally, it would be nice to have peer-review journals request that unaccessible file formats be published in two ways. First the raw, unadulterated files. Second, some form of a publicly accessible file format. This would enable the vast majority of users to freely access the data either with the vendor's own software or via the public format. Also, should eventually the plumbing code be established to actually access the raw file formats, then they would be usable.

In regards to the third argument, avoidance of junk. This is complete nonsense. It is not the place of a basic scientist to omit portions of data because they do not appear to be meaningful. A complete study might not be published due to it not supporting a desired effect; however, a portion of a published data set should never be omitted because the publishing scientist thinks that it is not informative. The peer-review process exists so that others can objectively re-evaluate data sets. Without the complete data set it is difficult if not impossible to do this task, and, the integrity of published data might be skewed. Computers continue to increase in processing power, algorithms continue to improve in performance, and bioinformaticians continue to grow our ability to mine data sets that would have seemed impossibly large previously. The task of determining junk from valuable data should be left to the community to decide in peer-review publication. It should also be freely questioned and critiqued same as any other basic science methodology.

Standardization of Statistical Methods – A related, final critique is that of standardization of statistical practices for determination of false discovery rates and objective high confidence peptide and protein identifications – a critique recently emphasized by several journals [44-46]. Several early proteomics publications and few more recent publications reported simple lists of peptide and protein identifications. Such lists of identifications are particularly difficult to validate given that one must guess at the parameters used by related software packages and that no statistical confidence was assigned to the identifications. Most journals have since migrated to a system of mandating that any peptide or protein identification be justified with objective statistics and that those statistics be clearly described. Due in part to these recommendations two false discovery rate estimation techniques have become commonly used, mixture models [47] and decoy analysis [48, 49]. Further efforts are also currently underway to standardize recommendations for statistical practices, namely the Human Protein Organization's Statistical Proteomics Initiative (HUPO SPI) [50].

In general, all of the aforementioned standardization of statistical methods have proven of questionable benefit to the proteomics community. The efforts may fruit in due time. However, any published standard will continually be subject to refinement and evolution. Much time will likely be required before anyone can objectively evaluate the success of existing standardization efforts. This opinion is in fact the primary motivation for the development of Tranche, Chapters 4 and 5.

The belief is that the most significant contribution that statistical technique standardization can provide is that of benchmark data sets and analyses. Given a proper tool to archive the raw data sets and results from proposed statistical analysis, others can much more quickly learn how to perform similar analyses. Additionally, journals can much more easily mandate comparisons and evaluate the techniques employed by researchers. Tranche seeks to properly serialized the raw data and analyses files for whatever statistical practices are proposed by the aforementioned studies. There is no reason that these files can not be saved now, and by saving the information, future studies can hopefully be greatly accelerated. The truth behind Tranche is that it reflects the belief that there will never be a final statistical standardization process. Instead, if the process of coming up with new MS/MS search algorithms and proposed statistical standards is made in to a commodity, then the community will more quickly reach truly significant developments. This final critique can more succinctly be stated as the following. The proteomics community seems preoccupied investing time in revising each others work versus accelerating community growth.

Introduction of Thesis Work

Motivation for the work presented in this thesis is clearly framed around mass spectrometry-based proteomics. By 2005, experimental procedures based on the original shotgun proteomics strategy had become widespread and related data sets were becoming much more accessible to researchers in the field of Bioinformatics. All work presented here was done with Dr. Andrews' basic science lab, National Resource for Proteomics and Pathways (NRPP), and Michigan Proteome Consortium (MPC). The Andrews lab conveniently had and continued to have several state-of-the-art mass spectrometers available during this thesis work. When discussing thesis projects it was clear that the project would involve gaining experience using mass spectrometers, high-throughput proteomics data processing, and dissemination of results to the community.

Early on it was decided that if possible the thesis work should avoid the obvious critiques of MS/MS proteomics at the time, most notably the lack data sharing and reinvention of MS/MS search algorithms. This mindset was of particular importance because both the NRPP and MCP were responsible for helping accelerate proteomics research and accessibility to proteomics services and software both in Michigan and nationwide. Implementation of restricted access tools, data sets, and reinvention of existing tools would likely not satisfy these goals well. Thus it was decided from the beginning that the thesis work would leverage personal prior experience, specifically open-source code development and web application (web site) development. Ideally, these skills could be used to start an objective survey for appropriate projects based on community feedback. Synchronously, training on analytical techniques and the experimental techniques related to shotgun proteomics could occur while the community survey was in progress.

Chapter 2 will introduce the initial survey step, namely ProteomeCommons.org [51]. Several related tools and sub-projects will also be summarized in that chapter. The work is important because it starts this thesis work off in a fashion that follows its own goals. Instead of diving directly in to development of the most obviously popular problem, processing MS/MS data sets, a fair evaluation of the community's resources and efforts is performed. Furthermore the entire evaluation is published in a public forum so that others may share in the knowledge I acquired only after building a complete website. Chapter 2 starts what will later be shown as a clear trend of attempting to accelerate the general proteomics community's growth versus pitting the author's intellect and ability to code software against that of other research groups.

Chapter 3 details an open-access reference data set, named Aurum [52], that was designed to aid in software algorithm development for shotgun proteomics. Aurum is also the first dataset published in Tranche [53] and a model peer-reviewed, public access data set. The entire process of generating and processing the Aurum data set is serialized on ProteomeCommons.org in a set of publicly accessible files. Same as with development of ProteomeCommons.org, the intention behind Aurum was to provide a data set that anyone else could easily use. Likewise, the result files from the MS/MS search algorithms used to process the Aurum data set are free for others to access. The intention is accelerate publications related to Aurum that claim benefits compared to the original analysis via reprocessing of published results. This type of benefit is in fact shown in Chapter 6 with discussion of the Bonanza manuscript.

Chapter 4 and 5 present Tranche and the technical details related to Tranche, which have become cornerstones of this thesis work. Tranche has thrived because it fills a critical niche in proteomics: making data sets accessible independent of file format or original software analysis. Tranche has been of particular importance because it also enabled the high-throughput data set analysis (Bonanza) in Chapter 6. Perhaps most importantly, Tranche enables any proteomics researcher to acquire and use the same data sets used for any study, not just the author of this thesis. This has proved extremely popular and resulted in rapid adoption by the community and journals. It has also made high-throughput proteomics data sets directly accessible to bioinformatics groups that do not even own a mass spectrometer.

The remainder of this thesis is broken into six chapters as mentioned previously. In short, the first chapter details ProteomeCommons.org, its contribution to nucleating an open-source community in proteomics, and work to survey the interest and needs of the proteomics community. The second chapter details work done in parallel with ProteomeCommons.org to develop an open-access, reference data set. The fourth and fifth chapters detail the implementation and design, respectively, of the Tranche project and the major impact it has had over a relatively short time period. The sixth chapter details a novel refinement to high-throughput proteomics data analysis, based on ideas designed to take advantage of the multitude of data sets stored in Tranche. The conclusion chapter brings closure to the discussion of this body of work. Three satisfying themes are full explored, first the successful inference of what type of tools would actually be widely used, second, the effect that Tranche has had on sharing public data and its consequences, and third, successful avoidance of 'reinventing' a MS/MS search algorithm in favor of refining and aggregating results from existing tools.

Chapter 2

ProteomeCommons.org

The diversity of biomedical problems to which proteomics technologies are being applied, coupled with the limited tools available, has lead many groups to develop their own scripts and software for analysis of proteomics data. This demand for new tools and the limited sources available has led to efforts by a number of laboratories to take advantage of the benefits derived from open source code development. This recent increase in open source projects for proteomics tools reflects the need for a broader range of tools and the reliance of proteome technology development on computational infrastructure. The open source effort has been paralleled by release of standard datasets and development of data format standards, both of which benefit algorithm and tool development. These aggregate efforts raise several issues currently being addressed by the proteomics community, including mechanisms for file standards development and support, data dissemination and annotation, and project organization and management. However, when I began my thesis work, all these efforts were in their infancies and we determined that a centralized resource could be developed that would help unify many of these efforts and provide access to the new resources that we and others were developing. ProteomeCommons.org was created as a tool to bring existing data archiving and dissemination functionality to the proteomics community in a fashion that requires minimal effort to use with existing code and data sets.

Introduction

ProteomeCommons.org [54] was placed on-line in 2004 and has grown to archive over 100 software projects and includes several hundred links to other resources. The basic site includes a simple web interface accessed via a web browser. The website currently receives over 5,000 visitors per week with over one thousand unique visitors.

ProteomeCommons.org was originally designed to help facilitate communication withing the proteomics community, act as a nucleation site for the open-source proteomics community, and to survey the field for existing tools, groups and data sets. Currently, ProteomeCommons.org is most often accessed for its aggregation of proteomics news, listing of available proteomics software, and indexing of proteomics software – primarily data hosted in Tranche. The site acts as a portal to many resources developed by our research group as well as many other groups and hosts several development projects.

Methodology

ProteomeCommons.org is designed to be as simple as possible to use. No registration is required and no extraneous information is mixed in with content. Users can use the entire website without needing anything more than their web browsers and access to all content is completely free. Combined with the free open access to content is an embedded peer review system. All content published on ProteomeCommons.org is manually verified by members of the proteomics community to ensure that the quality of content is as high as possible. This process is largely done by the National Resource for Proteomics and Pathways (NRPP) and volunteers. While it is not as stringent as most peer-reviewed journals, the process is designed to limit abuse of the resource and maintain a helpful level of service.

Key features currently available on ProteomeCommons.org include news aggregation, indexing of tools and links, free websites for projects, public archives for data, e-mail lists for communication, several open-source proteomics tools projects, distributed downloads and Google-like searching of content. The bulk of these features are based upon free tools and protocols that are commonly used in existing open-source software (OSS) communities. The website itself is coded in the Java programming language and JSP and hosted by the freely available Apache Tomcat web server, http://tomcat.apache.com. Most all of the projects supported by ProteomeCommons.org are also coded in Java with the exception of the Google provided e-mail lists, website usage tracking, and website indexing. Over the lifespan of ProteomeCommons.org it has become clear that many software packages exist for attempting to infer information from MS/MS data sets, and in contrast, a serious lack of software existed for interpreting results from such software, acquiring data to process by such software, and converting data into appropriate file formats MS/MS programs.

Largely due to the goals of the NRPP, ProteomeCommons.org has been able to become a hub for development of many desirable software services for the proteomics community. Several of these projects would have been very difficult, if not impossible, to justify independent funding. Primarily because many of the projects represent a logical step involved in solving or enabling research for the larger problem of high-throughput proteomics. Alone, the steps lack the luster associated with most basic science experiments, and when combined into a single projects, rarely were the steps polished significantly. Detailed here are three of the six primary sub-projects started on ProteomeCommons.org as part of this thesis work: the Java Analysis Framework (JAF) framework, the Input and Output (IO) Framework, and Peptide Finite State Machine (PFSM). Each of these projects was documented in publications in peer reviewed journals. The remaining projects, Tranche, Aurum and Bonanza, are covered in subsequent chapters of the thesis.

Java Analysis Framework (JAF)

The ProteomeCommons.org Java Analysis Framework (JAF) [55] provides a library of freely usable, open-source Java code that abstracts information regarding commonly used atoms, stable isotopes of atoms, residues, and modifications to residues. The code initially started as an application programming interface (API) for accessing this information and speeding up development of tools that relied on calculations such as the masses of peptides and proteins, SNPs of a protein sequence, theoretical isotope distributions of ions observed by mass spectrometry, and references for atomic weights and residue compositions. The JAF currently provides both the aforementioned programmer's API and several user tools.

The user tools provided by the JAF include mass spectrometrist-friendly HTML references for the common atoms and atomic isotopes, the common amino acids and known modifications of those amino acids and combinations of common amino acids, including mass shifts associated with residues on the N-terminus of C-terminus of peptides. In addition to the on-line HTML references the JAF provides a tool for dynamically finding combinations of residues that match a particular mass within a given mass tolerance. The JAF also provides a peptide calculator utility that looks just like a normal calculator, but can be used to calculate molecular weight of peptides, the mass of charged ions (allowing any charge) in mass spectrometry, the theoretical pI of peptides, and the fragments of the peptide's sequence assuming it was cleaved by any number of a given set of enzymes.

All of the user tools the JAF provides run directly on-line, through a web browser. The JAF takes advantage of the Java Web Start technology, which allows for robust, Java-based tools to automatically run directly on-line.

ProteomeCommons.org IO Framework

The ProteomeCommons.org IO Framework [56] is an freely usable, open-source framework for processing protein information and data produced by mass spectrometers. The framework initially started as a Java API that developers could use to convert between various mass spectrometer file formats, including MGF, PKL, DTA, mzData, mzXML, T2D, and more. The framework also provides a set of utilities for reading through sets of protein sequences saved in formats such as FASTA, and tools for manipulating proteins sequences in ways such as performing proteolytic digests, generating SNPs of protein sequences, and generating possible modifications of known protein sequences. In addition to the programmer's API the framework now provides user tools for performing conversion between different mass spectrometer output formats and dumping raw data into easily accessible formats such as mzData, mzXML, and plain-text. The primary data conversion tool is available on-line directly from ProteomeCommons.org and it requires no installation for users to be able to convert existing mass spectrometry data into a different format.

Peptide Finite State Machines (PFSM)

The peptide to protein inference is limited by many parameters with a significant hurdle being the time required to process a data set. Typically a MS/MS proteomics search engine will scan an entire library of protein sequences one at a time, modeled after Sequest [57], which can result in a prohibitively long period of data processing. This problem is of particular concern when considering larger data sets may contain hundreds of millions of protein sequences saved in a file that is of gigabytes in size. Ron Beavis and his optimization work with the X! Tandem search engine [58] is perhaps the most well known example of addressing this particular issue while not sacrificing the statistical sensitivity of the tool. Inspired by this work, research into creating a regular expression based pre-filter for MS/MS data analysis was done [59]. The work leveraged a practical computer science algorithm tactic involving suffix trees and construction of a regular expression to simultaneously search an entire set of spectra against a library of protein sequences in the same time as searching an individual spectrum. Figure 2-1 illustrates more intuitively the core concept with a cartoon.

If each MS/MS spectrum is treated as a fragmented set of amino acids, it is possible to convert an individual spectrum into a single regular expression [60] that accounts for all theoretical fragments. A convenient characteristic of regular expressions is that multiple regular expressions can be combined into one regular expression and applied to any set of input strings, e.g. protein sequences, with the same efficiency of a single regular expression. Thus, an entire set of MS/MS spectra can readily be combined down to a single regular expression that can filter a protein database for relevant sequences. The final step is for the MS/MS search engine to be applied to this filtered, presumably much smaller, set of sequences in order to infer the most correct peptide and protein identifications.

The PFSM strategy was demonstrated to work well on shotgun proteomics style data sets generated from an MALDI TOF/TOF mass spectrometer. Search time requirements were shown to drop from hours to a few minutes, and the majority of resulting peptide and protein identifications remained the same. The work is certainly successful; however, it was quickly discovered that many MS/MS search engines rely on statistics derived from the entire set of protein sequences analyzed. Utilization of PFSM as a pre-processing step for these search engines could significantly change the search results. In order to truly realize the benefit of PFSM pre-filtering use of the algorithm would have to be restricted to particular search engines or a custom statistical analysis would need to be developed. A strong desire still persists to avoid development of yet another MS/MS search engine, and this effort was not pursued with PFSM. Rather efforts were initiated to work with existing MS/MS search engines in a method that would leverage existing statistical scoring algorithms.

Conclusion

ProteomeCommons.org has established itself as a beneficial resource for the proteomics community. The website brings several modern tools for collaboration and data dissemination to researchers in proteomics. ProteomeCommons.org also acts as one of the largest and most comprehensive listing of existing proteomics on-line tools and software packages. Another significant benefit of the website is that it acts as a sponsor for several open-source software efforts aimed at building freely accessibly tools for common proteomics-related work. Many of these tools have been used by several different research groups. Most notable of such projects is the Tranche project, which currently acts as the repository for thousands of proteomics data sets that have been published and indexed by ProteomeCommons.org.

Chapter 3

Aurum Data Set

A current focus of proteomics research is the establishment of acceptable confidence measures in the assignment of protein identifications in an unknown sample. Development of new algorithmic approaches would greatly benefit from a standard reference set of spectra for known proteins for the purpose of testing and training. Here we describe an openly available library of mass spectra generated on an ABI 4700 MALDI TOF/TOF from 246 known, individually purified and trypsin-digested protein samples. The initial full release of the Aurum Dataset includes gel images, peak lists, spectra, search result files, decoy database analysis files, FASTA file of protein sequences, manual curation, and summary pages describing protein coverage and peptides matched via MS/MS followed by decoy database analysis using Mascot, Sequest, and X!Tandem. The data is publicly available for use at ProteomeCommons.org.

Availability

The Aurum Dataset is freely available for use in its entirety from ProteomeCommons.org. On-line versions of the data may be found at http://www.proteomecommons.org/current/553/index.html.

The ProteomeCommons.org Tranche network is used to provide fast downloads of the data and to get a verifiable, exact copy of the data described by this manuscript. The Tranche hash for the Aurum Dataset is given below.

HnxUzQuuP7BIqF10aetLtjwnffOwuOMAfDvg2BFmenNe9UeMgprBFh7+
wtpbcWnXqMk2KY8z9VjmwqXYDbQ0pTNqIx4AAAAAASJlaw==

Further information about Tranche and how to use this hash may be found on-line at http://www.proteomecommons.org/dev/dfs/.

Introduction

Tandem mass spectrometry (MS/MS) of peptides is currently the primary method to identify proteins in complex samples. Search programs such as SEQUEST [61], Mascot [62], X!TANDEM [63] are some of the most widely used software packages to identify the most likely peptide sequence to match an MS/MS spectrum. Development of better MS/MS identification tools is an active area of proteomics research [64-67], MS/MS de novo tools [68-70], MS/MS spectral search tools [71,72] and MS/MS search result refinement tools [73,74]. All of these tools rely on libraries of well-studied MS/MS spectra from a variety of instruments with accurate peptide assignments.

Accurate peptide assignments are essential but manual confirmation is a time-consuming process that is also subject to some degree of operator dependence, and it is not feasible for high-throughput proteomic analysis. Most commonly, MS/MS algorithms are trained on in-house generated data sets that have undergone a variety of selection criteria to verify their authenticity. These standard sets are often obtained from analysis of commercial protein preparations with limited criteria for purity or represent bootstrap efforts that set stringent criteria for results from existing search engines. Recently, several approaches have been proposed to accurately estimate false-positives and associate peptide identifications with MS/MS spectra with high levels of confidence [75, 76]. Development of MS/MS related algorithms and tools would greatly benefit from publication of third party data sets, particularly well-annotated data sets using these proposed approaches to estimating false-positives with as much manual confirmation as possible. Finally, the availability of well-verified sets of MS/MS spectra can provide the basis for direct spectral comparison, which has the potential to be a much more effective approach to peptide identification that existing engines that match against generated MS/MS spectra and obviates the need for an accurate fragmentation model.

Small reference sets of tryptic peptides have been made from known proteins [77,78] and larger datasets have been made from the yeast proteome [79,80] and human serum proteins collectively in the HUPO initiative [81]. While these are useful databases, they are time consuming to generate, and are not all publicly available as a reference set.

In this manuscript we describe a publicly available library of tandem mass spectra generated on an ABI 4700 MALDI TOF/TOF from 246 known purified and trypsin-digested protein samples using a work flow used for gel-purified proteins. The data are analyzed using the Mascot, X!Tandem, and Sequest search engines, and peptide identifications are adjusted to 99% true-positive confidence using the intuitive decoy database approach described by Elias et al. In addition to the peak lists and associated peptide identifications, the described data set is also published with the raw spectra, search result files, decoy database analysis files, Scaffold analysis files, and the gel images used when checking for protein purity.

Materials and Methods

Proteins - A selection of 300 sequence-verified recombinant human proteins (8-70 kDa.) were obtained from GenWay Biotech Inc (San Diego, CA). The proteins were selected by GenWay Biotech based on clones that could readily be over-expressed and purified. GenWay Biotech provided the sequence verification services and a report for each cloned sequence is included in GenWay's product documentation, which is referenced by the per protein report included in the on-line Aurum documentation. After purity analyses, 246 of the 300 proteins were used in the Aurum analysis. The proteins contained an N-terminal T7 tag (MASMTGGQQMG also observed as ASMTGGQQR) or His6 tag (HHHHHH) and were expressed in *E. coli*. Documentation provided with the proteins included the name, expressed length and the NCBI accession number. Proteins (2 µg/lane) were analyzed for purity by SDS-PAGE stained with colloidal Coomassie G-250 (Figure 3-1). The criteria for purity were that at least 50% of the protein was at the correct size, the gel lane contained no nearby unrelated protein, and at least 95% of the tryptic peptides corresponded to the anticipated protein. Images for each of the gels are included in the supplementary data, and shown directly on individual protein summary pages (Figure 3-2).

For each protein in the Aurum data set a unique GS-number was assigned where we used the letters “GS” followed by four digits representing a decimal number assigned to the protein. The supplementary data includes a table that maps this GS number to an appropriate GI number, NCBI accession number, and Swissprot accession number. The GS nomenclature is not intended to represent a new standard for referring to the associated protein sequence. Rather it is a convenient way to unique identify Aurum proteins for internal use within the dataset – independent of accession number or identifier changes that might occur in other databases. GS numbers will remain static throughout the lifespan of the Aurum dataset; however, use of the GS numbers outside the context of analyzing the Aurum data is discouraged when either NCBI or Swissprot identifiers are available.

Protein coverage calculations – The entire protein sequence is included in each summary file as shown in one file in Figure 2-2. The sequence is further colored in order to indicate the peptides that were identified and the portions of protein sequence that are not expected to be identified. The portions that are not expected to be identified are those that have a m/z at +1 charge of less than 900 Da or more than 2500 Da, i.e. the range that the mass spectrometer is configured to ignore. Data analysis for this manuscript is based off a MALDI instrument, and the +1 charge state is almost ubiquitously observed for ionized peptides. Thus the theoretical m/z of an ionized peptide is well approximated to be its molecular mass. The range of 900 to 2,500 Da is selected for three primary reasons. First, MALDI instruments are prone to ionizing matrix clusters that can dominate the lower mass region, which often makes it very difficult to identify anything below the mass of 900 Da. Second, the instrument used for this analysis is tuned to most accurately identify ions with a m/z of 1,800 Da. Ions with much less or much greater m/z may report an incorrect m/z to the point where it is difficult to use in data analysis. Third, the MALDI TOFTOF does not detect higher mass peptides as well as lower and the trade off between higher m/z and the amount of sample required to detect the ion appears to be non-linear. Thus, peptides with masses higher than 2,500 are problematic due to both relatively inaccurate m/z measurements and relatively poor signal strength.

Explanation of protein coverage is important because the results section and included protein reports present two types of protein coverage information. The first type of protein coverage is a strict percentage of the total protein sequence that is covered by observed peptides. The second type of coverage, named 'expected protein coverage', is the percentage of tryptic peptides that fall within the 900 Da to 2,500 Da range, i.e. the peptides one might expect to see based on the Aurum data acquisition parameters.

In-Gel Tryptic Digestion - Excised gel plugs (0.67 µg protein) were placed in 96-well plates and were processed using a MassPrep robotic workstation (Waters). The plugs in the presence of 50 mM ammonium bicarbonate underwent the following steps: wash/dehydration with 50% acetonitrile; reduction with 10 mM DTT; alkylation with 55 mM iodoacetamide; wash/dehydration with 50% acetonitrile; digestion for 4 hours with trypsin (200 ng, porcine, modified, Promega). Peptides were extracted from the gel plug with 1% formic acid/2% acetonitrile and concentrated using C-18 ZipTips (Millipore). Digests were spotted (4 replicates) on a MALDI target using α-cyano 4-hydroxy cinnamic acid (2 mg/ml in 50% acetonitrile, 0.1% TFA containing 10 mM ammonium phosphate) as matrix. Dilutions of the digests were made at 1/8 and were spotted in the same manner.

MS/MS acquisition - Spectra were acquired on a 4700 MALDI TOFTOF mass spectrometer (Applied Biosystems). Spectra were acquired for the 8 most intense ions. In a replicate well, after excluding the 7 most intense ions, the next 8 most intense ions were analyzed.

Similarly, the next set of 8 ions was analyzed for wells 3 and 4. Known trypsin auto-digestion peptides were excluded. This process resulted in acquisition of a maximum of 32 spectra per digest, theoretically 29 unique spectra if sample and MS intensities do not change between spottings.

Data curation – Default peak lists from Applied Biosystems GPS software were taken from replicate wells and were concatenated into a single Mascot Generic Format (.mgf) file. In order to map spectra back to the original files the the base 16 encoding of individual file's MD5 [82] hash was set as the MGF file's TITLE field. Additionally, all peak lists were converted into a set of .dta files by the ProteomeCommons.org IO Framework [83] for subsequent analysis by the Sequest. Four different initial searches were performed, each using 0.5 Da for the parent and fragment ion mass accuracy and with oxidation (M,H,W), deamidation (N, Q) variable modifications. Iodoacetamide (C) was specified as a static modification. Four follow-up searches were performed using the similar parameters but without iodoacetamide as a static modification in favor of setting iodoacetamide and propionamide as variable modifications for the side-chains of cysteine residues. Three of the four searches used different search engines in an attempt to identify as many of the spectra as possible. The two same-search engine searches both used Sequest but included the variable n-term protein modifications for each of the two purification tags.

The MS/MS searches were performed using Mascot, X!Tandem, and Sequest. Two of the initial searches used the concatenated .mgf file. One search on Mascot 1.9 and the other search on X!Tandem 06_9_15. X!Tandem did by default include N-term pyro-glu from N and

Q as modifications. The other two initial searches used Sequest. One search assuming each protein had an N-term T7 tag and the other search assuming each protein had a N-term HIS tag. All searches were performed on a decoy database version of the IPI Human FASTA file version 3.14. The decoy database was the exact IPI Human 3.14 FASTA file with a concatenated reverse version of the same database. Each protein in the reverse sequence is noted by appending an “R” to the protein's accession number and each protein is changed by reversing the order of the amino-acid residues. The ProteomeCommons.org IO Framework was used to generate the reverse database.

Identification of peaklists was based on the decoy search strategy outlined by Elias et al. [84] and described briefly here. Each search engines peptide identifications were individually ranked according to the respective following scores: Mascot's ion score, X! Tandem's hyperscore, and Sequest's XCorr. Each sorted list is then filtered to only include matches that scored above a 99% confidence threshold determined as follows. All peptides above the score are binned into two categories. Those that are from the normal FASTA sequences (i.e. matches without a “R” in the accession) and those that are from the decoy sequences (i.e. matches with a “R” in the accession). The false positive rate is estimated to be twice as much as the ratio of decoy sequences versus normal sequences – twice because the decoy sequences only represent half the total database thus approximately as many normal sequences are likely inaccurate. An example would be the case where 198 normal sequences were identified per every 1 decoy sequence, where (1 * 2) / 200 yields a 1% false positive rate aka 99% confidence in individual peptide identifications. This strategy is presented as an appropriate objective analysis of the data set that takes advantage of individual expertise present in different search engines while still normalizing all search results to approximately 99% confidence in true positives.

Files used by the respective search engines, including search parameter files, peak list files, and FASTA files are included with the on-line download as described in the availability section.

Results and Discussion

A well documented set of purified human recombinant proteins has been procured and analyzed using a routine gel-based protocol to generate a library of mass spectra referred to as the Aurum Dataset. At present the Aurum dataset consists of 246 recombinant human proteins that have been trypsin-digested and characterized by MALDI TOF/TOF. The MS/MS dataset further underwent what is intended to be an objective, community-standard based analysis to generate spectra-associated peptide identifications and protein coverage information.

The recombinant proteins were expressed in *Escherichia coli* and initial isolation performed by the vendor. Upon receipt, the proteins were analyzed by SDS-PAGE for purity and the dominant bands were excised for in-gel tryptic digestion. Figure 1-1 shows a representative gel, where samples GS0372 and GS0376 represent the highest purity provided and samples GS0312 and GS0256 represent a moderate-low purity. Of the 246 proteins, 181 were represented as a single band and were classified as high purity. An additional 21 were represented as evenly distributed doublets. Analysis of both bands of the doublets confirmed that both were forms of the target protein and could be placed in the subset of high purity proteins that would be suitable for future in-solution digests. The remaining proteins had varying degrees of contaminating bands ranging from possible truncation products to *E.*

coli proteins. Only the predominant band was excised and characterized. Only proteins for which all tryptic peptides returned the correct ID were included in the final protein list. Gel images are included for each protein in the protein's summary page found with the on-line documentation for the Aurum Dataset.

MS/MS analysis for the selected gel bands was carried out according to the standard analysis procedures described in the methods section. Seven different MALDI plates were used, found in the documentation with the names “T10467”, “T10475”, “T10622”, “T10645”, “T10707”, “T10739”, and “T10761”. Each protein was spotted individually at least four times to help ensure the best chance of acquiring high-quality spectra for as many of the peptides associated with each protein as possible. At least 32 spectra were acquired for each protein by collecting data from four separate spots of the protein as described in the methods section. Default peak lists of the spectra were then extracted for analysis using the MSExtractor tool (http://www.proteomecommons.org/current/489) and concatenated using the ProteomeCommons.org IO Framework. 9,987 total peaklists are included in the resulting .mgf file. Each peak list is identified by the original file's MD5 hash listed in the TITLE field of the associated peak list in the .mgf file.

Decoy database analysis targeting 99% true-positive confidence and using Mascot, X!Tandem and Sequest were performed according to suggested guidelines published by Elias et al. The analysis is not intended to be a comparison of the search engines used, rather it normalizes the results of each search engine to an approximated 99% true-positive confidence. All of the peptides from the 99% true positive results were aggregated to make the set of all identified spectra. 5,054 unique peptide sequences (>50% of all peak lists) were identified at 99% true-positive confidence with the peptides being identified coming from the initial and follow up searches described in the following format (initial search)/(follow up search). Note that the follow up search is not intended to identify a superset of peptides and the notation does not indicate a fraction. Mascot identified 1,682/1,847 peptides, X! Tandem identified 441/424 peptides, and Sequest identifying 2,939/2,921 peptides for the T7 tag and 2,937/2,920 peptides for the HIS tag searches. No search engine identified a superset of all others and a significant gain in highly-confident identifications was obtained by combining the three search engines. These results appear to support the use of decoy database analysis with multiple search engines as an approach to identify more spectra from a dataset; however, it is worth emphasizing that these results are not intended as a basis for comparison of the search engines used. Various search results are expected as each search engine performs analysis differently even with the similar settings we used in each search. Additionally each search engine has a disparate range of settings that might be optimized to change the results of the analysis. The set of search engines used is intended only to help increase the number of unique spectra identified. For further analysis, the same search result files used for decoy database analysis were imported into the Scaffold software package (ProteomeSoftware Inc., Portland, OR) for PeptideProphet and ProteinProphet-like analysis. Similar results as for the decoy database analysis were found and the free Scaffold Viewer program may be used to examine the Scaffold files included with this manuscript.

Of the 246 purified proteins 242 proteins had at least one peptide identified to the expected cloned sequence and 233 peptides had more than 2 peptides matched to the expected cloned sequence. At the most, up to 19 unique peptides matched to a cloned protein. The average protein sequence coverage from this analysis was 32% and the average protein coverage of theoretically detectable peptides is 63%. Theoretically detectable peptides include those that have a unmodified m/z of more than 900 Da and less than 2500 Da – m/z restrictions specified at the time of data collection. A summary of protein coverage and matched peak lists are provided for each protein in the supplementary files described by the availability section. Figure 2 shows an example protein summary. Analysis of the summary files illustrates that the majority of proteins have several peptides that may be used to identify them from a biological sample assuming that similar quantities of the purified protein and or or peptides can be obtained. Although some of the proteins have very few, if any, peptides that are readily observed. These proteins are of interest for further study as they may represent proteins that are difficult to identify from a potentially more complex sample using a similar MALDI TOFTOF based approach. A simple explanation for several of these difficult to analyze proteins is that they are relative small proteins with very few, if any, tryptic peptides that fall within the 900 to 2500 Da cutoff used when analyzing this data set. Potentially a different mass spectrometer such as a ESI-based instrument or a different digestion enzyme would provide a more favorable analysis. Other plausible explanations could account for the difficulty in analysis of other proteins such as poor ionization of the peptides, unfavorable experimental protocols for analyzing the particular protein, or even experimental error. In any case, the set of poorly identified proteins may be of interest for further analysis to identify if they are indeed poor candidate proteins for mass spectrometry analysis.

Further data analysis and summary reports were generated to check for common contaminants in mass spectrometry experiments. The crap (pronounced “cee-RAP”) 1.0 list of proteins maintained at TheGPM.org was searched against the unidentified peak lists. The crap list contains approximately 100 proteins including common laboratory proteins, proteins added by accident through dust or physical contact, and proteins commonly used as molecular weight standards. The crap analysis was performed using just the X!Tandem search engine, and 28 proteins were found with more than 2 peptides matching. In all, 37 proteins were identified by 1 or more peptides. The proteins primarily identified included many keratin proteins and several recombinant E. Coli proteins. BSA and Serotransferrin where also found. The complete list of crap proteins, the X!Tandem search results, and a set of summary pages similar to Figure 2 for the crap proteins are included with the supplementary data.

Conclusions

The Aurum Dataset is a high quality dataset of known proteins analyzed by a MALDI TOFTOF. The proteins are all human proteins expressed in E. coli and purified by N-terminus T7 and HIS tags. The proteins further purified using SDS PAGE, individually digested with trypsin, and individually spotted 4 times on a MALDI plate. Data was acquired to represent at least the top 29 most intense MS peaks, and published decoy database analysis was used to identify more than 50% of the acquired spectra, approximately 5,000 unique peptides. Based on this analysis the majority of proteins can readily be identified, but a range exists where some proteins are not as easily analyzed. The low end of this range is of particular interest for further analysis as it might be helpful for identifying why certain proteins are more difficult to identify from complex samples.

The Aurum Dataset is a valuable contribution for testing existing MS/MS algorithms and tools, and the Aurum Dataset will be helpful as an objective third-party data set for developing new tools and algorithms. The published data set contains all raw and curated data used to generate the analysis described by this manuscript and all of this data is openly and freely available for use.

Chapter 4

Tranche

Facile access to scientific data is a general problem in research that is of particular concern to post-genome fields, including Proteomics. Current technologies can generate very large quantities of data and this rate of data production is rapidly increasing. Most proteomics studies are targeted to specific goals and information extraneous to these goals, yet present in the datasets, are not pursued. A key question has been, how can these very large and useful data sets be shared and properly cited? The field of Proteomics provides a clear example of coping with this data sharing issue, and the tactics used are potential solutions for other fields facing similar problems. Presented here is a research project, that addresses the scientific data sharing problem from the perspectives of open-access, community based distributed storage. The software implementation of these concepts, named “Tranche”, represents a radical change from previous approaches for data sharing in the field of Proteomics. Tranche provides a scalable, secure mechanism for partitioning the responsibility of the data sharing problem across available bioinformatics resources in the entire proteomics community. These properties enable two critical features: very large data sets can now be shared and any data set can be accurately cited and validated as unchanged since publication. Tranche also allows individual laboratories to comply with guidelines on data accessibility proposed by leading proteomics journals.

Introduction

Sharing large amounts of data and software is a legitimate need in the field of proteomics because replication of search results and reanalysis of data rely on access to the original data. Proteomics studies are increasingly large, relatively expensive, and, when patient samples are involved, often deal with irreplaceable samples. Additional information can be gleaned from these large datasets that can be valuable for other research efforts. It is important to archive and share such data, particularly publicly funded data, in order to allow replication of results and reanalysis with new proteomics software, which itself is rapidly evolving and improving. Logically this point is straightforward to argue; however, researchers in the field of Protemics have additional motivation due to recent guidelines by three of the leading journals: Nature Biotechnology, Molecular and Cellular Proteomics (MCP), and Proteomics.

A Nature Biotechnology March 2007 editorial [85] succinctly stated, “Beginning this month, Nature Biotechnology is recommending that raw data from proteomics and molecular-interaction experiments be deposited in a public database before manuscript submission.”

Carr et al. in Molecular and Cellular Proteomics have repeatedly emphasized similar points, but most clearly the 2004 [86] and 2006 guidelines [87] emphasize, “MCP strongly encourages (but does not at present require) the submission of all MS/MS spectra mentioned in the paper as supplemental material.” This is in addition to the conceptual guidelines outlined requiring sufficient information to document the search engines used and how peptides were identified.

Proteomics also elaborated on similar recommendations in guidelines described by Wilkins et al. [88] where academic databases and software used must be freely available for use, and furthermore “Supplementary material is encouraged. This includes protein identification results, expression data, and mass spectrometry peak lists.” Concluding with the point that such material will not appear in print but on-line via the journal's website.

A point not clearly detailed in these guidelines is how exactly proteomics research data (this data being a potential mega-data set) can be placed on-line and later accessed as needed. The general trend is that a few megabytes of data, ideally annotated spectra and peak lists, can be submitted as supplemental data with a manuscript, but the journals generally do not provide a mechanism to archive gigabytes of supplemental peak lists, raw mass spectra, and related files for indefinite public access. Individual researchers are left to solve this problem themselves in a variety of ways. Nature Biotechnology does recommend a few possible public databases, including Tranche, but no concise requirement for what constitutes published supplemental data and how to cite and access such data is provided.

Clearly and simply the goals of the Tranche project are now stated in relation to these journal guidelines.

1. Tranche can freely and publicly host data sets of any size. Downloading and uploading data to the Tranche network is limited by the speed of an individual's internet connection.

2. Every data set has a single, unchanging “Tranche Hash” that should be used for citation. This identifier is independent of the

physical storage location of the data (i.e. not a URL) and the identifier verifies that data has not changed since publication. The identifier can also be calculated using standard hashing algorithms, independent of Tranche if desired.

3. Data can be archived for a reasonable amount of time in the Tranche repository. At least several years, but potentially indefinitely.
4. Storage is independent of file format or directory structures.

Conceptually, these points are all a researcher need know about Tranche to use it to publish data with a proteomics manuscript. Obviously, hosting “data sets of any size” comes with the caveat that physical disk space must exist to store the data; however, Tranche has been in operation for over a year and has easily hosted thousands of data sets including 2005's “largest and most ambitious” [89] data sets. Finally, point 2 is of particular note. Tranche not only hosts data but does so in an ideal way for scientific citation that also proves data hasn't changed since publication – an obvious benefit to peer-reviewed data, yet one that is not explicitly mentioned in any of the aforementioned guidelines.

Mass Spectrometry Proteomics Data Repositories

Several efforts exist for hosting proteomics data and annotation information. These efforts significantly differ from Tranche although not necessarily in incompatible ways. Notable centralized repositories such as PeptideAtlas.org [90], and the Open Proteomics Database (OPD) [91] host various amounts of raw and annotated proteomics data. In addition, some information management systems, CPAS [92] and PRIME (https://prime-sdms.org) allow dissemination of discrete datasets. In these examples, the data hosted often comes primarily from local collaborators and local installations of information management systems. More importantly, all data are hosted by a centralized framework. Scaling a centralized database to handle many terabytes of raw data can be a challenge, requiring large investments in both storage capacity and network bandwidth to maintain effective performance characteristics. Additionally, all data sets hosted by such a centralized repository have the disadvantage of relying on the owner of the repository(s) to keep it properly on-line. Other notable centralized repositories such as TheGPM [93], PRIDE [94], and the HPRD [95] exist for hosting filtered versions of raw data and annotations of raw data. These resources are quite valuable, but for practical reasons, distance themselves from coping with the problem of storing the raw data associated with a proteomics experiment. The overhead associated with maintaining a centralized database for raw spectra is significant, can interfere with performance and compete for resources better dedicated to development of higher level functionality.

Tranche was developed in June 2006 and similar to all of the previous tools, the goal was to share proteomics data; however, Tranche is not tied to a specific software package for processing data or to any particular file formats. Rather Tranche was designed for the sole purpose of sharing large sets of files and providing a citation mechanism suitable for peer-reviewed scientific literature. Furthermore, the intention of Tranche was that it be freely used to both mirror and provide data to the repositories previously mentioned, ideally letting groups more interested in higher-level data processing to not have to worry about storage and transportation of raw data. Currently Tranche is used in various ways, including as a data repository, by the majority of the above resources as indicated in the example collaborations section of the on-line Tranche documentation.

Results

The ProteomeCommons.org Tranche network went on-line in June 2006 and after a year of use has approximately 4,500 data sets on-line representing close to 2 terabytes of compressed data occupying physical disk space. The network consists of 17 dedicated servers spanning the globe with an aggregate storage capacity of more than 60 terabytes. Development of Tranche is funded by NCRR and primarily performed at the University of Michigan; however, it is an open source project and multiple labs have established Tranche servers that participate in the ProteomeCommons.org network. Summarized here is the existing core network of Tranche computers, the data hosted on the Tranche network, and adoption progress of Tranche in the proteomics community. More comprehensive documentation regarding all of these topics is maintained in the on-line documentation at http://tranche.proteomecommons.org.

Core server development of the Tranche network

Tranche is a Free Open Source Software (FOSS) project where the model of use is that one group takes primary responsibility for code development and maintenance but use of the Tranche software is completely free and others are strongly encouraged to participate and reap the benefits of a system designed by bioinformatics experts [96]. Initial computer clusters were established around the State of Michigan as part of the National Resource for Proteomics and Pathways (NRPP). Subsequent storage resources were established by other proteomics groups including those working on PeptideAtlas.org [97], GFS [98], and the Human Proteinpedia (humanproteinpedia.org). Additionally large scale data collection efforts including the ABRF, HUPO, and NCI MMHCC and CPTAC have provided resources for expanding the storage capacity of existing Tranche servers in order to adequately store data for the associated study.

The Tranche homepage keeps an updated Google map of the servers currently participating in the Tranche network, along with approximate storage capacity and use per site. Figure 4-1 provides a snapshot of this map as an example. It generally does not display users of Tranche that primarily seek to download data. Such users can also act as servers to share copies of data that are downloaded, but generally the map only shows dedicated servers.

Data currently available from Tranche

Tranche has no specific file format restrictions, but the majority of the data sets currently in Tranche are derived from mass spectrometry. This includes files generated directly by mass spectrometers and vendor-specific software programs, processed peak list files, output files from MS/MS search algorithms, lists of identified peptides and proteins, and other files related to mass spectrometry based proteomics studies. Typically all files or a subset of files associated with a single peer-reviewed manuscript are put into a single directory and that directory is uploaded as a data set. Figure 4-1 provides a breakdown of the file types by size that are hosted by Tranche. The majority of files are comprised of .raw files (Thermo Finnigan Scientific, Waltham, MA), .mzXML files [99], .DAT files in .raw directory structures (Waters, Milford, MA), and .tgz files that are primarily the output of Sequest [100] searches. All well-known file formats related to mass spectrometry based proteomics experiments.

Tranche adoption for proteomics data collection and sharing

Throughout the past year, the development version of Tranche has been used for multiple large-scale data collection efforts. Tranche currently contains a complete publicly accessible copy of the ambitiously large HUPO PPP data set [101] and the even larger NCI Mouse Proteomics Technologies Initiatives (MPTI) data set (http://proteomics.cancer.gov/programs/mouse/), including the ability to download portions of the entire data sets. Tranche was used as the primary tool for data collection in the Association of Biomolecular Resource Facilities (ABRF) 2005 and 2006 sPRG studies (http://abrf.org) and is currently being used for data collection in the NCI Clinical Proteomic Technologies for Cancer (CPTAC) project (http://proteomics.cancer.gov/) and the HUPO 2007 study (http://hupo.org).

Tranche has also aided in the collection and publication of data sets referenced in peer-reviewed literature. The recommended use is that researchers publish their data prior to submission of a manuscript for peer-review, similar to proposed journal recommendations. Many groups have elected to do this and several examples are maintained in the on-line Tranche documentation. Many more data sets have been added post-publication of manuscripts. A data collection effort continues at ProteomeCommons.org where many proteomics journal research articles are continuously scanned for new data sets that could be added to Tranche. The set of journals scanned includes Nature Biotechnology, Molecular and Cellular Proteomics, Bioinformatics, The

Journal of Proteome Research, and Analytical Chemistry. Currently more than 1,200 articles have been scanned for data sets and several hundred of the datasets have been requested. A database representing this effort is maintained at http://www.proteomecommons.org/data.jsp.

Discussion

It is rapidly becoming feasible to share and maintain large amounts of proteomics data in the public domain – potentially indefinitely. Tranche clearly illustrates that the majority of public proteomics data can be hosted on-line in both secure and public access forms. Additionally, Tranche is free to use either on dedicated computer hardware or as a free data sharing service supported by participants in the ProteomeCommons.org network. Tranche should be considered one viable mechanism for publishing proteomics data sets, even very large data sets, that complement peer-reviewed manuscripts. Tranche certainly does not satisfy all the aforementioned guidelines set forth by proteomics journals; however, Tranche provides a critical capability in disseminating as little or as much proteomics data necessary to satisfy those guidelines.

Open-access is emphasized heavily in the design of Tranche. This holds true to both how Tranche shares data and how the source-code was developed and is maintained. Data can easily be accessed from Tranche by anyone, and if desired, all data on Tranche can be openly migrated as desired to other resources – potentially even a superior data sharing system. Within Tranche, users are able to chose if they wish to support the network, including the ability to host a copy of all of their data and more, or if they wish to take advantage of resources provided by others in the community. No centralized repository is required and the network itself does not lose data if a server goes off-line. It represents a unique tool for open-access sharing of scientific data. Any focused scientific community can freely use Tranche to efficiently share data to the limits of current computer storage and network speeds, and while Tranche supports flexible annotation and revision of data, nothing locks data into staying only in Tranche. The entire system is a clear example of true open-access scientific data sharing for data sets of all types and sizes.

Methods

Tranche was developed as a free open-source software package programmed in the Java programming language. Tranche is based on agile development philosophies for creating robust code and standardized e-commerce encryption algorithms to guarantee privacy of shared data and to prevent abuse of servers supporting the Tranche network. Complete source-code and documentation for all of the features included in Tranche are available from the ProteomeCommons.org Tranche website, http://tranche.proteomecommons.org. A primer describing how the core functionality in Tranche works is included in the supplemental text for this manuscript.

Bi-monthly user meetings are held on-line for individuals interested in using Tranche, developing Tranche code, reporting problems with Tranche, or in learning more about how Tranche works. These meetings are also podcast and are available from Tranche or iTunes. Specific topics of interest are also recorded and published in Tranche and on YouTube.com. See the Tranche website for details.

Chapter 5

Tranche Implementation and Technical Details

Tranche is a peer-to-peer (P2P) network for sharing scientific data. This does mean that expanding Tranche's capacity to share data is as straight-forward as adding more computers that run Tranche. Theoretically, virtually unlimited amounts of storage space can be made available at extraordinarily fast transfer speeds; however, practically, the P2P paradigm enables a good storage and data transfer mechanism that works with existing computers and can easily grow as improvements are made in data storage technologies and data transfer media. Tranche's use of P2P does not mean that it should be equated to using Bittorrent, Napster, or other popular P2P programs associated with non-scientific file sharing, often inclusive of illicit data sharing. Tranche was designed from first principles to leverage the benefits of P2P for scientific data publication but avoid the stigmas and potential risks associated with certain popularized P2P networks. Furthermore, Tranche is designed specifically for sharing scientific data, namely: the ability to explicitly cite published data, the ability to verify that data hasn't changed since publication, and the ability to prevent potential abuse of networked computers including illicit or dangerous files.

This primer is intended to be a comprehensive introduction to how Tranche works; however, many technical details intended solely for computer programmers are left to the on-line Tranche documentation at http://tranche.proteomecommons.org. Additionally, the Tranche source-code and associated unit tests provide ideal examples for those interested in how particular features are implemented.

How Data is Evenly Spread Across the Tranche Network

Computers participating in a Tranche network may have various amounts of disk space, potentially less space than is required to host a complete proteomics data set. In order to take advantage of all possible disk space and to evenly spread data across the Tranche network, all files are split into one megabyte (1024*1024 byte) chunks – a size smaller than modern hard drives. These chunks are identified by a scheme that essentially creates a random identifier (see *How Tranche Verifies Data*) for each chunk that is a fixed length and essentially represents a number between 0 and 10^{181} (approximately 8 exabytes). Computers on the Tranche network are split to take spans of all possible identifiers based on the relative amount of available disk space. For example, assume three computers are on-line with the first two having twice as much disk space as the third. Tranche would split the data that these computers store in the following way. The first computer would handle all data with any of the first 2/5 possible identifiers. The second computer would store all chunks that have next 2/5 of possible identifiers. The third computer would store all chunks that have the final 1/5 of remaining identifiers.

The assumption is that if the Tranche chunk identifiers are random, this means that each computer on the network has a chance of being required to store the chunk relative to the size of the span of possible chunks it is responsible for. Thus, all computers will fill up equally with 1MB chunks of data regardless of the actual size of space available on any one computer. Chunk identifiers are named “Tranche Hashes” or “Tranche Hash Strings” in the Tranche documentation and the span of all possible chunk identifiers a computer will hold is likewise named a “Tranche Hash Span”. Normally, all data on Tranche is replicated at least 3 times. In order for this to occur, hash spans are split so that at least three different computers on the network will be responsible for any given hash span.

Data uploaded to Tranche is split into 1MB chunks using the scheme illustrated in Figure 5-1. In order to make the most efficient use of space, data is always compressed, normally using the GZIP algorithm. In cases where data must be kept private for a period of time, data is next encrypted using the NIST AES 256 standard for data encryption. Finally, the resulting files are split into 1MB chunks and evenly spread across all Tranche servers as described above.

It is worth noting that the compression and encryption encodings shown in Figure 5-1 are not the only encodings Tranche can provide. Identifiers in Tranche are based solely on the un-encoded data, meaning that at a later time Tranche can arbitrarily change compression, encryption schemes, or any other encoding without invalidating existing Tranche hashes. This abstraction of encodings is purposely done so that user data can later be re-compressed, re-encrypted, or re-encoded if a more appropriate algorithm is desired. This even allows for encodings that are not yet invented.

How Tranche Quickly Finds Data Shared On The Network

Use of Tranche hash spans does more than provide an elegant mechanism for evenly distributing the contents of files. The hash spans also provide a convenient mechanism for quickly determining what servers should have particular chunks of data. Periodically the Tranche tools query servers on the Tranche network to obtain a list of what hash spans are configured for all servers. This list is typically very small, a hash span or two per server, and generally needs to be downloaded only once and cached for reuse. Based on the list of hash spans, the Tranche tools can very quickly look up the location of the data on the network. Instead of having to ask all servers if they have a particular chunk of data, the Tranche tools can simply scan the cached hash spans and ask exactly which servers should have a particular chunk of data. Those severs can then be queried appropriately to access those data. Likewise, when uploading data, the same cached hash spans can be used to very quickly determine exactly what servers should get a copy of particular data chunks.

We refer to this technique as an “index-less” approach because there is no requirement for a single server to maintain an index of where all of the files in Tranche currently are. For example, consider how Google works when looking up data. Google periodically indexes as many websites as possible. When a user wants to find a particular web page, e.g. a proteomics website, the user would enter a query on Google's homepage. The magic behind Google's search then takes place where hundreds of databases are likely queried to find the most relevant websites and a list of those websites is presented to the user. Google is required in this process because web servers are allowed to store any file. There is no way to know where a file is without consulting a third party index. In Tranche, there is no need for a tool such as Google when trying to download or upload data. Instead each server has a defined range of files that it hosts. When the Tranche code wants to download or upload it can look at the list of pre-defined hash spans and know immediately where to upload or download data. Figure 5-2 enforces this concept with a figure that uses the English alphabet as model for Tranche hashes.

Note, a tool like Google can be generally helpful on a Tranche network for finding semantic information and this is why ProteomeCommons.org provides the meta-data indexing service, which is conveniently also indexed by Google. While Tranche can quickly download and upload data appropriately, the hash span mechanism doesn't allow for queries such as “show me all data from the organism Mouse” or “show me all data from ThermoFinnigan Scientific LTQs”. A third party tool can enable such queries by scanning all of the data in Tranche and examining the contents for meaning. Tranche also has a general mechanism for associated meta-data with files on the network. This allows arbitrary standardized annotations, e.g. ProteomeCommons.org's format or mzXML or MIAPE, to be directly linked to files hosted in Tranche. The benefit being that third-party index tools can systematically see these links and choose to read the associated, standardized annotation if the format is understood. Furthermore, data standards in proteomics (and general) are in a state of flux. By linking arbitrary annotations to files, Tranche provides a mechanism for accommodating any new data annotation scheme.

How Tranche Verifies Data

It is the author's opinion that Tranche provides an invaluable mechanism for sharing raw and processed data files. Aside from handling large data sets for publishing scientific data, most researcher's simply don't know that they can ask for verification that data hasn't changed since publication. Tranche hashes assure with an extremely high level of certainty that cited data has not changed since publication. Furthermore, this is done using standard digital hashing algorithms that aren't specifically tied to Tranche. This means that regardless of whether Tranche is used as a repository for a specific data set, a Tranche hash can be recalculated easily by anyone and be used to prove that data hasn't changed since publication. Stated in a different way, if data is initially published in Tranche and cited using a Tranche hash, it doesn't matter if the data is migrated later to a different repository. The data can still be checked to ensure that it hasn't changed since publication with or without Tranche.

A Tranche hash is nothing more than the physical bytes of output from three independent hashing algorithms concatenated with an 8 byte representation of an unsigned long (8 bytes or $2^{\hat{}64}$) appended to the end. Figure 5-3 provides an example. The three specific algorithms used are MD5 [102], SHA-1 and SHA-256 [103]. The unsigned long is in little endian format. It is worth noting that both MD5 and SHA-1 are currently considered theoretically "broken", loosely meaning that relying solely on them is no longer appropriate for digital hashing. However, practically, MD5 and SHA-1 exploits are largely tied to content bloating where the hash can be replicated by appending more, sometimes ridiculous amounts, of data to the end of a file. Inclusion of the file length in the Tranche hash prevents such abuse. Additionally, regardless of MD5 or SHA-1 potential faults, the Tranche hash includes the SHA-256 algorithm which is currently considered safe and one of the best digital hashing algorithms available.

Finally, one minor point needs clarification. The aforementioned hash algorithms work on one file at a time. This poses a problem if one desires to upload multiple files as a data set, which is commonly expected with proteomics data. The solution Tranche provides is that if multiple files are upload the following is done. First, each individual file is uploaded and the hash is recorded. Second, all of the recorded hashes are saved to a file that in Tranche terms is named a “project file” (top right in Figure 5-1). Third and finally, the newly made project file itself is uploaded to Tranche. In order to cite or download the entire data set one need only cite the project file. The exact algorithms previously described can be used to validate that the file's contents haven't changed, and thus the list of hashes in the file are inferred to be legitimate. The files represented by that list of hashes can then be downloaded and validated automatically to rebuild the entire data set.

Why Tranche hashes are well suited for publication and why they look like long strings of gibberish

Tranche hashes are almost always encoded in either Base64 or Base16. This is because the raw bytes that represent a Tranche hash are not suitable for publication in a peer-reviewed manuscript. Often the publication's character encoding does not allow all byte values to be shown, but more practically, it is of questionable benefit to present readers with characters they have never seen before or cannot easily type on a standard keyboard. The solution is to convert the raw Tranche hash bytes into a set of English-friendly characters, which is exactly what Base64 and Base16 provides. Figure 5-3 provides an example of such a Base64 representation.

Another reason Tranche hashes are presented in Base64 or Base16 is that these two encodings specify exactly what characters can be used to represent the underlying data, Base 16 allows a-f and 0-9 (16 different characters) and Base64 allows a-z, 0-9, and a several more (64 different characters in all). Both encodings do not allow the hyphen “–“ or blank space “ “ or the greater-than symbol “>”, which are by far the most commonly inserted characters in formal publications or e-mails. For example, if a line of text is too long, normally a formal publication hyphenates it and makes two lines. Similarly, an e-mail will often break up long text and insert “>” to symbolize a reply to a previous message. Both of these cases are logical ways that a Tranche hash will get munged via publication; however, both cases are easily prevented if the Tranche hash was originally in either Base16 or Base64 format. The fix is as simple as throwing away known bad characters (i.e. “–“, “ “, “>”) and reconstructing the valid Tranche hash.

Finally, it is worth noting that URLs, the most commonly used mechanism for currently citing data, have no such ability to be reconstructed if damaged by e-mail or publication formatting. Additionally, URLs have the critical flaw that they normally provide no mechanism for formally checking that data has not changed since publication. The data a URL points to can easily be changed by whoever owns the corresponding web server. Considering these two faults it can be inferred that URLs are not appropriate for publication of peer-reviewed scientific data compared to a Tranche hash or similar solution.

How Tranche prevents abuse of computers on the network

Tranche provides a unique best effort system for preventing abuse of the network. The term “best effort” is used because abuse prevention is an ongoing process and should not be inferred to mean that Tranche can guarantee that illegal MP3s will never be published on publicly funded servers. Further explanation is required. What Tranche provides is support for e-commerce grade digital signatures of data. More plainly put, Tranche uses the same mechanism used by the on-line banking industry to encrypt web page traffic that contains sensitive information. This is possible because the related algorithms are public standards. Thus data that is hosted on Tranche is allowed on-line and archived if and only if it comes from a trusted source. Furthermore, every bit of data in Tranche is digitally signed by at least one person. This allows data to be revoked if it is ever discovered that an individual is abusing Tranche resources.

The entire digital signature support is based on the X.509 standard [104] and public key cryptography [105], which means Tranche users must have a X.509 key in order to add data, delete data, or do anything other than download data. That is how trust is established when data is being uploaded and it works because only the owner(s) of a particular Tranche network can create new upload keys. In the case of the ProteomeCommons.org Tranche network, only the group collaborating at ProteomeCommons.org can make new X.509 keys that will work on all servers. Occasionally, customized tools are provided that make it appear like no upload key is required, but the tool itself is simply hiding use of an appropriate key. Finally, these X.509 keys have a limited lifespan, which means that it is trivial to let a user upload or modify data for a fixed period of time, say pre-publication, and ensure that the user will lose such privileges later, say post-publication.

What cannot be provided by Tranche or the X.509 standard (or on-line e-commerce web sites for that matter) is that illegal or illicit information is blocked from signing by a valid X.509 key. That brings discussion back to the MP3 example. A legitimate Tranche user can sign an illicit MP3 and upload it to Tranche. Certainly that file can later be revoked at will and the owner of the key abused is unambiguously known, but that is the best that can be done at this time. There is no method available to currently scan all files and determine if they contain illegal content or not. The file format itself cannot simply be excluded. MP3s for example are commonly used to publish podcasts, including the Tranche podcast that gets published in Tranche. There is no method of automatically inspecting all MP3s (or similar potentially illicit files), interpreting what the data represents, and finally deciding if the information is illegal or not. Some file formats attempt to allow for this type of functionality, but the vast majority of scientific data file formats, specifically proteomics file formats, do not have such functionality. Thus it is impossible to strictly prevent a trusted user from uploading an illegal file, although Tranche makes allowances for easily dealing with this situation once it is identified.

Tranche provides a best effort solution to preventing abuse. In general, the community behind Tranche has little interest in sharing inappropriate data; however, Tranche still provides what we argue is the most practical and robust mechanism for preventing the sharing of illicit data. Most popularized P2P programs do not have similar mechanisms (often purposely so), which is why Tranche is particularly well-suited for sharing scientific data. Resources supported by public funds, particularly government grants, should be as difficult as possible to abuse. Tranche does provide a practical approach of ensuring this.

Conclusion and Further Information

The Tranche system acts as a data repository, allowing very large data sets to be efficiently shared by partitioning (slicing) responsibility for sharing the data across several computers. Should one large computer exist that can handle all data, it can easily be used with Tranche to do just that; however, currently, and of more importance, is Tranche's ability to aggregate multiple smaller computers to accomplish the task of storing and archiving vast quantities of data – more than any one group may wish to be responsible for. Tranche accomplishes this goal through use of several simple tactics described in this manuscript. Tranche is Free Open Source Software (FOSS) and the implementations of all of these concepts are freely available from http://tranche.proteomecommons.org. Additionally, Tranche was developed using agile software development philosophies, meaning all features are strictly checked via additional code named “unit tests”. These unit tests themselves are all freely available with the Tranche source-code and provide an ideal mechanism for proving important concepts are implemented as expected and for providing code snippets that others can learn from.

The source-code itself is by far the most appropriate documentation of how things work, but interested individuals are encouraged to ask questions and participate in the Tranche e-mail lists and bi-monthly remote Tranche user meetings. Tranche is intended to be an open, free tool for benefiting the field of proteomics and science in general. The intention is that others will reap the rewards of this work, ideally without having to learn the details of how Tranche works, and that at least one complete solution to the data-sharing and publication problem in proteomics will be available for future use, discussion, and improvement.

Chapter 6

Bonanza

Unidentified tandem mass spectra typically represent 50% to 90% of the spectra acquired in proteomics studies. These idiopathic spectra may fail to yield results for several reasons, including low signal to noise ratios, incomplete fragmentation, differences in the chemical structures of the peptides, co-fragmentation, among others. The class of unidentified spectra representing chemically modified peptides are of particular biological interest. This manuscript describes a novel algorithm, “Bonanza”, for clustering spectra without knowledge of peptide or protein identifications. It also represents a new approach that specifically matches MS/MS spectra independently of precursor mass to identify both identical spectra and spectra that are otherwise identical but have an m/z shift for the precursor ion (i.e. potential modification or amino acid substitution). Furthermore, the presented algorithm works independently of a spectral library, allowing for additional identifications to be mined from existing datasets and approximating trends in both chemical and biological peptide modifications. Also described here is a probability-based scoring method and a high efficiency search process. Application of Bonanza to a collection of MALDI TOFTOF tandem mass spectra obtained from approximately 250 recombinant human proteins expressed in E. Coli identified biological modifications as well as chemical artifacts. A

similar global analysis performed on isotopically tagged human embryonic stem cell extracts also identified trends in biological modifications and chemical artifacts. The approach described here significantly increases the number of spectra identified, improves identification of post-translational modifications or amino acid substitutions, provides a global quality assessment, and could be used to filter spectra prior to database searches to reduce computational times.

Introduction

Database searches of peak lists from tandem mass spectrometry datasets rarely results in unambiguous identification of more than half of the collected spectra using current extant search engines. The current focus is on the fraction of spectra that can confidently be identified above a scoring or probability threshold which leaves many spectra unaccounted for. The extremely large number of MS/MS spectra acquired in a typical experiment makes it impractical to individually account for the unidentified spectra by expert *de novo* analysis. These idiopathic spectra include peptides that fragment poorly or have low signal-to-noise levels, but also include unexpected post-translational modifications, amino acid substitutions, splice sites, artifactual modifications, and the co-fragmentation events that occur in analysis of high complexity samples, particularly with tandem TOF instruments. All of these latter categories can be of vital biological or analytical importance and represent an opportunity for new approaches to intelligently recognize previously unidentified spectra that could boost the total number of peptides, proteins, and post-translational modifications identified.

Database search algorithms such as X!Tandem [106], Mascot [107], Sequest [108], and similar algorithms [109-116] identify a significant portion of spectra generated by a MudPIT style proteomics experiment. These search tools generally work by matching the observed fragment masses for particular parent ions to the theoretical fragment masses calculated from an organism-specific proteome database. It is not uncommon to have the majority of the spectra left unidentified. Explaining why this subset of spectra remains unidentified, and, more importantly, creating tools that assist in their identification, benefits the proteomics community as a whole by exposing additional information that search algorithms may have difficulty finding. Because existing search algorithms match experimental MS/MS spectra against theoretical MS/MS spectra for peptides having the same parent mass, they can miss peptides whose primary structures are discrepant due to chemical modifications, amino acid substitutions, or other reasons. Most search engines attempt to address this problem by allowing post-translational modifications and substitutions to be specified. When variable modifications are specified, the search space rises exponentially, with increases in search time and false positive rates which has led to development of a number of approaches that attempt to minimize this effect, including iterative approaches that generate a smaller search library that is subsequently evaluated for modifications. For these reasons, it is generally advisable to include a limited number of variable modifications during database searches.

Spectral matching tools, as described in this manuscript, can also be useful for extending classical database searches by identifying unexpected post-translational modifications. Recent projects by Stein et. al (http://chemdata.nist.gov/mass-spc/ftp/mass-spc/PepLib.pdf), Craig et. al [117], Frewen et al. [118], Bandeira et al. [119] and Lam et al. [120] illustrate the benefits of examining MudPIT data through use of spectral matching algorithms. The approach used in those studies is an orthogonal method to thos described above. In many of the above studies (Stein, Craig, Frewen, and Lam) a library of existing, identified spectra is condensed into consensus or representative peak lists and a subset of the library, often spectra with similar precursor *m/z* values, is then compared to an unknown peak list in an effort to identify matches to the existing library. Unlike database search algorithms, no organism-specific proteome database (e.g. FASTA file) is required and users need not specify particular peptide modifications. If a modified peptide spectrum is present in the spectral library, it can be matched against the equivalent spectrum from another experiment. Additionally, since these libraries hold data collected from laboratory experiments, they will exhibit more accurate fragmentation patterns and ion intensity values than an *in silico* peptide fragmentation. The improved quality of peptide identifications using spectral matching was documented in a preliminary study indicating greatly improved ROC profiles for spectral matching over classical search engines [120]. Other research such as that by Bandeira et al. begins to explore the concept that related spectra, specifically modified and unmodified forms of a peptide, often generate similar MS/MS spectra. Through use of spectral graphs and looking for long, shared sequences of amino acids, MS/MS spectra can be compared for similarity. In cases of high similarity the spectra can further be examined for potential modifications.

The work described in this manuscript is most similar to recently described spectral search algorithms cited above, particularly the Bandeira, *et al.* work, but differs from previous efforts by addressing the large number of unidentified spectra and the relationship of these unidentified spectra to spectra that are easily identified by classical search engines. Presented here is the design and implementation of a novel approach that clusters spectra regardless of precursor *m/z*, amino acid residue modifications, or whether the peptide sequence can be identified by a database search algorithm. Like the Banderia work the intention is to identify potential modifications through comparison of spectra, independent of a protein database; however, the approach described by this manuscript does not use spectral graphs. Instead, result files from existing MS/MS database search engines are post-processed to infer high-confidence identifications. Subsequently, these identifications are clustered to other spectra in order to infer similarity. Once similar spectra are identified tentative peptide identifications are assigned along with potential modifications that can account for differences in the spectra.

Our approach to spectral searching extends existing spectral searching in at least the following two unique aspects. First, no restriction is placed on the precursor mass when comparing peak lists. MS/MS spectra with different precursor masses may be clustered and identified. Second, the peak list comparison is based on both the observed fragment ions and on the precursor m/z subtracted by the observed peaks, named the “parent-minus peak list”. The entire Bonanza algorithm is further documented in the Methods section along with a description of performance characteristics in the Results section.

Evaluation of the Bonanza algorithm is performed on two different datasets acquired on Applied Biosystems/Sciex model 4700 and 4800 MALDI TOF/TOF mass spectrometers. The first data set is the Aurum reference data set by Falkner et al. [121]. This data set represents a controlled set of approximately 250 human proteins that have been expressed in E.Coli. The peak lists are well documented based on decoy database analysis and provide a good trued test data set. The second data set is a human embryonic stem cell data set collected by A. Yocum in collaboration with the laboratory of Dr. Katherine O'Shea at the University of Michigan. This data set represents a more realistic set of peak lists obtained in the process of a biological experiment.

Experimental Procedures

Pre-processing of Peak Lists – The Bonanza algorithm can be applied to any peak list, but preprocessing was applied to the data sets analyzed in this manuscript in an attempt to reduce low intensity noise peaks and improve memory requirements and processing time requirements of the algorithm. Preprocessing was applied to all peak lists to take at most the top two most intense peaks per every 100 Da m/z. Of this list only the top 30 peaks were kept per peak list. Peak lists with less than 10 peaks were discarded from the analysis.

Peak List Comparison Score – A modified dot product is used for comparing MS/MS peak lists. Four key modifications exist. First, peaks are matched within an arbitrary cutoff, Δ, by default 0.3 Da for MALDI TOF/TOF data. Second, the intensities of each peak in each peak list is converted to the portion of total intensity in the particular peak list. Third, peaks are partitioned into two groups: matched and unmatched. The dot product of the matched peaks is calculated without change where $\vec{a}$ and $\vec{b}$ are the peak lists being compared and $\bar{a}$ and $\bar{b}$ are respective pairs of matching peaks within the Δ cutoff.

$$m(\vec{a},\vec{b},\Delta)=\sum \bar{a}_i\bar{b}_i=\bar{a}_1\cdot\bar{b}_1+\bar{a}_2\cdot\bar{b}_2+\ldots+\bar{a}_n\cdot\bar{b}_n$$

The dot product of the unmatched peaks is improvised to be the squared intensity of the unmatched peak, either a or b respectively. This practice allows for unmatched peaks to penalize the final calculated bonanza score.

$$u(\vec{a},\vec{b},\Delta)=\sum a_i^2+\sum b_i^2=(a_1^2+a_2^2+\ldots+a_n^2)+(b_1^2+b_2^2+\ldots+b_n^2)$$

The final score, named bonanza score, is the matched dot product divided by the matched dot product plus the unmatched (modified) dot product.

$$bonanza\,score=\frac{m(\vec{a},\vec{b},\Delta)}{m(\vec{a},\vec{b},\Delta)+u(\vec{a},\vec{b},\Delta)}$$

The fourth and final modification is of particular importance, and it is how the bonanza algorithm is capable of finding unexpected peptide modifications, typically single residue side-chain modifications. When determining if two peaks from different peak lists match or not, e.g. for inclusion in the matched partition, two checks are performed. First, a check is performed to see if the *m/z* ratio reported for the two peaks is less than or equal to the Δ cutoff. Second, a check is performed to see if the respective precursor mass minus the m/z ratio of two peaks is within the Δ cutoff. If either of the two checks is satisfied for a pair of compared peaks, then that pairing is added to the set of matched peaks. Otherwise the remaining peaks are considered unmatched.

Confidence Calculation for Valid Clusters – Clusterings presented in this manuscript are preformed by comparing all peak lists against all other peak lists for each data set, regardless of if the *m/z* of precursor ions is similar or not. For each peak list, the set of scores of it compared against other peak lists is sorted from highest to lowest. The sorted scores are then used to approximate the confidence of valid clusterings as follows. It is assumed that the highest score will represent the most similar other peak list and subsequent scores will represent less similar peak lists up to the lowest bonanza score. Based on this assumption, the distribution of the 1st best bonanza scores should include a mix of both valid clusterings and invalid clusterings. Likewise, the distributions of the next best bonanza scores, e.g. 10th, 20th, 50th, and 100th, for each peak list will also be a mixture of invalid and valid clusterings; however, the worse the ranking of the bonanza score, e.g. 100th best score, the more prominent the distribution of invalid clusterings should be. This idea is predicated on prior experience with mass spectrometer data sets and the observation that rarely will the same peak list be collected more than a few times. This observation is particularly true with the common practice of using a dynamic MS/MS acquisition algorithm that purposely tries to avoid repetitive collection of the same MS/MS scan; however, it is worth noting that occasionally the same peak list will be acquired numerous times (10 or more) but very rarely 50 times or more even on a large-scale experiment.

In order to approximate an arbitrary rate of incorrect versus correct clustering, the 1st best clustering score distribution is compared against a aggregate of the 10th, 20th, 50th, and 100th best clustering score distributions. The aggregate is calculated by using the mean of the distributions plus two standard deviations, as illustrated in Figure 6-1. Given any cluster score threshold an approximate rate of estimated invalid clusterings versus estimated valid clusterings can be made. For this manuscript, results are report for an estimated 95% valid versus invalid clusterings. Conveniently, this objective measure of confidence will automatically account for parameterized changes in the delta *m/z* used when assigning matched and unmatched peaks in peak list comparisons.

Estimating modification trends – Bonanza clustered peak lists do not necessarily need peptide identifications to provide information about the data set. The clusters alone can be used to provide an estimate of the peak lists that are being observed multiple times, which implies reproducible artifacts in the analysis. Additionally, the clusters can serve as an approximation of trends in the data, e.g. common modifications on peptides. Figure 6-2 summarizes this second point. Provided are plots of the m/z differences between peak list clusterings with bonanza scores above the approximated 95% confidence ratio. Without knowing anything about the data it is clear that some *m/z* differences appear much more often than others. Not surprisingly, for the Aurum data set, nominal *m/z* differences of +/-16, +/-17, +/-18, and +/-32 Da dominate the clusters. Using the HUPO-MS terminology (used throughout the manuscript), these changes could easily be argued to be oxidation of peptides and Glu->pyro-Glu and Gln->pyro-Glu of N-terminal residues. Analysis of the MS/MS data and associated peptide identifications is later provided along with similar analysis describing dethiomethyl as the source of the clusters of nominal mass differences at +/-42, and +/-64 Da. However, juxtaposition of the Yocum data set, an iTRAQ experiment, with the Aurum set provides significant support for the idea that obvious trends in Figure 2 may be taken at face value. Dominating the Yocum cluster trends are nominal *m/z* differences at +/-144 Da. These *m/z* differences correspond well to the known iTRAQ4plex modification (~144.1).

Subsequent information regarding MS/MS analysis and peptide identifications is provided later in the manuscript to further analyze these trends.

Peptide Identifications – The Bonanza algorithm is not intended as a search engine *per se* and thus does not have a component that performs MS/MS database searching that identifies peptides to peak lists. Instead Bonanza relies on other search engines to provide this functionality and the Bonanza algorithm is restricted to inferring peptide identifications based on clusters of peak lists where at least one of the peak lists is identified. This design feature allows Bonanza to work with existing MS/MS identification software as a tool to help account for more of the observed peak lists. Furthermore, Bonanza analysis can also be performed on existing bioinformatics analysis, as demonstrated with the Aurum data set, which is convenient if post processing a large set of previously analyzed data. Bonanza also lessens the requirement to explicitly specify potential modifications when performing an MS/MS database search. If both modified and unmodified forms of a peptide are acquired in the MS/MS analysis then only the unmodified form needs to be identified by the search engine. Bonanza can cluster the modified form with the unmodified and help infer the appropriate identification. This practice is appealing because most MS/MS search engines degrade significantly in performance, both in the accuracy of peptide identifications and the speed of the searches when multiple partial modifications are specified.

Results from three and four MS/MS search engines were incorporated into the analysis of the Aurum and Yocum data sets, respectively, i.e., Mascot [121], X!Tandem [122], X!Tandem with the pluggable k_score algorithm [123], and Sequest [124]. In the case of the Aurum data set, the pluggable k_score algorithm was not used because this manuscript is reanalyzing the original search results, which do not include k_score. All searches were performed with similar parameters (see supplemental data for details) and on the same FASTA file. A decoy database search strategy was used as previously detailed by Falkner et. al [125]. In short, the strategy combined the August 2006 Human IPI database with a reversed version of of the same protein sequences. Each reversed entry is noted by including “R” in the protein's accession number. Searches were performed normally by each of the software packages and then filtered to keep only matches above a 95% confidence threshold. The 95% confidence threshold was determined by ranking the respective search engine results by score and counting the number of known decoy matches present. The threshold was used where a ratio of 190 (95 * 2) peptide identifications exists for every 5 known decoy peptide. The resulting lists of peptide identifications are the ones used in the analysis presented by this manuscript. The complete search results are included in the supplemental information included with this manuscript. These search results do contain all search parameters used.

It is important to comment that the decoy strategy we used in this manuscript is not an adequate method for comparing the individual search engines nor do we suggest the approach as a superior peptide identification method. No attempt was made to optimize individual performance of the search engines, nor were any enhanced search features used to help find unusual potential modifications, point mutations, or the like. The aggregate set of identifications is only intended to represent a reasonable base analysis of the data, something that also represents normal practice for initial searches of similar data sets.

Inferred peptide Identifications – Inference of peptide identifications was provided to demonstrate that the vast majority of Bonanza clustered peak lists to represent a logical modification of a amino acid side chain. In the simple case where only one unique peptide identification exists, all other peak lists in the cluster are assumed to have the same amino acid sequence. If two peak lists do not have the same precursor *m/z* then the difference between the two peak lists is applied as a potential modification that might have occurred to any of the residues in the peptide. The “best” match was found by summing the intensities of the peaks that match the theoretical b- and y-ion series for that peptide. The highest aggregate intensity match is considered the best match. In cases where multiple candidate peptide identifications were present in the same clusters, all were considered when determining the best match. The unchanged, original peak lists were used in this intensity comparison. Not the filtered peak lists as described previously.

The resulting identifications aggregated from the individual decoy analysis and inferred by Bonanza are provided as comma-delimited files in the supplemental data. The results also include statistics for each match that allow for manual examination of the inferred peptide sequences. The data tables may easily be opened, viewed and column-sorted by either Microsoft Excel or the free OpenOffice.org software for simplified manual inspection.

Description of datasets – The Aurum data set is a published reference data set by Falkner et al. [125]. Approximately 246 human proteins were expressed in E.coli, purified, checked for purity by SDS PAGE, the gel bands were individually digested by trypsin, and analyzed individually via a ABI 4700 MALDI TOF/TOF. The data set includes a similar decoy database analysis as described under Methods. The same decoy database analysis is used by this study.

A human embryonic stem cell data set collected by Anastasia Yocum in collaboration with Dr. Kathy O'Shea was used as an experimental data set to test the performance of the Bonanza algorithm. The data set consisted of three biological replicates that were each analyzed in triplicate for a total of nine 2D LC runs with each sample being spotted across two MALDI target plates. A total of 18 plates were used in MALDI-MS/MS analysis via an ABI 4800 MALDI TOF/TOF. The data set is of particular interest, because it is expected to contain many replicates of the same spectra and derived peak lists. Like the Aurum data set, the Yocum data set was digested with trypsin prior to spotting on the MALDI target plate; however, unlike the Aurum data set the Yocum data set was blocked with MMTS at cysteine residues, labeled with the iTRAQ reagent, and the source protein samples were not checked for purity via 1D PAGE. Further description of the Yocum hESC data set is included in a manuscript pending publication. Details regarding the Yocum hESC manuscript may be obtained from the communicating author.

Performance Characteristics – The Aurum data set (10,000 peak lists) was analyzed by Bonanza in approximately 5 minutes, and the Yocum data set (42,000 peak lists) required approximately 1 hour and 8 minutes. Data analysis was performed on a 2.0 Ghz Pentium M computer with 1GB of RAM. The source-code demonstrating the Bonanza algorithm is not particularly optimized for speed; however, it is designed to take similar, if not much less, time than the related MS/MS analysis. The time requirements for the two data sets presented in this manuscript demonstrate that it is feasible to run the Bonanza algorithm in-line with an existing proteomics data analysis pipeline.

Availability

Source-code and documentation for this project as well as the Aurum dataset are made freely available under the Apache 2.0 license. Copies of these files may be requested from the authors. The data sets and result files for this manuscript are made available through Tranche (http://tranche.proteomecommons.org).

The files related to the Aurum data set can be downloaded using the following Tranche hash.

96rx5lCBh6SNpGyuAsE1fSEn3sDxwmHITFfC9uQMNob12r36Xqg2+uFHJ46Jd
VrZB2/UwdbWvizBfigbzJMtpxV9/AQAAAAAAAAFCg==

The hESC data set (Yocum data set) can be found in Tranche using the following hash.

ClX0eNVtoXZrFA6oixm6tImsBvGtrJi7bZCwwJjohqBGaGZDruH0KkntDx9Mw
CXSDRfNLuajYHTtp90/2WYivOjhCxQAAAAAAAAALew==

The hESC data set is not included with the public Bonanza data and source-code because public release of the data is pending final acceptance of Dr. Yocum's hESC manuscript. The Tranche hash provided here references the encrypted project and the files will be released for public access when the stem cell manuscript is in press.

Results and Discussion

Several conclusions regarding use of spectral clustering to aid analysis of MS/MS data can be drawn from our analysis. The first observation is that Bonanza allows significantly more spectra to be identified (484 in the Aurum data set and 3,650 in the Yocum data set) than high-confidence decoy analysis alone. Expanding upon the plots presented in Figure 6-2, it is clear that many of these newly identified spectra are modified forms of identified peptides. While identifying these spectra may not contribute to more unique peptide identifications, Bonanza does provide a valuable quality control mechanism through summation of trends present in clusters of peak lists. For example, techniques such as iTRAQ or even reducing and blocking disulfide bonds can be globally evaluated for completion – a complete reaction should primarily yield a single form of the peptide, leaving minimal observable other forms via Bonanza analysis. In addition, sample quality can be globally evaluated for extreme oxidation, Glu->pyro-Glu, Gln->pyro-Glu and similar, commonly observed modifications. It is important to point out that Bonanza's ability to provide this global view of data sets is in addition to any use of Bonanza to do same-dataset or cross-dataset spectral searches. Identification of significantly more spectra can also contribute significantly to the quality of quantitative studies by providing considerably valid identifications. In the same manner, it could also be used to contribute to peptide scores.

Table 6-1 summarizes the Bonanza analysis including peak lists kept after filtering, unidentified clustered peak lists, identified peak lists and inferred peptide identifications. This analysis indicates that Bonanza is finding matches that were not identified by search engines alone, and the included listings of identifications in the supplemental material support that these identifications are of similar quality to that of the search engines. However, the caveat is that Bonanza is not actually finding any new peptide identifications compared to the decoy analysis. The inferred identifications are either modified forms of the same peptide or similar spectra that were not identified using search engines alone.

Although Bonanza makes no new peptide identifications, it provides significant insight into both of the data sets described in this manuscript. It does so by allowing a larger fraction of the spectra to be accounted for, leading to identification of experimental artifacts and potential post-translational modifications. For example, initial Bonanza analysis of the Aurum data yielded the global view in Figure 6-2 a where the differences in mass within clusters are binned in one dalton increments. The data shown has major peaks at +/-14, +/-16, +/-17, +/-18, +/-32, +/-46, +/-64, and +/- 128 Da. These major peaks represent abundant modifications to peptides in this preparation. Note that the mass discrepancies in Figure 6-2 occur in pairs equidistant from the origin because the modified and unmodified versions of these peptides are not distinguished. Actual modification trends and corresponding amino acids can be identified by looking at the trends of the decoy analysis and inferred peptide identifications. Table 6-2 provides a summary of the top residue side-chain modifications for both data sets of interest. Clear trends at methionine oxidation (+16) not only on methionine residues but also tryptophan and histidine residues. Additionally, dioxidation (+32) modifications are also present in lower abundances. The potentially mysterious +/-14 Da trend is also readily explained by the abundance of propionamide modifications, which would have prevented the expected carbamidomethyl modification (71 – 57 = 14). The Aurum data set was purified using PAGE.

Figure 6-3 summarizes the reasoning behind correctly interpreting clusters with -46 Da precursor differences. Based on Lagerwed et. al's [126] analysis. Neutral loss of -64, or -80 is routinely observed for the unmodified, singly and doubly oxidized forms of methionine respectively. Further examination of the MS/MS peak lists, Figure 6-4b provides an example that contains both -46 and -64 losses indicating a 2 Da shift is required to best match the observed b- and y-ion series. Closer examination of the MS spectrum suggests that the neutral loss must be due to incomplete metastable decay prior to MS/MS fragmentation, and the intermediate form observed in MS mode incorrectly appears 2 Da higher than the neutral loss with only a minor peak present at the appropriate neutral loss mass. Bonanza analysis uncovered this situation, and interestingly, it represents a modification that the MS/MS search engines used in this study can not be told to look for. No known mechanism exists in these algorithms to specify that a potential modification should be considered that appears to be 2 Da lighter in the MS scan.

However, a small fraction of the spectra mapping to the 14 Da adduct peak do not contain cysteine but do contain many high quality b/y ion series that confirm methylation of E and D residues. The source of the methylation could be endogenous methyltransferase activity in *E. Coli* which has been reported for expression of recombinant proteins in *E. coli* [127,128]. It more likely arises from the colloidal Coomassie Blue gel staining procedure which is performed in acidic methanol. No clear evidence for simple amino acid substitutions were observed for this data set. The +/-14 Da trend alone yielded two significant insights into this dataset. While most search engines allow inclusion of propionamide as a variable modification, it is not always chosen. Additionally, variable methyl esterification of Asp and Glu residues is also provided for, but is not expected and rarely selected to keep search times lower and minimize false positives. It is also worth noting that the mass shift for a Glu for Asp substitution is the same as for methyl esterification of an Asp residue.

More, obvious expected trends in the Aurum dataset including +/-17 and +/-18 are explained by cyclization of N-terminal Glu and Gln residues respectively. Other minor, but significant trends were also identified at +/-128 and -/+156 Da and explained by missed tryptic cleavages at lysyl and arginyl residues respectively. These minor trends were validated by MS/MS analysis and by inspection of the original protein sequence FASTA file for dibasic sites. Almost all of these were semitryptic peptides or were due to incomplete cleavage at dibasic sites (KK, KR, RK, RR). Trypsin can cleave between the two residues or on the C-terminal side of the dibasic site. If Trypsin cleaves the latter site, little additional cleavage occurs because trypsin is not an efficient exopeptidase.   These minor trends were also observed for the Yocum data set, but with lysyl residues being iTRAQ modified.

The global view of the Yocum data set provided in 6-2b has several different features than for the Aurum data set 6-2a. Interestingly, the significant methionine side-chain neutral losses are not present, nor is the significant nominal +/-14 Da peak present. The experimental protocol supports both of these observations as the proteins were not expressed in E.Coli and the sample was not subjected to polyacrylamide gel electrophoresis (PAGE). However, the Yocum data set is dominated by nominal peaks at +/-144.1 and +/-272 Da. The +/-272 Da trend is easily identified as iTRAQ modified lysine residues – similar to the +/-128 Da trend in the Aurum data set. The +/-144.1 trend corresponds to the mass of the iTRAQ4plex modification. Analysis of the MS/MS data determined that incomplete

N-terminal modification by iTRAQ explains the majority of the trend, but variable modification of lysine, tyrosine, threonine and serine residues (K, Y, T, and S) are also clearly present. Variable modification of tyrosyl, threonyl, and seryl side chains has been previously reported as known side reactions of the iTRAQ reagent and are included in ABI's MS/MS search engine software Paragon [129]. The extent of incomplete N-term modification by iTRAQ was anomalous and is attributed to the slightly lower pKa value of N-terminal amino groups relative to the epsilon amino group of lysyl residues. Slight changes in pH values of reaction buffers can have significant effects on the relative degrees of nucleophilicity for these two classes of amino groups. This result could also arise from limiting concentrations of the iTRAQ acylating reagent. Variable N-term modification by iTRAQ does not seem to affect the quantification in this case, but is a cause for general concern since it can reduce the number of peptides identified if the search engine is not informed to treat N-terminal iTRAQ tags as variable and also because slight changes in labeling conditions could potentially lead to significant changes in reactivity.

It is interesting to observe that Bonanza analysis did not account for every peak list collected (Table 6-1). Excluding the singleton spectra that are not assigned to clusters, approximately 42% (2,274) of the clusters in Aurum and 58% (14,023) of the clusters in the Yocum data set are not identified. We propose that these while these peak lists represent reproducible fragmentation patterns, they may correspond to peptides that could not be identified due to features of the search engines or the search parameters selected. They could also represent contaminating compounds and peptides, repetitive electronic noise, or other analysis artifacts. It is unknown if all of these unidentified clusters represent peptide spectra that existing MS/MS search engines can be enhanced to identify or that the FASTA database employed in this study does not contain these genes; however, the clusters do represent a reasonable set of peak lists to examine for further novel peptide identifications.

Concluding Remarks

This manuscript describes the use of spectral clustering to effectively identify protein modifications through a non-targeted approach. When used to identify general trends in mass shifts, it can be an effective tool for quality control, identifying chemical artifacts and incomplete chemical reactions. It can also reliably identify post-translational modified peptides in a non-targeted way without significantly increasing the search time. This approach differs from targeted approaches that find only modifications that are specified. The quality control application of this approach can be used to help improve existing protocols for sample preparation – ideally leading to reductions in undesired side reactions or incomplete modifications.

Spectral clustering allows unanticipated modifications to be detected with good efficiency even when they are infrequent events. Application of an objective scoring threshold to spectral clustering provides an effective data-set specific method of determining correctly clustered peak lists. It is important to consider, however, that the more abundant a modification is, the more apparent it will be in the mass difference plots used to visualize trends in precursor mass shifts.

Beyond the applications highlighted in this manuscript, it is worth considering other potential applications of Bonanza-style spectral clustering. Many spectral clusters were observed in this study using the Bonanza software for which no member provided a significant peptide identification despite relative confidence in the validity of observed clusters. Often these clusters included common mass shifts. Such clusterings are intuitively a good place to attempt and improve a MS/MS search algorithms performance because other peak lists in the same data set have the same trends.

Another approach that could also be taken advantage of is to improve data acquisition by dynamic exclusion of precursor masses corresponding to the major m/z difference trends observed using this algorithm. Omitting known analytical artifact peaks can allow acquisition of more unique spectra. Alternatively, one could target known analytical modification peaks in order to increase the confidence of peptide identifications. These tactics could be particularly helpful for any data set being analyzed repetitively or that has been split into several fractions.

An important assumption that Bonanza makes is that the algorithm requires both a modified and unmodified form of a peptide in order to identify a modified peptide. This is a reasonable assumption for many post-translational modifications, particularly those used to modulate function. However, there is a significant fraction of post-translational modifications that are stoichiometric that would be missed in an intra-dataset search. For the analyses presented in this manuscript, clustering was performed within a single data set. When stoichiometric modifications are present, this problem can potentially be alleviated through clustering across multiple data sets or by use of a spectral library of proteins not post-translationally modified. Finally, unclustered spectra have some value because they will include most of the noise spectra which can be useful in diagnosing the properties of random spectral noise.

The Bonanza algorithm allows us to discover a large portion of the previously unidentified peak lists in the Aurum and Yocum data sets. These results indicate that this approach can be of benefit for analysis of MALDI tandem mass spectra. Application of this approach to electrospray ionization (ESI) MS/MS spectra is also practical and will be pursued in future studies. We found Bonanza's analysis particularly useful because while we did attempt to provide a robust, multi-search engine analysis of the initial data set that is limited to high-confidence identifications, the use of Bonanza allowed many more unidentified spectra to be confidently assigned. Bonanza unambiguously identified many analytical modifications of identified peptides without requiring the MS/MS search engines to explicitly search for them. The use of spectral clustering represents a considerable improvement in the identification of modifications because incorporating potential modifications into MS/MS search engines can result in lengthy analysis times and an increased number of false positives. The non-targeted nature of spectral clustering makes no assumptions about the presence of specific modifications and so will allow detection of unexpected modifications or even previously unknown modifications.

Bonanza successfully found many expected modifications, including oxidation and formation of N-terminal pyroglutamyl residues. Bonanza also found a number unanticipated but retrospectively likely modifications, including methyl esterification, certain variable modifications by iTRAQ, partial metastable decay during neutral loss of oxidized methionine side-chains, and polyacrylamide adducts. This implementation of spectral clustering successfully found these modifications in a non-targeted way. Additionally, it is clear that the non-targeted approach used by Bonanza can provide valuable quality control feedback regarding experimental protocols that existing targeted approaches used by MS/MS search engines are not designed to find. The application of spectral matching to accommodate the multiple charge states observed for ESI data is a logical extension to the algorithm. It is also worth noting that Bonanza is essentially a spectral comparison tool. From this perspective, the very efficient core algorithms could also be applied to cross data set analysis, acting as a spectral matching tool similar to X!Hunter, SpectraST, and the NIST peptide fragmentation library tool.

Acknowledgments

The authors wish to thank Alexey Nesvizhskii and Thomas Blackwell for helpful discussions regarding statistics and expectation values, Katherine O'Shea for collaboration in analysis of the hESC samples, and the many others in or associated with Philip Andrews lab for helpful discussions regarding the Bonanza algorithm and implementation.

This work was supported by NCRR Grant # P41-18627 to the National Resource for Proteomics and Pathways.

Chapter 7

Conclusion

The body of work presented in this thesis presents several novel advancements in the field of proteomics with a focus is on the development of MS/MS search engine algorithms and data sharing. By focusing on these topics, the overall field of proteomics was reduced to several tractable problems, which were addressed by this work. Clear themes of open-access, open-source, and building upon existing tools are present in each project. These themes are similar in importance to the scientific work itself because it makes the overall body of this thesis work more palatable for others, potentially even extend. Certainly the success of the Tranche project is largely due to the openness of both the source-code and the data that is shared by the tool.

The ProteomeCommons.org website continues to act as a community resource for general dissemination of proteomics information, including tools such as the JAF, IO Framework, PFSM, Aurum data set, and Tranche. The website itself is in a maintenance mode with regular updates relating to news, tools, and data sets. It does not appear as the web site's more developer orientated resources, namely the subversion repository, are tools that many individuals in the proteomics community desire; however, the news, tools, and data sets continue to drive more and more traffic to the website. It is fully expected that ProteomeCommons.org will remain operating in its current capacity for as long as the National Resource for Proteomics and Pathways (NRPP) exists. Maintenance of the resource is relatively easy compared to the amount of use the site gets and the benefits it has for the proteomics community. Overall, a successful web site and lasting resource for the community.

Out of all of the tools developed on ProteomeCommons.org the Tranche Project is perhaps the best example, and it certainly gets the majority of use. Prior to Tranche it was difficult to easily publish and associate a complete data set with a manuscript. Now, it is relatively easy to accomplish this task and resources such as ProteomeCommons.org even provide an index of all such resources. This represents a fundamental shift in the way that scientific data sets are shared. Data transfer is greatly accelerated. It is now possible for anyone to access the raw published data and it is relatively easy to ensure that the data persists indefinitely. A further, very important, point is that Tranche does not fundamentally change the concept of data sharing in science. Rather Tranche's role is to greatly accelerate the process of sharing data and resources. Previously the plumbing did not exist for easily sharing files. Yet, if sharing was desired groups would agree upon a collaboration and cumbersomely figure out a way to send the required information. Currently with Tranche, if data sharing is desired, groups still agree upon terms, then simply click upload or download for the appropriate data sets. Nothing is fundamentally different about the data itself or the negotiation of collaboration. This point is important because it is often missed. Primarily, I believe, due to the technophobia of researcher's that have long been entrenched in their computer-less discipline. This is likely the largest obstacle to overcome for widespread adoption of Tranche and the acceptance of a revolutionary approach to scientific information sharing. There is likely no easy way to bypass this obstacle; however, ease of use and open-source and open-access are all key features of Tranche that will facilitate its adoption. It is likely that the technology will always remain a mystery to many users, but the integrity of the tool's code and freedom of use are readily accessible concepts that all seem to easily consume.

The success of Tranche is not intended to dwarf the results of Bonanza. Work on the Bonanza algorithm shows great promise for leveraging the vast amount of data in Tranche and refining existing analyses to discover many previously unidentified peak lists and unexpected protein and peptide modifications. Given enough time it is likely that Bonanza style analysis will be both common place for MS/MS data sets and the technique will occupy much more mind share amongst scientists compared to Tranche. Several critical omissions existing in contemporary MS/MS database search algorithms. Most notably are those of neglect to do cross spectra comparisons and the inability to identify post translational modifications that were not previously expected. The initial Bonanza example work clearly shows that both of theses omissions can readily be resolved. Perhaps not completely, but to a very significant extend. Furthermore, obvious future work exists where Bonanza's generic scoring function is optimized to account for specific, fragmentation-altering post translational modifications such as phosphorylation and glycosolation. Continuing this work would be a fantastic opportunity for future researchers, and certainly the most logical continuation of this thesis work.

It is with great satisfaction that the software created during this thesis work is all available as free open-source software (FOSS). This type of tool development is possibly the best method of enabling both replication of previous work performed and complete public critique of the work – two key components of the scientific method. Conveniently, the FOSS model also enables anyone to freely access and use the software both in its compiled form and in source-code. This greatly aids in allowing other researchers to try and use the software. It also lets software developers openly critique the design of aspects of the code base and or or modify the code to fit specialized needs. Aside from the aforementioned benefits, one of the most satisfying aspects of the FOSS development for this thesis work is that it was directly supported and encouraged by the National Resource for Proteomics and Pathways (NRPP), including sponsors in part the National Center for Research and Resources (NCRR) and National Cancer Institute (NCI). Hopefully the work presented in this thesis, regardless of its scientific value, will serve as a good example of how FOSS development can both benefit scientific research and accelerate adoption of software tools.

A final, yet clearly present, theme in this thesis is that of working with the active areas of the proteomics community, not in competition with. It seems that far too many researchers are still attempting to create the next best MS or MS/MS search algorithm that will deprecate the reset. If not an obvious fault with this mentality, the Sequest algorithm developed in the early 1990's is still considered one of the best viable tools. Certainly room exists for competition and improvement but it is important to emphasize that the work presented in this thesis purposely avoided an obvious reinvention of the wheel – primarily due to foresight based on ProteomeCommons.org. The major software tools of this thesis work, Tranche and Bonanza, both work with existing search algorithms. Tranche enables facile access to data sets for both reproduction of prior analyses and better testing against a variety of data. Bonanza leverages the work that has been put in to existing MS/MS search algorithms and aids greatly in refining the parameters used by the tools so that easily identifiable data is not missed.

In conclusion, it seems most appropriate to comment on future uses and potential development of the tools described by this thesis, specifically Tranche and Bonanza. Tranche has matured into a relatively stable and widely used tool. The NRPP has funding to continue support of Tranche for several years to come, and it is well within reason to expect Tranche to thrive as a data sharing tool for proteomics via the NRPP. An obvious extension to Tranche would be to extend use to other disciplines of science outside of proteomics. Tranche itself is not tied specifically to proteomics. Existing efforts are underway to use Tranche in the fields of glycomics and glycoproteomics, metabalomics, and 2D gels; however, only time will tell how successful those efforts are. Continued use of Bonanza on the majority of data sets present in Tranche is an appealing concept. Currently, Bonanza has shown that single data set analyses and a handful of data sets from a single mass spectrometer can yield significant insights into artifacts present in both the mass spectrometer and the experimental protocols used. Automated use of Bonanza with most any existing MS/MS search algorithm is clearly of benefit; however, this type of automated use would be of particular benefit to the proteomics community if the majority of data sets in Tranche were analyzed. Such results would provide an excellent approximation of search parameters appropriate in specific mass spectrometer and MS/MS search engine combinations. Furthermore, Bonanza can accurately estimate unexpected peptide modifications and unidentified portions of MS/MS data sets that are repeatedly observed. It would be intriguing to have such large-scale multi-dataset Bonanza results to work with, and the Bonanza work presented in this thesis grows into such use or inspires other tools to be used in similar ways.

Figures

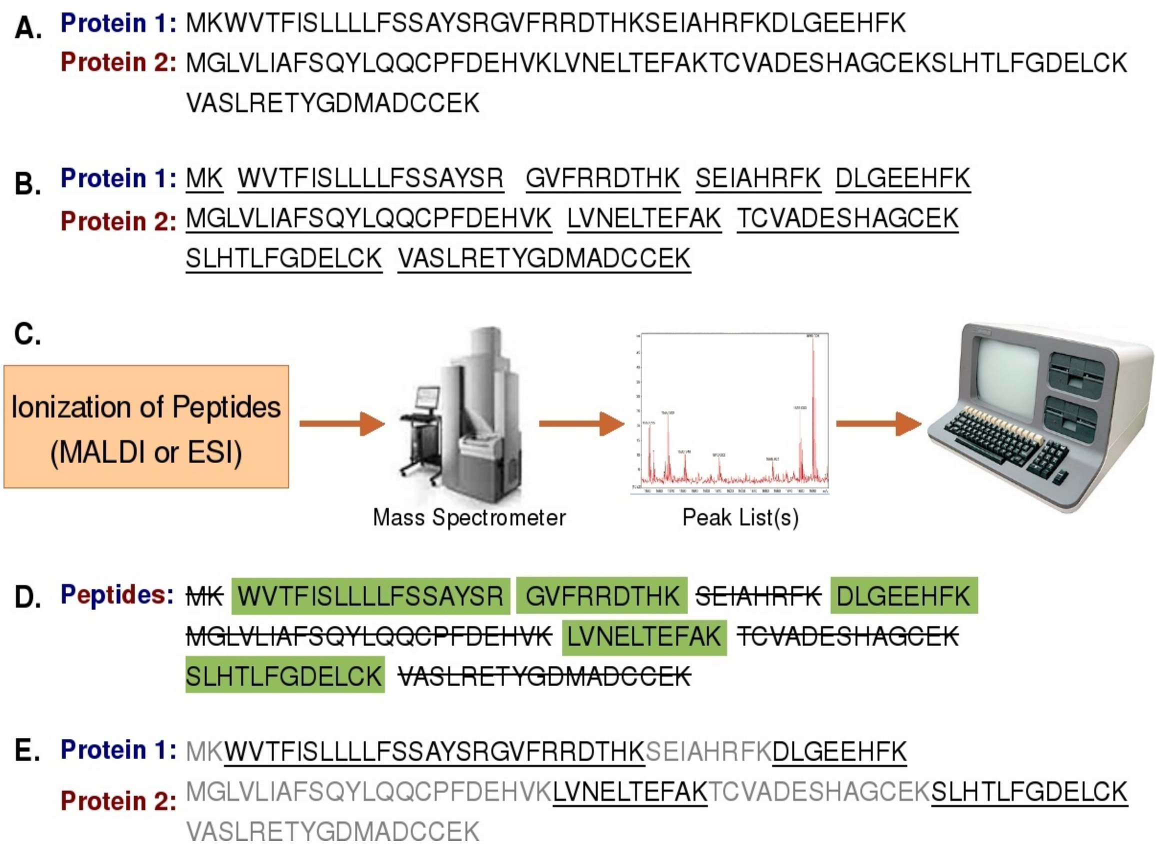

A. Protein 1: MKWVTFISLLLLFSSAYSRGVFRRDTHKSEIAHRFKDLGEEHFK
Protein 2: MGLVLIAFSQYLQQCPFDEHVKLVNELTEFAKTCVADESHAGCEKSLHTLFGDELCK VASLRETYGDMADCCEK
B. Protein 1: MK WVTFISLLLLFSSAYSR GVFRRDTHK SEIAHRFK DLGEEHFK
Protein 2: MGLVLIAFSQYLQQCPFDEHVK LVNELTEFAK TCVADESHAGCEK SLHTLFGDELCK VASLRETYGDMADCCEK
C.
Ionization of Peptides (MALDI or ESI)
Mass Spectrometer
Peak List(s)
D. Peptides: ~~MK~~ WVTFISLLLLFSSAYSR GVFRRDTHK ~~SEIAHRFK~~ DLGEEHFK
~~MGLVLIAFSQYLQQCPFDEHVK~~ LVNELTEFAK ~~TCVADESHAGCEK~~
SLHTLFGDELCK ~~VASLRETYGDMADCCEK~~
E. Protein 1: MKWVTFISLLLLFSSAYSRGVFRRDTHKSEIAHRFKDLGEEHFK
Protein 2: MGLVLIAFSQYLQQCPFDEHVKLVNELTEFAKTCVADESHAGCEKSLHTLFGDELCK
VASLRETYGDMADCCEK

Figure 1-1 – An informatics focused overview of a shotgun proteomics experiment. (a) Isolated and relatively purified proteins are prepared. (b) Trypsin is used to cleave the proteins into smaller peptides. (c) Peptides are ionized in to a mass spectrometer for analysis. (d) Peak lists are analyzed with software to infer likely peptides present. (e) Further software-based analysis infers likely proteins present.

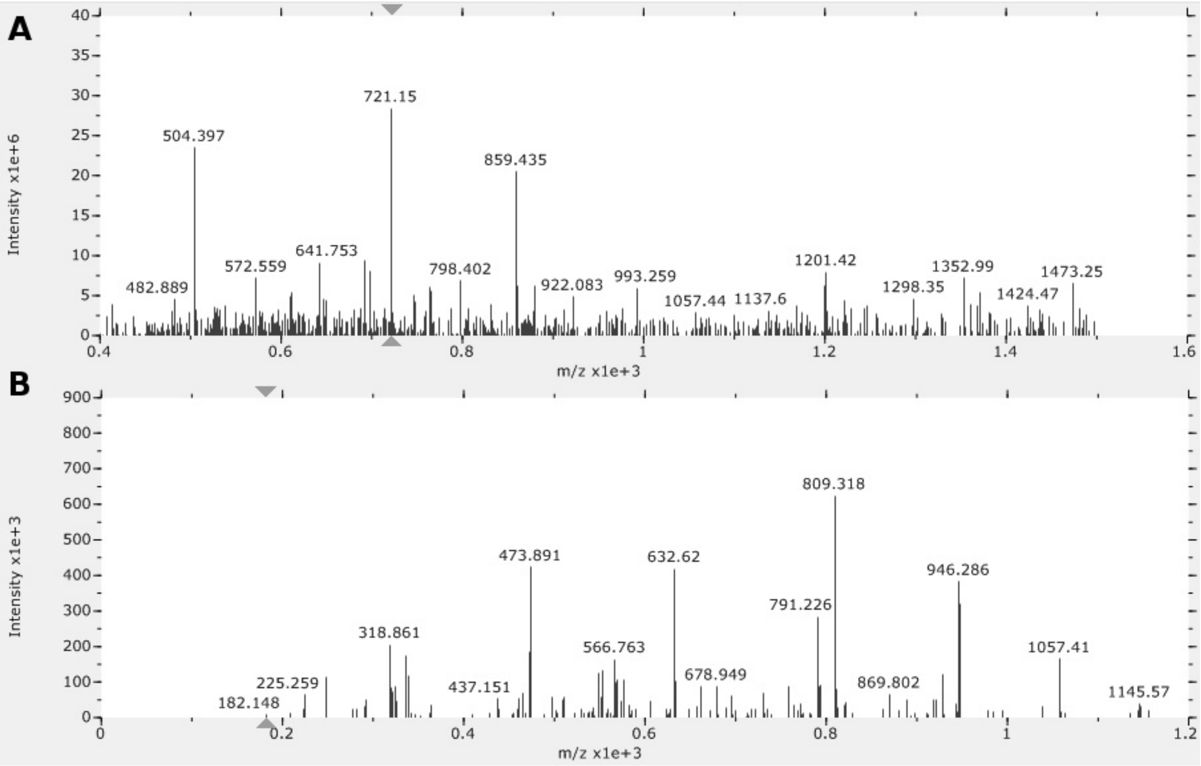


Figure 1-2 – Example Mass Spectra. (a) MS and (b) MS/MS. MS data typically represents the mass of ionized tryptic peptides. MS/MS data typically focuses in on a particular ionized peptide to help determine the amino acid sequence.

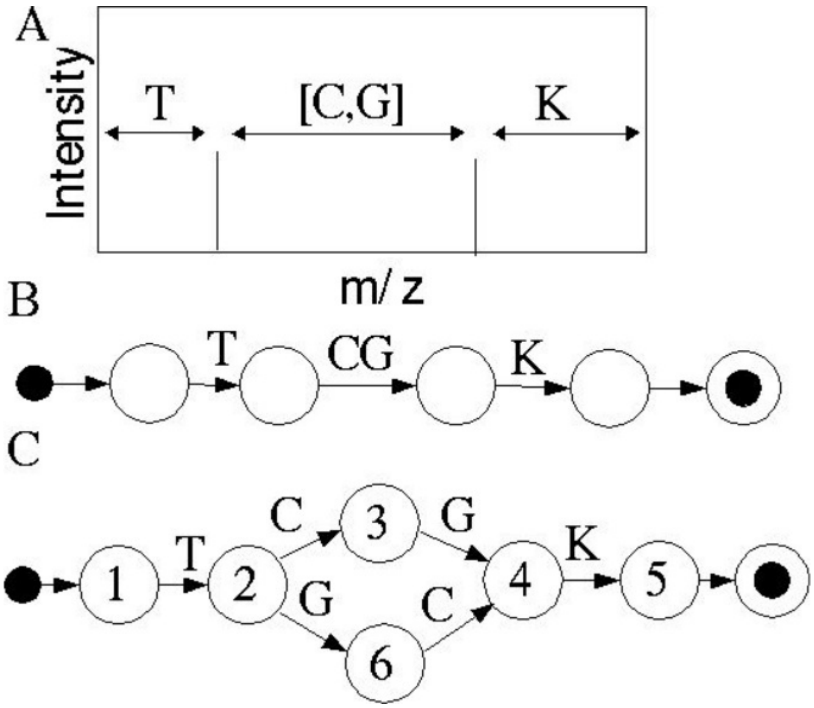


Figure 2-1 – Example PFSM figure(A) A peak list missing the second ion in the ion series TCGK – or since it is ambiguous, TGCK. The m/z difference is annotated with [C,G] because it is assumed that the m/z of adding the two residues gives the appropriate m/z to bridge the ion series. (B) The graph conversion of the given peak list allowing arcs to have multiple residues. (C) The NDFA conversion of the graph in part B, illustrating how to convert multiple residue to a single residue transitions. The solution is to create all combinations of the residues.

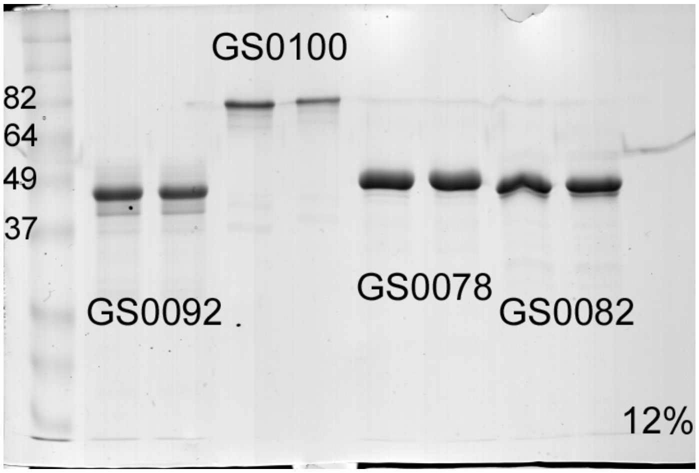
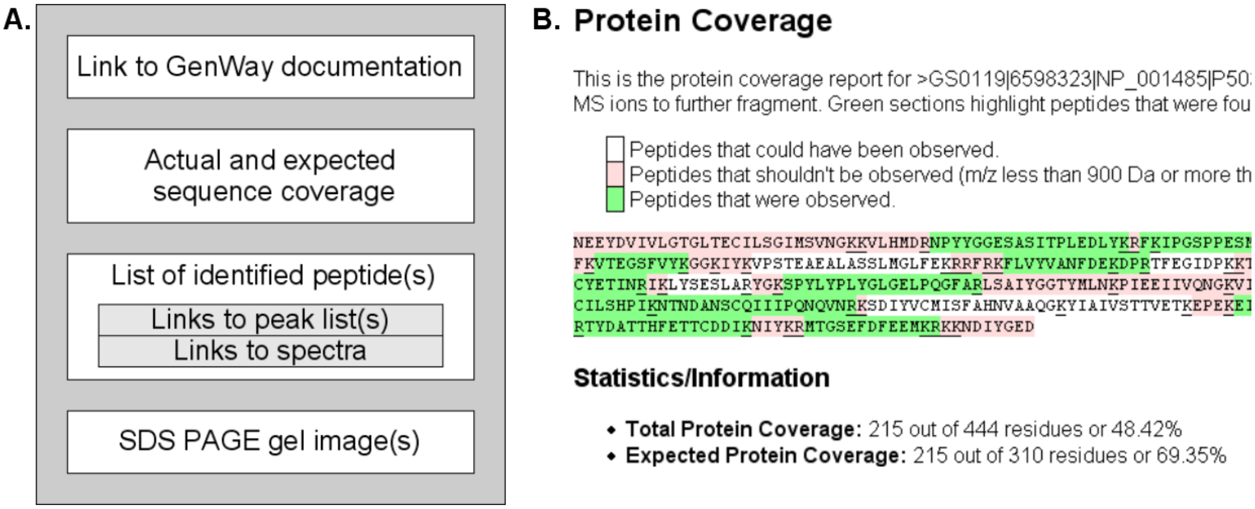


Figure 3-1 – An example SDS PAGE check for protein purity. Each protein in the Aurum Dataset was checked for purity using a hand cast polyacrylamide gel run under protein denaturing conditions (SDS PAGE). Multiple proteins were run on the same gel and each protein was run in two lanes. Proteins are labeled by their Aurum identification number. Only the predominant band was excised, digested, and analyzed using MS/MS. A gel image is included with this manuscript's data for every protein analyzed and the gels are linked in the protein summary report (Figure 2). Gel images also include a protein standard ladder in the first lane (approximate kDa labeled) and the percent polyacrylamide used to cast the gel in the bottom-right.


A.
Link to GenWay documentation
Actual and expected sequence coverage
List of identified peptide(s)
Links to peak list(s)
Links to spectra
SDS PAGE gel image(s)
B.

Protein Coverage

This is the protein coverage report for >GS0119|6598323|NP_001485|P50
MS ions to further fragment. Green sections highlight peptides that were fou

□ Peptides that could have been observed.
🟥 Peptides that shouldn't be observed (m/z less than 900 Da or more th
🟩 Peptides that were observed.

```
NEEYDVIVLGTGLTECILSGIMSVNGKKVLHMDRNPYYGGESASITPLEDLYKRFKIPGSPPESM
FKVTEGSFVYKGGKIYKVPSTEAEALASSLMGLFEKRRFRKFLVYVANFDEKDPRTFEGIDPKKT
CYETINRIKLYSESLARYGKSPYLYPLYGLGELPQGFARLSAIYGGTYMLNKPIEEIIVQNGKVI
CILSHPIKNTNDANSCQIIIPQNQVNRKSDIYVCMISFAHNVAAQGKYIAIVSTTVETKEPEKEI
RTYDATTHFETTCDDIKNIYKRMTGSEFDFEEMKRKKNDIYGED
```

Statistics/Information

- **Total Protein Coverage:** 215 out of 444 residues or 48.42%
- **Expected Protein Coverage:** 215 out of 310 residues or 69.35%

Figure 3-2 – Protein report file structure and partial example. (A) Block diagram describing the information in each of the protein report files included with the supplementary information. The intention of these pages is to provide a human-friendly summary for each protein analyzed. (B). Example protein coverage information from a protein summary page. This is only a portion of the summary page highlighting total protein coverage and coverage of expected peptides. Not shown in the figure are the complete statistics, peptide matches, and links to spectra, peak lists and other documentation.

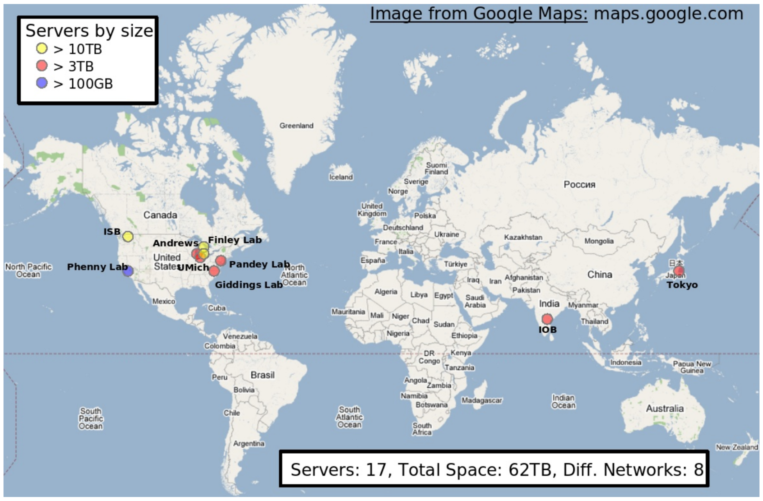


Figure 4-1 A snapshot of the Tranche core servers. The Google map widget was used to show the following snapshot of the core Tranche servers. These are servers dedicated to sharing data on the Tranche network.

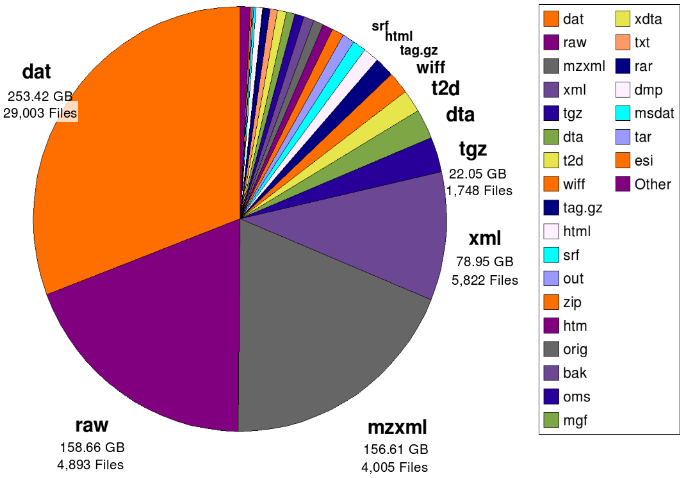


Figure 4-2 Snapshot of file types in Tranche. Files in Tranche listed by size with information about the number of files. Files with .dat are primarily from Water's .raw directory structures, and .tgz are primarily the results of Sequest searches.

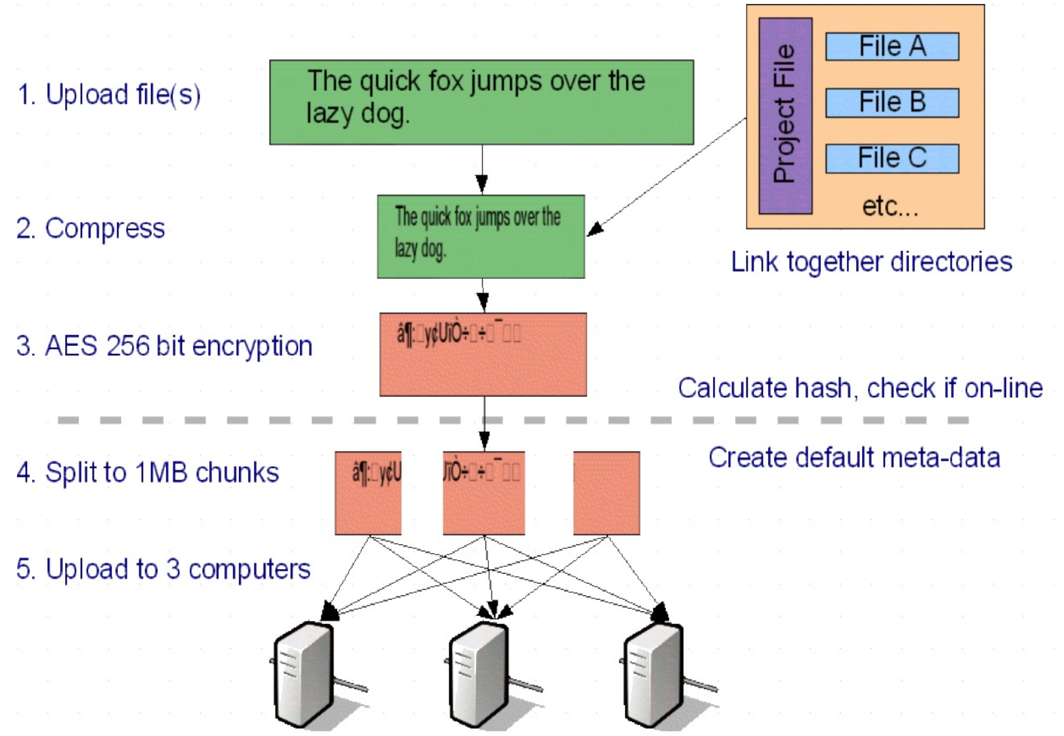

1. Upload file(s)
The quick fox jumps over the lazy dog.
Project File
File A
File B
File C
etc...
Link together directories
2. Compress
The quick fox jumps over the lazy dog.
3. AES 256 bit encryption
Calculate hash, check if on-line
4. Split to 1MB chunks
Create default meta-data
5. Upload to 3 computers

Figure 5-1 – Overview of the data upload process in Tranche. Tranche automatically handles compression, encryption, splitting of files, and replicating data on multiple servers.

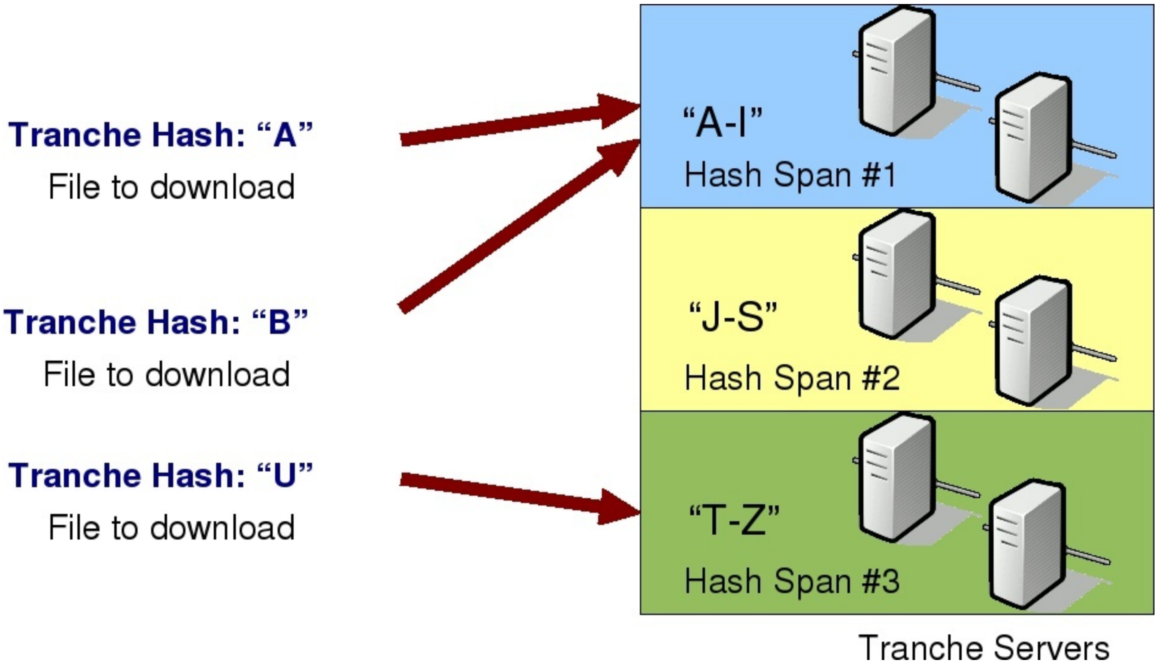


Figure 5-2 – Illustration of a Tranche hash span. Content is evenly spread across servers in the Tranche network by pre-configured “hash spans”. This ensures that at least a certain number of servers, normally 3, get a copy of each bit of data.

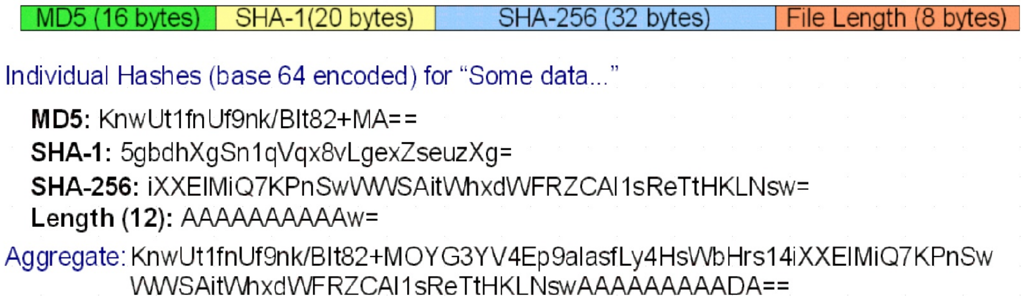

MD5 (16 bytes)
SHA-1(20 bytes)
SHA-256 (32 bytes)
File Length (8 bytes)
Individual Hashes (base 64 encoded) for “Some data...”
MD5: KnwUt1fnUf9nk/BIt82+MA==
SHA-1: 5gbdhXgSn1qVqx8vLgexZseuzXg=
SHA-256: iXXEIMiQ7KPnSwWWSAitWhxdWFRZCAI1sReTtHKLNsw=
Length (12): AAAAAAAAAAw=
Aggregate: KnwUt1fnUf9nk/BIt82+MOYG3YV4Ep9alasfLy4HsWbHrs14iXXEIMiQ7KPnSw
WWSAitWhxdWFRZCAI1sReTtHKLNswAAAAAAAAADA==

Figure 5-3 – Structure of a Tranche Hash. Tranche uses existing hashing algorithms to create a secure and unique identifier for data sets. The scheme is a combination of the MD5, SHA-1, and SHA-256 hashes plus the length of the file.

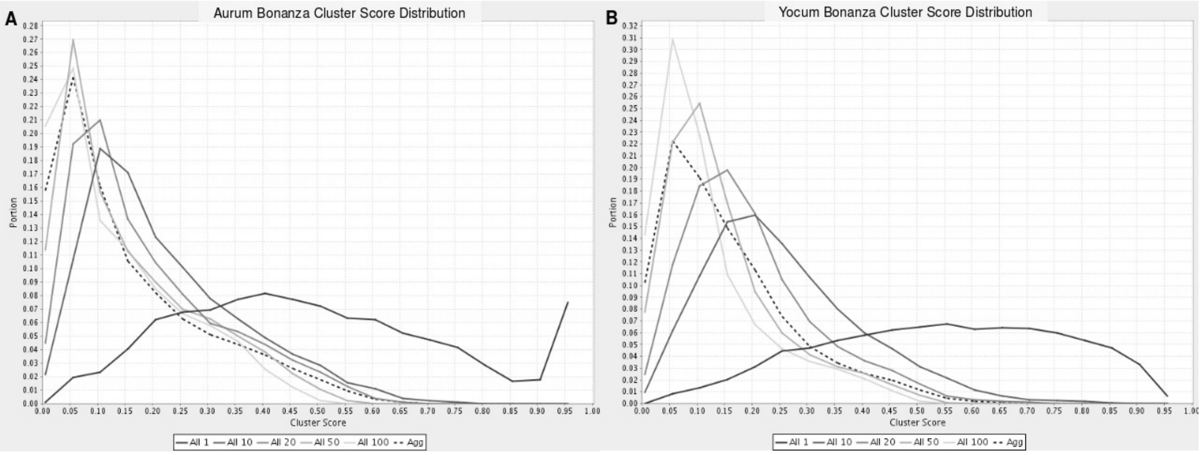


Figure 6-1 – Distribution of bonanza scores. Distributions of Bonanza scores for the all peak lists compared against all other peak lists for the Aurum and Yocum data sets, respectively. The ratio of the 1st best cluster score compared to the aggregate cluster score distribution is used for approximating valid peak list clusterings. Lower ranked cluster scores (10th, 20th, 50th, 100th) quickly converge to a presumed distribution of invalid clusterings.

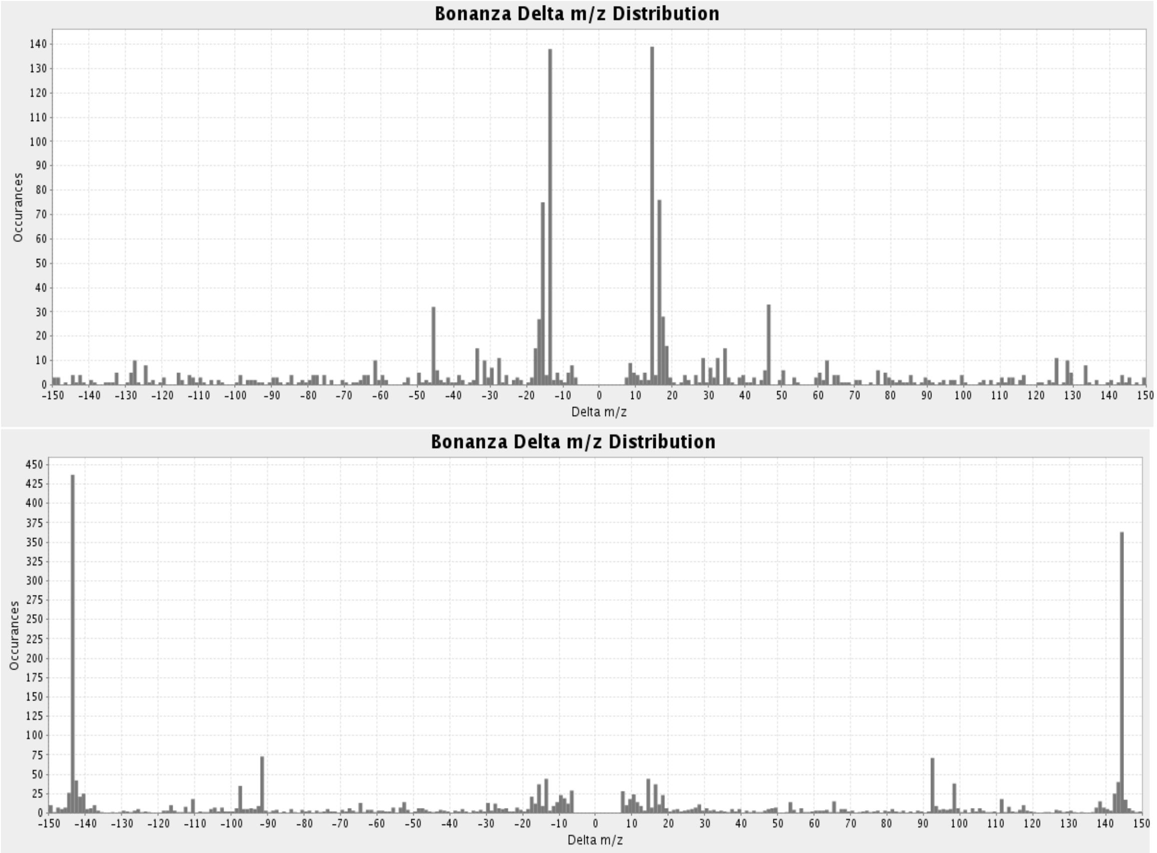


Figure 6-2 – Histogram of m/z precursor differences in Bonanza clustered peak listsHistogram of *m/z* difference between clustered peak lists. Only clusterings where the bonanza score is above threshold are plotted, and the +/-6 Da range is omitted. In both plots 0 Da is the highest bin (see appendix A for complete plots). Significant trends can be observed, which support the proposed use of Bonanza for identifying unexpected modifications.

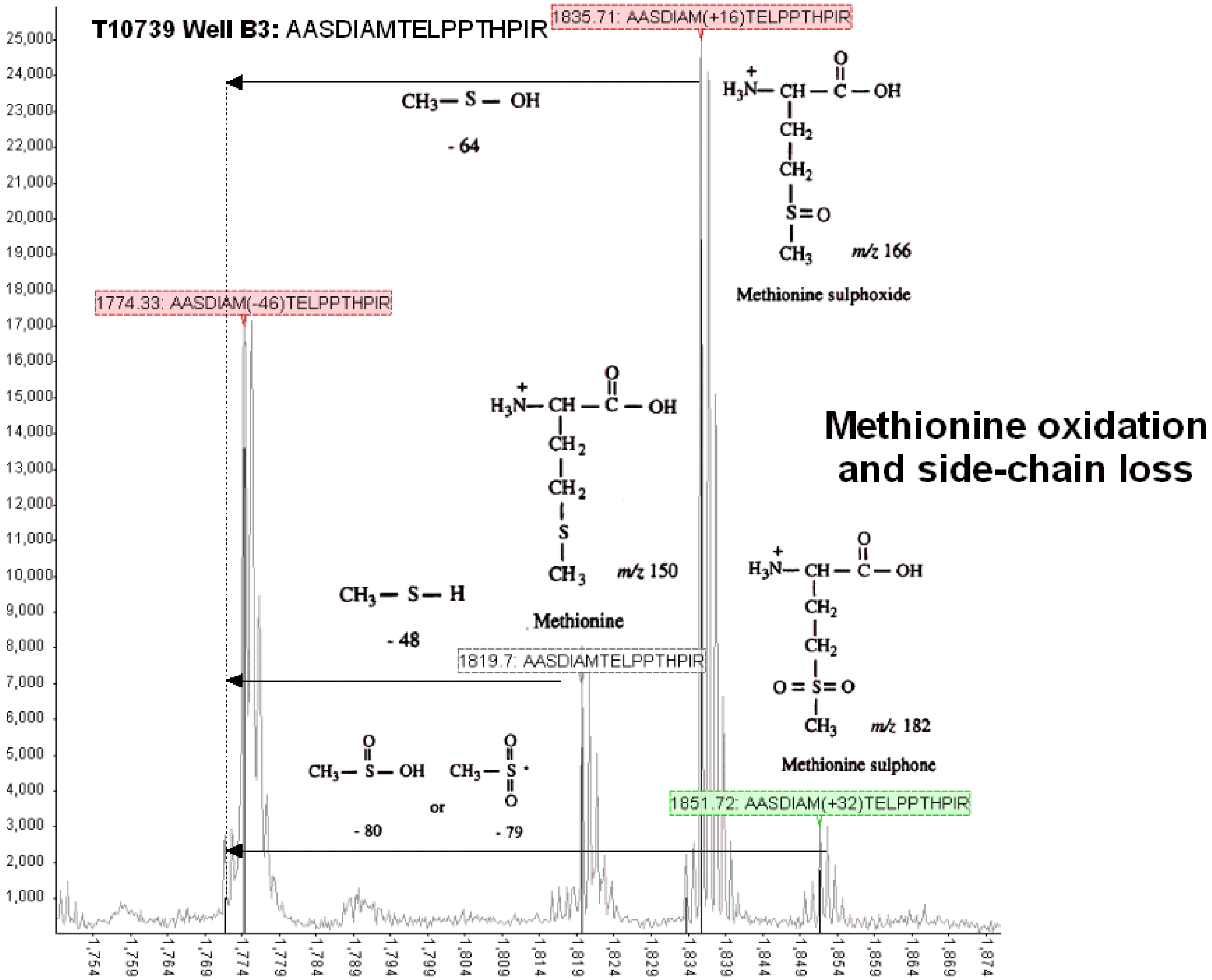


Figure 6-3 – Example of observed methionine side-chain loss in MS. Examination of the Aurum MS data helps explain an unusual neutral loss. In light gray is the actual MS spectra. The black lines are the called peaks from the peak list. The masses 1,189.7, 1,835.7, and 1,851.7 are the respective unmodified, oxidized and doubly oxidized peptides. If any form of the peptide has a neutral loss of the side-chain a peak appears at a net nominal loss of -48 Da from the unmodified peptide.

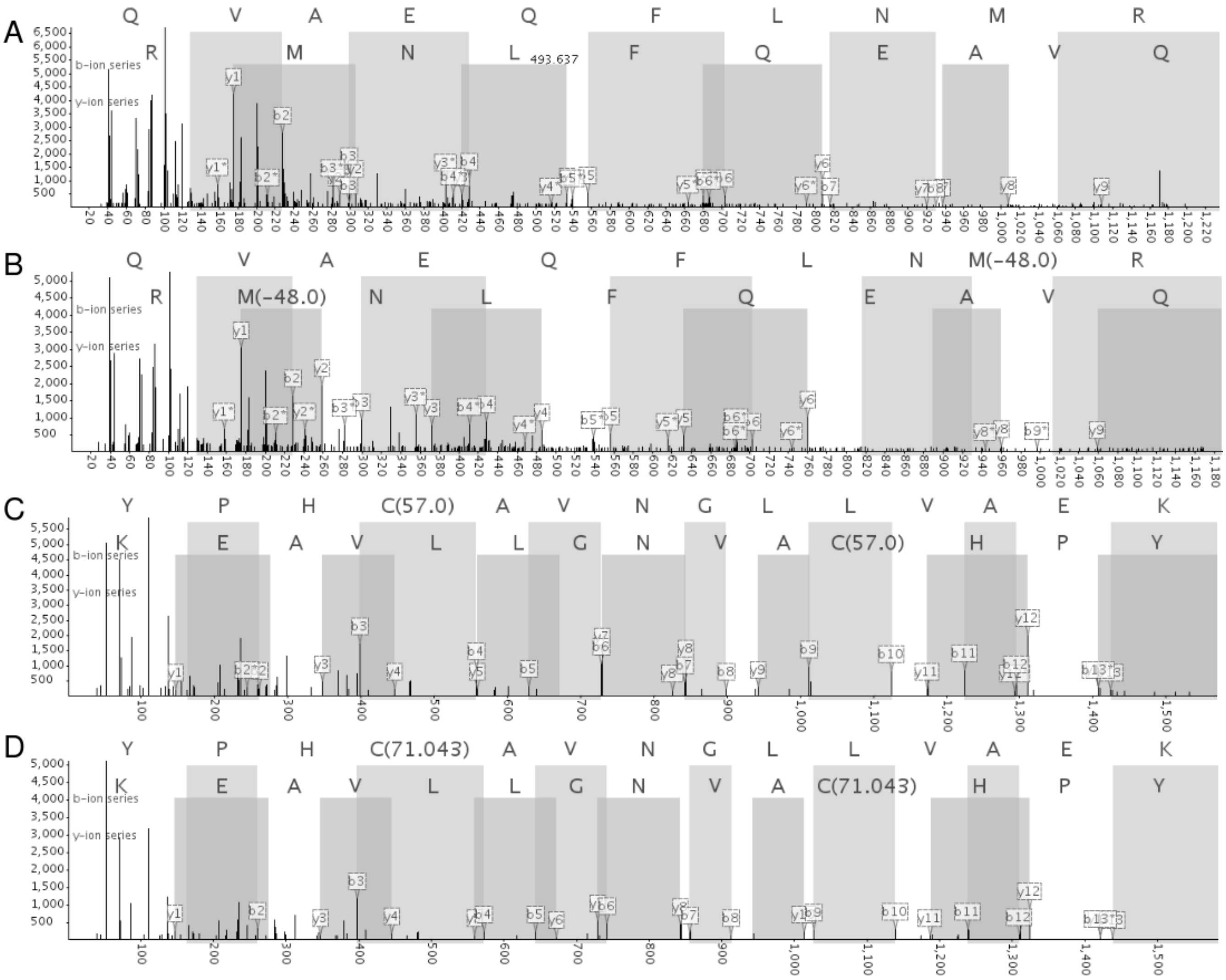


Figure 6-4 – Select modifications identified by Bonanza and confirmed by MS/MS peak lists. (A) and (B) are peak lists clustered together. The decoy database search identified the unmodified peak list (A) to be QVAEQFLNMR and Bonanza inferred (B) with the neutral loss of the methionine side chain. (C) and (D) are another set of peak lists clustered together. The decoy database analysis identified (C) as YPHCAVNGLLVAEK with a carbamidomethylated cysteine. Bonanza inferred the artifactual acrylamide adduct due to PAGE used for purification.

Tables

	Aurum Dataset	Yocum Dataset
Starting Peak Lists	9,987	41,942
Filtered Peak Lists (>10 peaks)	9,350	37,772
Peak Lists w/o Clustering	3,998	13,727
Unidentified Clusters	2,274	14,023
Decoy Analysis Peptide Identifications	2,594	6,372
Bonanza Inferred Peak Lists	484	3,650

Table 6-1 – Overview of peak lists and clustering identifications during Bonanza analysis for the Aurum and Yocum data set. The Yocum data set represents several MudPIT experiments, and is much more representative of a shotgun proteomics experiment compared to the purified proteins used in Aurum.

	Bonanza *(m/z); count*	Yocum *(m/z); count*
1	C(57); 760	K(144); 4,607
2	M(16); 342	Y(144); 1,345
3	W(16); 82	C(46); 731
4	C(71); 56	N-term (144); many*
5	H(16); 45	M(16); 148

Table 6-2 – Summary of the top 5 modifications based on identified and bonanza-inferred peptide identifications. Two interesting observations. First, the Aurum data set appears to have many more oxidation events, especially considering it has 1/4th the amount of total peak lists. Second, modifications in the Yocum data set confirm obvious trends observed in Figure 6-2, and modifications in the Aurum data set confirms not so easily explained trends (+/-14 Da is due to propionamide) in respective plots in Figure 6-2.

* Many different residues with N-term iTRAQ were omitted from the list.

Bibliography

1. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. (1989). "Electrospray ionization for mass spectrometry of large biomolecules.". Science (journal) 246: 64-71. doi:10.1126/science. 2675315. PMID 2675315

2. Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T. (1988). "Protein and Polymer Analyses up to m/z 100 000 by Laser Ionization Time-of flight Mass Spectrometry". Rapid Commun Mass Spectrom 2 (20): 151-3.

3. Strupat K, Karas M, Hillenkamp F (1991). "2,5-Dihidroxybenzoic acid: a new matrix for laser desorption-ionization mass spectrometry.". Int. J. Mass Spectrom. Ion Processes 72 (111): 89-102.

4. Markides, K; Gräslund, A. Advanced information on the Nobel Prize in Chemistry 2002

5. Jimmy K. Eng, Ashley L. McCormack, and John R. Yates, III (1994). "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database". J Am Soc Mass Spectrom 5: 976-989.

6. Wysocki VH, Resing KA, Zhang Q, Cheng G., "Mass spectrometry of peptides and proteins." Methods. 2005 Mar;35(3):211-22. Epub 2005 Jan 20.

7. McDonald WH, Yates JR 3rd, "Shotgun proteomics: integrating technologies to answer biological questions." Curr Opin Mol Ther. 2003 Jun;5(3):302-9.

8. Alexey I. Nesvizhskii and Ruedi Aebersold, "Interpretation of Shotgun Proteomic Data", Molecular & Cellular Proteomics 4:1419-1440, 2005.

9. Hanno Steen and Matthias Mann, "THE ABC’S (AND XYZ’S) OF PEPTIDE SEQUENCING ", Nature Reviews: Molecular Cellular Biology 5:699, 2004

10. Kermit K. Murray, Robert K. Boyd, Marcos N. Eberlin, G. John Langley, Liang Li and Yasuhide Naito, "STANDARD DEFINITIONS OF TERMS RELATING TO MASS SPECTROMETRY " International Union of Pure and Applied Chemistry Analytical Chemistry Division 2006

11. Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins" Nucleic Acids Res. 2007 January; 35(Database issue): D61–D65.

12. Kersey P. J., Duarte J., Williams A., Karavidopoulou Y., Birney E., Apweiler R., "The International Protein Index: An integrated database for proteomics experiments." Proteomics 4(7): 1985-1988 (2004).

13. Jayson A. Falkner, James A. Hill, Philip C. Andrews, "Proteomics FASTA Archive and Reference Resource" Proteomics (in publication) 2008

14. Pappin DJ, Hojrup P, Bleasby AJ, "Rapid identification of proteins by peptide-mass fingerprinting" Curr Biol. 1993 Jun 1;3(6):327-32.Click here to read

15. Eriksson J, Chait BT, Fenyˆ D, "A statistical basis for testing the significance of mass spectrometric protein identification results." Anal Chem. 2000 Mar 1;72(5):999-1005.

16. Huang Y, Tseng GC, Yuan S, Pasa-Tolic L, Lipton MS, Smith RD, Wysocki VH., "A Data-Mining Scheme for Identifying Peptide Structural Motifs Responsible for Different MS/MS Fragmentation Intensity

Patterns." J Proteome Res. 2008 Jan;7(1):70-9. Epub 2007 Dec 4.

17. Huang Y, Triscari JM, Tseng GC, Pasa-Tolic L, Lipton MS, Smith RD, Wysocki VH., "Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns." Anal Chem. 2005 Sep 15;77(18):5800-13.

18. Tabb DL, Huang Y, Wysocki VH, Yates JR 3rd., "Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides." Anal Chem. 2004 Mar 1;76(5):1243-8.

19. Jimmy K. Eng, Ashley L. McCormack, and John R. Yates, III (1994). "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database". J Am Soc Mass Spectrom 5: 976-989.

20. David N. Perkins, Darryl J. C. Pappin, David M. Creasy, John S. Cottrell, Probability-based protein identification by searching sequence, databases using mass spectrometry data, Volume 20, Issue 18 , Pages 3551 - 3567

21. TANDEM: matching proteins with mass spectra, Robertson Craig and Ronald C. Beavis, Bioinformatics, 2004, 20, 1466-7.

22. Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. "OLAV: towards high-throughput tandem mass spectrometry data identification," Proteomics, Vol. 3, No. 8, August 2003, pp. 1454-1463.

23. Tabb DL, Fernando CG, Chambers MC., MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res. 2007 Feb;6(2):654-61.

24. Searle BC, Turner M, Nesvizhskii AI., Improving Sensitivity by Probabilistically Combining Results from Multiple MS/MS Search Methodologies., J Proteome Res. 2008 Jan;7(1):245-53.

25. Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM, Schaeffer DA., The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra., Mol Cell Proteomics. 2007 Sep;6(9):1638-55. Epub 2007 May 27.

26. Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, Gilles Lajoie. PEAKS: Powerful Software for Peptide De Novo Sequencing by MS/MS. Rapid Communications in Mass Spectrometry, 17(20):2337-2342. 2003. Early version appeared in 50th ASMS Conference 2002.

27. Clauser, K. R., Baker, P., Burlingame A. L. (1999) Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal Chem. 71, 2871-82.

28. Bafna, V., Edwards, N. (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. Bioinformatics 17 Suppl. 1, S13-21.

29. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., Bryant, S. H. (2004) Open mass spectrometry search algorithm. J Proteome Res. 3, 958-64.

30. Keller A, Eng J, Zhang N, Li XJ, Aebersold R., A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol. 2005;1:2005.0017. Epub 2005 Aug 2.

31. Zhang B, Chambers MC, Tabb DL., Proteomic parsimony through bipartite graph analysis improves accuracy and transparency., J Proteome Res. 2007 Sep;6(9):3549-57. Epub 2007 Aug 4.

32. A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes, David Fenyo and Ronald C. Beavis, Anal. Chem., 2003, 75, 768-774.

33. Interpretation of shotgun proteomic data: the protein inference problem., Nesvizhskii AI, Aebersold R., Mol Cell Proteomics. 2005 Oct; 4(10):1419-40. Epub 2005 Jul 11.

34. Bandeira, N., Tsur, D., Frank, A., and Pevzner, P.A. (2006) Protein identification by spectral networks analysis. PNAS April 10th, vol 104, no. 15

35. Craig, R., Cortens, J.P., Fenyo, D., and Beavis, R.C. (2006) Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. Proteome Res 10, 1021/pr0602085

36. Frewen, B., Merrihew, G., Wu, C.C., Noble, W.S., MacCoss, M.J. (2006) Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries Anal Chem 78, 5678-5684

37. Lam, H., Deutsch, E.W., Eddes, J.S., Eng, J.K., King, N., Stein, S.E., Aebersold, R. (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 7(5), 655-67.

38. Jayson Falkner and Philip C. Andrews, A solution to sharing and citing large scientific data sets based on developments in the field of proteomics, in publication (2008)

39. Desiere, F, et al., The PeptideAtlas project. Nucleic Acids Res. 34: D655-D658 (2006)

40. Martens, L, et al. PRIDE: the proteomics identifications database. Proteomics 5 : 3537-3545 (2005)

41. Craig, R, Cortens, JP, and Beavis, RC, Open source system for analyzing, validating, and storing protein identification data. J. Proteome Res. 3 : 1234-1242 (2004)

42. TANDEM: matching proteins with mass spectra, Robertson Craig and Ronald C. Beavis, Bioinformatics, 2004, 20, 1466-7.

43. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH *Open mass spectrometry search algorithm* J Proteome Res. 2004 Sep-Oct;3(5):958-64

44. Carr, S, et al. The Need for Guidelines in Publication of Peptide and Protein Identification Data. Molecular and Cellular Proteomics 3.6. (2004)

45. Bradshaw, RA, Burlingame, AL, Carr, S, and Aebersold, R, Reporting protein identification data: the next generation of guidelines. Mol. Cell. Proteomics 5 : 787-788 (2006)

46. Wilkins, MR, et al. Guidelines for the next 10 years of proteomics. Proteomics 6 : 4-8 (2006)

47. Choi H, Nesvizhskii AI., Semisupervised model-based validation of Peptide identifications in mass spectrometry-based proteomics. J Proteome Res. 2008 Jan;7(1):254-65. Epub 2007 Dec 27.

48. Kall L, Storey JD, Maccoss MJ, Noble WS., Assigning significance to peptides identified by tandem mass spectrometry using decoy

databases., J Proteome Res. 2008 Jan;7(1):29-34. Epub 2007 Dec 8.

49. Elias JE, Gygi SP., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry., Nat Methods. 2007 Mar;4(3):207-14.

50. Martin McIntosh, Developing and Disseminating Advances in Computational and Statistical Proteomics, Journal of Proteome Research Vol. 7, No. 1: January 2008

51. JA Falkner, PJ Ulintz, PC Andrews, ProteomeCommons. org: Code and Data Archive and Dissemination for the Proteomics Community, American Biotechnology Laboratory (ABL), in publication, Dec, 2005

52. Jayson A. Falkner, Maureen Kachman, Donna M. Veine, Angela Walker, John R. Strahler and Philip C. Andrews, Validated MALDI-TOF/TOF Mass Spectra for Protein Standards, Journal of the American Society for Mass Spectrometry Volume 18, Issue 5, May 2007, Pages 850-855

53. Jayson Falkner and Philip C. Andrews, A solution to sharing and citing large scientific data sets based on developments in the field of proteomics, in publication (2008)

54. JA Falkner, PJ Ulintz, PC Andrews, ProteomeCommons. org: Code and Data Archive and Dissemination for the Proteomics Community, American Biotechnology Laboratory (ABL), in publication, Dec, 2005

55. Falkner JA, Falkner JW, Andrews PC, ProteomeCommons.org JAF: reference information and tools for proteomics Bioinformatics. 2006 Mar 1;22(5):632-3. Epub 2006 Jan 24.

56. Falkner JA, Falkner JW, Andrews PC ProteomeCommons.org IO

Framework: reading and writing multiple proteomics data formats Bioinformatics. 2007 Jan 15;23(2):262-3. Epub 2006 Nov 22

57. Jimmy K. Eng, Ashley L. McCormack, and John R. Yates, III (1994). "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database". J Am Soc Mass Spectrom 5: 976-989.

58. A Method for Reducing the Time Required to Match Protein Sequences with Tandem Mass Spectra, Robertson Craig and Ronald C. Beavis; Rapid Commun. Mass Spectrom., 2003, 17: 2310-2316.

59. Falkner J, Andrews P. Fast tandem mass spectra-based protein identification regardless of the number of spectra or potential modifications examined. Bioinformatics. 2005 May 15;21(10):2177-84. Epub 2005 Mar 3

60. Sipser, Michael. "Chapter 1: Regular Languages", Introduction to the Theory of Computation. PWS Publishing, 31–90. ISBN 0-534-94728-X.

61. Eng, JK, McCormack, AL, and Yates JR III. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5, 976–989.

62. Perkins, DN., Pappin, DJC., Creasy, DM., et al. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3551–3567.

63. Craig, R. and Beavis, R.C., "TANDEM: matching proteins with mass spectra", Bioinformatics, 2004, 20, 1466-7.

64. Clauser, KR, Baker PR., Burlingame AL. (1999) Role of accurate

mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal Chem. 71(14): 2871-82.

65. Narasimhan C, Tabb DL, Verberkmoes NC, Thompson MR, Hettich RL, Uberbacher EC. (2005) MASPIC: intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. Anal Chem. 77(23):7581-93.

66. Colinge J, Masselot A, Cusin I, Mahe E., Niknejad A., Argoud-Puy G, Reffas S, Bederr N, Gleizes A, Rey P, Bougueleret L, "High-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics”, Proteomics, Vol. 4, Issue 7, Pages 1977 – 1984.

67. Tabb DL, Narasimham C, Strader MB, and Hettich RL, “DBDigger: Reorganized Proteomic Database Identification That Improves Flexibility and Speed”, Anal. Chem.2005, 77,2464-2474

68. Taylor JA and Johnson RS, "Implementation and Uses of Automated de Novo Peptide Sequencing by Tandem Mass Spectrometry", Anal. Chem. 2001, 73, 2594-2604

69. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, and Lajoie G, "PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry", Rapid Comm. In Mass Spec., 2003, Vol. 17, Issue 20, Pages 2337 – 2342

70. Frank A and Pevzner P, "PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling", Anal. Chem. 2005, 77, 964-973

71. Craig R, Cortens JC, Fenyo D, and Beavis RC, "Using Annotated Peptide Mass Spectrum Libraries for Protein Identification", J. Proteome

Res.; 2006; 5(8) pp 1843 – 1849

72. NIST, "NIST Library of Peptide Ion Fragmentation Spectra: Version June 2006", http://chemdata.nist.gov/mass-spc/ftp/mass-spc/PepLib.pdf

73. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. (2002A) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 74(20):5383-92.

74. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 75(17):4646-58.

75. Elias, JE, Haas W, Faherty, BK, and Gygi SP, “Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations”, Nature Methods 2, 667 - 675 (2005)

76. Purvine S, Kolker N, Kolker E. (2004) Spectral quality assessment for high-throughput tandem mass spectrometry proteomics. OMICS. 8(3):255-65.

77. Keller A, Purvine S, Nesvizhskii AI, Stolyar S, Goodlett DR, Kolker E. (2002B) Experimental protein mixture for validating tandem mass spectral analysis. OMICS.6(2):207-12.

78. Martin DB, Eng JK, Nesvizhskii AI, Gemmill A, Aebersold R. (2005) Investigation of neutral loss during collision-induced dissociation of peptide ions. Anal Chem. 77(15):4870-82.

79. Tabb DL, Smith LL, Breci LA, Wysocki VH, Lin D, Yates Jr, III (2003) Statistical characterization of Ion trap tandem mass spectra from doubly charged tryptic peptides. Anal Chem 75, 1155-1163.

80. Kapp EA, Schutz F, Connolly LM, Chakel JA, Meza JE, Miller CA,

Fenyo D, Eng JK, Adkins JN, Omenn GS, Simpson RJ. (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics. 5(13):3475-90.

81. Omenn GS et al. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. Proteomics. 2005 Aug;5(13): 3226-45.

82. Rivest R., "The MD5 Message-Digest Algorithm", Network Working Group, Request for Comments: 1321, April 1992

83. Falkner, J.A., Falkner, J.W., Andrews, P.C., "ProteomeCommons.org IO Framework: reading and writing multiple proteomics data formats", Bioinformatics, in publication

84. Elias, JE, Haas W, Faherty, BK, and Gygi SP, “Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations”, Nature Methods 2, 667 - 675 (2005)

85. Democratizing proteomics data. Nat. Biotech. 25 (1). (2007)

86. Carr, S, et al. The Need for Guidelines in Publication of Peptide and Protein Identification Data. Molecular and Cellular Proteomics 3.6. (2004)

87. Bradshaw, RA, Burlingame, AL, Carr, S, and Aebersold, R, Reporting protein identification data: the next generation of guidelines. Mol. Cell. Proteomics 5 : 787-788 (2006)

88. Wilkins, MR, et al. Guidelines for the next 10 years of proteomics.

Proteomics 6 : 4-8 (2006)

89. Lennart, M, et al., Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. Proteomics 5, 3501–3505 (2005)

90. Desiere, F, et al., The PeptideAtlas project. Nucleic Acids Res. 34: D655-D658 (2006)

91. Prince, JT, Carlson, MW, Wang, R, Lu, P, and Marcotte, EM, The need for a public proteomics repository. Nat. Biotechnol. 22 : 471-472 (2004)

92. Rauch, A, et al., Computational Proteomics Analysis System (CPAS): An Extensible, Open-Source Analytic System for Evaluating and Publishing Proteomic Data and High Throughput Biological Experiments. J. Proteome Res. 5 : 112-121 (2006)

93. Craig, R, Cortens, JP, and Beavis, RC, Open source system for analyzing, validating, and storing protein identification data. J. Proteome Res. 3 : 1234-1242 (2004)

94. Martens, L, et al. PRIDE: the proteomics identifications database. Proteomics 5 : 3537-3545 (2005)

95. Peri, S, et al. Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res. 32 : D497-D501 (2004)

96. Field, D, et al., Open Software for biologists: from famine to feast. Nat. Biotech. 24: 7 (2006)

97. Desiere, F, et al., The PeptideAtlas project. Nucleic Acids Res. 34: D655-D658 (2006)

98. Wisz, MS, Suarez, MK, Holmes, MR, and Giddings, MC, GFSWeb: A web tool for genome-based identification of proteins from mass spectrometric samples. J. Proteome Res. 3(6): 1292–1295. (2004)

99. Pedrioli, PGA, et al., A common open representation of mass spectrometry data and its application to proteomics research. Nat. Biotechnol. 22 : 1459-1466 (2004)

100. Eng, JK, McCormack, AL, and Yates, JR, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. JASMS 5 (11)

101. Omenn, GS, et al., Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. Proteomics 5(13) 3226 (2005)

102. Rivest R. (1992) The MD5 Message-Digest Algorithm. Network Working Group, Request for Comments: 1321, April 1992; http://tools.ietf.org/html/1321

103. National Institute of Standards and Technology (2002) Secure Hash Standard. FIPS 180-2

104. Housley R, Polk W, Ford W, and Solo D (2002) Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile. Network Working Group: Request for Comments: 3280, April 2002; http://tools.ietf.org/html/3280

105. Rivest R, Shamir A, Adleman L (1978) A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. Communications of the ACM 21(2): 120–126

106. Craig, R., Beavis, R. (2004) TANDEM: matching proteins with tandem mass spectra Bioinformatics 20(9), 1466–1467

107. Perkins, D.N., Pappin, D.J.C., Creasy, D.M., Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20 (18), 3551-3567

108. Eng, J.K., McCormack, A.L., Yates, J.R. III. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database Journal of the American Society for Mass Spectrometry 5, 976-989

109. Tsur, D., Tanner, S., Zandi, E., Bafna, V., Pevzner, P.A. Identification of post-translational modifications by blind search of mass spectra. Nat Biotech 23 (12)

110. Tabb, D.L., Narasimhan, C., Strader, M.B., Hettich, R.L. (2005) DBDigger: reorganized proteomic database identification that improves flexibility and speed. Anal Chem 77(8), 2464-74

111. Tabb, D.L., Fernando, C.G., and Chambers, M.C.(2007) MyriMatch: Highly Accurate Tandem Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis. J. Proteome Res. 6 (2), 654 -661 10.1021/pr0604054 S1535-3893(06)00405-2

112. Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., Bryant, S.H. (2004) Open Mass Spectrometry Search Algorithm. J Proteome Res 3(5), 958-964

113. Searle, B.C., Dasari, S., Wilmarth, P.A., Turner, M., Reddy, A.P., David, L.L., Nagalla, S.R. (2005) Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment

algorithm. J Proteome Res. 4(2), 546-54.

114. MacLean, B., Eng, J.K., Beavis, R.C., and McIntosh, M. (2006) General framework for developing and evaluating database scoring algorithms using the TANDEM search engine Bioinformatics 22(22), 2830-2832; doi:10.1093/bioinformatics/btl379

115. Colinge, J., Masselot, A., Giron, M., Dessingy, T., Magnin, J. (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics 3(8), 1454-63.

116. Hernandez, P., Gras, R., Frey, J., Appel, R.D. (2003) Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. Proteomics 3(6), 870-8.

117. Craig, R., Cortens, J.P., Fenyo, D., and Beavis, R.C. (2006) Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. Proteome Res 10, 1021/pr0602085

118. Frewen, B., Merrihew, G., Wu, C.C., Noble, W.S., MacCoss, M.J. (2006) Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries Anal Chem 78, 5678-5684

119. Bandeira, N., Tsur, D., Frank, A., and Pevzner, P.A. (2006) Protein identification by spectral networks analysis. PNAS April 10th, vol 104, no. 15

120. Lam, H., Deutsch, E.W., Eddes, J.S., Eng, J.K., King, N., Stein, S.E., Aebersold, R. (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 7(5), 655-67.

121. Perkins, D.N., Pappin, D.J.C., Creasy, D.M., Cottrell, J.S. (1999)

Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20 (18), 3551-3567

122. Craig, R., Beavis, R. (2004) TANDEM: matching proteins with tandem mass spectra Bioinformatics 20(9), 1466–1467

123. MacLean B, Eng JK, Beavis RC, McIntosh M, General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. Bioinformatics. 2006 Nov 15;22(22):2830-2. Epub 2006 Jul 28.

124. Eng, J.K., McCormack, A.L., Yates, J.R. III. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database Journal of the American Society for Mass Spectrometry 5, 976-989

125. Falkner, J.A., Kachman, M., Veine, D.M., Walker, A., Strahler, J.R., Andrews, P.C. (2007) Validated MALDI-TOF/TOF Mass Spectra for Protein Standards J Am Soc Mass Spectrom

126. Lagerwed, F.M., Marc van de Weert, Heerma, W., and Haverkamp, J. (1996) Identification of Oxidized Methionine in Peptides. Rapid Comm Mass Spec 10, 1905-1

127. Clarke, S. (1985) PROTEIN CARBOXYL METHYLTRANSFERASES: TWO DIST]INCT CLASSES OF ENZYMES. Annu Rev Biochem 54, 479-506.

128. Clarke, S. (1993) Protein Methylation Current Opinion in Cell Biology 5, 977-983

129. Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP,

Hunter CL, Nuwaysir LM, Schaeffer DA. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. Mol Cell Proteomics. 2007 Sep;6(9):1638-55. Epub 2007 May 27.