Delineating structural characteristics of viral capsid proteins critical for their functional assembly

by

Shanshan Cheng

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2014

Doctoral Committee:
Professor Charles L. Brooks III, Chair
Assistant Professor Barry J. Grant
Professor Kerby A. Shedden
Professor Janet L. Smith
Professor Yang Zhang

To my parents,
my husband
and my daughter

ACKNOWLEDGEMENTS

I am very fortunate to have been advised by Dr. Charles Brooks, who has been a truly wonderful mentor for my graduate career. His genuine interest in advancing understanding of the basic sciences, as well as his scientific vision, has always been a source of inspiration for me. The instrumental guidance he provided me with has shaped the way I approach science, and I will be grateful for all that he taught me in many years to come.

I would also like to acknowledge my committee members, Drs. Barry Grant, Kerby Shedden, Janet Smith and Yang Zhang, for all the insightful suggestions and feedback they gave that helped improve the quality of my work. In addition, I would like to thank Dr. Georgios Skiniotis, who is not on my thesis committee, but from whom I have learned a lot about cryo-electron microscopy.

I am also thankful to all current and past members of the Brooks lab. Working in such a diverse and vibrant environment has given me new perspectives and has complemented my Ph.D. training in many ways. In addition, I would like to particularly thank David Braun, who manages the computer resources in the lab, and Kate Dyki, our admin assistant, for providing full support throughout my graduate studies.

My gratitude also extends to all the friendly faculty and staff in the Bioinformatics program. Dr. Margit Burmeister was the one that recruited me to this wonderful program.

Dr. Daniel Burns was always approachable for academic advice. Julia Eussen, our graduate coordinator, was very patient with my endless questions.

I want to express my heart-felt gratitude to my parents, Jingxuan Cheng and Liqun Zhang. I would simply not be where I am now without their unconditional love for me and their self-sacrifices to support me to pursue my dreams. Special thanks also goes to my dear husband, Chaolong Wang, whose continuous support and encouragement gave me the strength to overcome every hurdle. Thank you for being in my life. Finally, to my beloved daughter, Zixuan Ann Wang, for coming to this world 10 days after my thesis defense and for bringing so much joy in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Viral capsids exhibit elaborate and symmetrical architectures of defined sizes and remarkable mechanical properties not seen with cellular macromolecular complexes. The limited coding capacity of viral genome necessitates economization upon one or a few identical gene products known as capsid proteins for shell assembly. The functional uniqueness of this class of proteins prompts questions on key structural features critically important for their higher order organization. In this thesis, I develop the statistical framework and computational tools to pinpoint the structural characteristics of viral capsid proteins exclusive to the virosphere by testing a series of hypotheses, providing understanding of the physical principles governing molecular self-association that can inform rational design of nanomaterials and therapeutics. In the first chapter, I compare the folded topology of capsid proteins with those of generic proteins, and establish that capsid proteins are segregated in structural fold space, highlighting the geometric constraints of these building blocks for tiling into a closed shell. Second, I develop a software program, PCalign, for quantifying the physicochemical similarity between protein-protein interfaces. This tool overcomes the major limitation of current methods by using a reduced representation of structural information, greatly expanding the structural interface space that can be investigated through inclusion of large macromolecular assemblies that are often not amenable to high resolution experimental techniques. As an application of this method, I propose a computational framework for template-based protein inhibitor design, leading to the prediction of putative binders for a therapeutic target, the influenza hemagglutinin. In silico evaluations of these candidate drugs parallel those of known protein binders, offering great promise in expanding therapeutic options in the clinic. Lastly, I examine protein-protein interfaces using PCalign, and find strong statistical evidence for the disconnectivity between capsid proteins and cellular proteins in structural interface space. I thus conclude that the basic shape and the sticky edges of these Lego pieces act concertedly to create the sophisticated shell architecture. In summary, the novel tools contributed by this dissertation work lead to delineation of structural features of viral capsid proteins that make them functionally unique, providing an understanding that will serve as the basis for prediction and design.

CHAPTER I

Introduction

1.1. General background:

Viruses, a frequently overlooked domain of pseudo-living organisms, are gaining attention in the scientific community due to the unparalleled reach and diversity of these systems, with an abundance that is estimated to outnumber living cells by an order of magnitude [1]. Since the 1980s there has been a paradigm shift from the traditional view of viruses as a long-standing foe of human health [2], to their being a rich source of information on biodiversity [3], on their role in shaping evolution through genetic exchanges with hosts [4], and on the insights they provide into the design principle of nanoscale biological containers [2]. Our current understanding of the virosphere is still relatively limited, and growing efforts are now channeled into studying this important section that is an indispensable and integral part of the biosphere.

Traditionally, viruses have been classified based on a variety of phenotypical features, such as their host organisms (animal viruses, plant viruses, etc.), their overall morphology (helical, icosahedral or complex), and the types of genetic materials they carry (dsDNA viruses, ssDNA viruses, dsRNA viruses, +ssRNA viruses, -ssRNA viruses, ssRNA-RT viruses, and dsDNA-RT viruses). A major obstacle in deriving a systematic taxonomy for viruses, like with the domains of cellular life, is the lack of a unifying genetic marker, compounded with the fact that viruses frequently undergo lateral gene transfer as opposed to vertical transmission. With recent advances in experimental techniques that have determined the structures of viral capsids, there is accumulated evidence that viruses previously thought to be unrelated, including those infecting different domains of hosts [5-7], can share remarkable structural and architectural similarity in their viral capsid proteins, providing support for a newly formulated hypothesis that attempts to group together viruses by the structural similarity of the viral coat proteins to define viral lineage [8]. This proposal appears reasonable, given that viral capsid proteins are arguably the single hallmark feature present in all viruses, in the same way that ribosomes are signature of the cellular empire.

This thesis work is centered on viral capsid proteins, which are a functionally unique class of proteins in the virosphere. They play a critical role in the life cycle of viruses, providing a protective coat that encapsulates the viral genome to be delivered to host cells they infect. In carrying out this central function, capsid proteins first have to self-assemble into a closed shell of a defined size inside the highly crowded cellular environment at the stage of replication, undergo a series of maturation steps to stabilize the shell in the case of some viruses, and later dissociate from one another in the newly infected host cells to release the genetic materials, manifesting remarkable dynamic and mechanical properties [9]. Other than the central function of protecting the genetic material, there are other more refined roles of capsid proteins that are differentiated among different viruses, depending on their detailed mechanism of infection. For instance, enveloped viruses, which are membrane-covered viruses, typically have glycoproteins (here broadly considered as coat proteins or capsid proteins) that specialize in inducing membrane fusion with the endosome upon endocytosis, in addition to their structural role.

Most of the currently solved structures of viral capsids are icosahedral, and we have focused our studies on these available structural data, which are collectively available in the database VIPERdb [10]. An icosahedron contains 20 triangular faces, with 2-, 3- and 5-fold symmetry, and consists of 60 copies of an icosahedral asymmetric unit (IAU). Given the limited coding capacity of viral genomes, viruses economize upon a single type or a few types of capsid proteins that tile on an icosahedron lattice. Within each IAU, there are T-number of capsid proteins, where the T-number is the triangulation number that takes discrete values given by $T=h^2+hk+k^2$, h and k being non-negative integers. An icosahedron has twelve vertices, and thus an icosahedral virus consists of 12 pentamers placed at the vertices and 10(T-1) hexamers tiled on the flat surfaces. For T=1 viruses, all capsid proteins are placed in an identical environment. For T>1 viruses, the sequence-wise identical capsid proteins are placed in different chemical environments, and interact with one another in slightly different fashions. This construction principle, termed ‘quasi-equivalence’ by Caspar and Klug [11], allows large viruses with more than 60 protein subunits to form, while still obeying icosahedral symmetry. In this case, quasi-symmetry axes are introduced, which do not necessarily coincide with the icosahedral symmetry axes, depending on the classification of different T-numbers as discussed in detail in [12].

While most icosahedral viruses have symmetries that are explained by the quasi-equivalence theory, a few exceptions exist. The best known example of deviation from the quasi-equivalence theory is the polyoma viruses. Instead of having 12 pentamers and 10(T-1) hexamers, polyoma viruses consist of only pentamers, which occupy positions on an icosahedral lattice that correspond to T=7 symmetry [13] and classified as such. The total number of capsid protein subunits is thus 360, which implies a forbidden T-number based on the formula, contrary to the expected 420 for a T=7 virus. Another case that disobeys the quasi-equivalence rule is L-A viruses, which are comprised of 120 protein subunits that again points to a forbidden T-number of 2. It is therefore commonly viewed as having a symmetry of T=1, except with dimers making up the IAU [14]. Finally, different capsid protein subunits in picornaviruses have different peptide sequences, yet all adopt the same structural fold, which can also be regarded as forming pseudo-quasi-equivalent interfaces [15]. Despite these aberrations, the quasi-equivalence theory remains an important cornerstone that has wide applicability in describing the architecture of most icosahedral viruses, and the T-number is generally regarded as a useful metric in quantifying the size of viral capsids.

1.2. Aim of the study and motivation:

One observation about viral capsids is of interest to us; although symmetric macromolecular assemblies are not lacking in the cellular domains of life, exemplified by ferritin cages and the chaperone GroEL/ES complexes, they are not found to be on such a large scale as viral capsids. In other words, viral capsid proteins represent a functionally unique class of proteins dedicated to make large cages for the protection and transportation of other macromolecules. Given that function follows form, we ask the question of what structural features of viral capsid proteins are key to determining their higher order organization, with the goal of understanding the fundamental physical principles governing assembly of biological containers of defined sizes.

Prior to our work, an important finding about the structure-function relationship of viral capsid proteins has established the role of topological constraints- specifically the trapezoidal shape adopted by most viral capsid proteins- in tiling into a closed spherical shell with no holes, no overlaps and no gross structural variability [16], which are essential properties for capsid shells designed to offer protection to the viral genome with maximal genetic economy. With the aim to gain a deeper understanding to fully address the question raised above, I set out to investigate in this thesis work how various structural features of viral capsid proteins compare with generic, non-capsid proteins. The answer to our question has three-fold important implications; first is to provide guidelines for the design of nano-packaging tools in various biomedical applications and materials science, second is to inform the rational design of antiviral drugs that minimize toxic effects, and lastly to enrich our knowledge of protein-protein interactions in general by mining through the myriad of structural data for viral capsids. These three aspects are presented as follows.

The ability of viral capsid proteins to self-assemble into cages of defined sizes offers great promise in a range of biomedical applications, including vaccine development, gene therapy, bio-imaging and drug delivery. In fulfilling these functionalities, viral capsid proteins are typically expressed in the absence of the infectious genomic material they encapsulate, with the exception of the use of viruses for transfection in gene therapy, and these empty viral shells are often called virus-like particles (VLPs) without virulent activity. Since as early as in the 1990s, immunogenic peptides have been genetically fused with viral capsid proteins to be presented on the surface of the assembled capsids to enhance recognition by antibodies to elicit immune response [17-44], with some of the

VLP-based vaccines making it to the clinical trial stage. The use of viruses as vectors of gene therapy also has a long history, and has been known for their efficiency for the reason that viruses are naturally designed to transfect living cells with their own genetic material, via sophisticated mechanisms selected by evolution. These vectors are generated as recombinant viral genomes packaged by viral capsid proteins, incorporating the therapeutic gene of interest, allowing self-replication in some cases (especially in the treatment of cancer) while prohibiting replication in others for a ‘one-hit’ transient expression of the therapeutic gene using replication-defective viruses [45]. Different types of viral vectors including retroviruses, adenoviruses, adeno-associated viruses and herpes simplex viruses have been developed [46], with notable successes in the treatment for inherited monogenic diseases [47-51]. In the area of bio-imaging, VLPs can either be convalently conjugated to dye molecules on the viral surface to take advantage of the large display area for maximized signal, as well as the precise positioning to prevent the dye molecules from fluorescent quenching, or they can encapsulate fluorescent cores as cargos, leaving the outer surface for modifications to achieve specific targeting [52, 53]. These techniques have been widely applied in various biomedical studies, including but not limited to imaging of tumor angiogenesis in vivo with fluorescent cowpea mosaic virus sensors [54-56], packaging quantum dots and infrared chromophores into simian viruses and brome mosaic viruses for in vitro imaging [57, 58], and using cowpea mosaic viruses, cowpea chlorotic viruses and MS2 viruses as carriers of MRI contrast reagents for diagnostic and therapeutic decisions [59-62]. Lastly, formulations of small drug molecules delivered by viral containers have also been developed for treatment of cancer

[63-68] and bacterial infections [69], greatly prolonging the activity of the drug molecules which otherwise undergo fast clearance.

In materials science applications, viral capsids are equally popular as bio-templates given their remarkable mechanical properties and diversity in sizes and shapes. Phage display libraries, for one, have been used for identifying peptides with specific binding to materials, which are in turn engineered onto viral capsid proteins for directing nucleation of inorganic nanocrystals of different sizes, composition and morphology [70-74]. The interior of VLPs has been explored to confine nanomaterials synthesis within, with examples ranging from mineralization of anionic polyoxometalate salts [75], to formation of nanowires in helical viruses [76, 77], to creation of light harvesting systems by chemically attaching chromophores to viral coat proteins [78, 79]. Selective engineering of exposed residues on the outer surface of viral capsid proteins also allows spatially controlled crosslinking patterns in various nanodevices such as sensors and electronic circuits [51, 80-82]. Lastly, altering assembled architectures by tuning inter-subunit interactions of viral capsid proteins [83] enables some degree of control over magnetic properties of the magnetic materials synthesized within [84].

The need for precise control in all of the above design applications is highly important; on the exterior surface, functional moieties are arranged with specific geometries (such as defined spacing) for optimally targeting specific sites of interest. At the interface between protein subunits of the capsid shell, the slightest modifications can imply altered stability and overall architecture. For materials encapsulated in the viral interior, the shell has to assume the right size to tailor to different cargos or has to be decorated with modules internally to direct the assembly. Understanding the physical principles that govern viral shell assembly is essential to achieve predictable results when such new functionalities are imparted onto the viral templates. What specific structure is required for the building block to satisfy geometric constraints in forming a closed shell? In what ways do these Lego pieces associate with one another? We can glean some preliminary insights into these unanswered questions by comparing the structural features of capsid proteins that form large protein cages and the non-capsid proteins that do not form large protein cages, and delineate characteristics that differentiate the function of capsid proteins.

A second reason for us to study viral capsid proteins is the increasing recognition of their being attractive therapeutic targets in viral infections. Traditionally, efforts on developing antiviral drugs have largely focused on viral enzymes [85], which are involved in important steps in the life cycle of viruses. Examples of these targets include viral DNA polymerases and reverse transcriptases, which are essential for replication of the viral genome, with known drugs such as acyclovir, brivudin, zidovudine, nevirapine and other nucleoside analogues [86, 87]; viral proteases, which process newly synthesized viral proteins for maturation, targeted by saquinavir, ritonavir and other peptidomimetic substrate analogues [88]; neuraminidases, the glycoproteins found on influenza viruses that allow newly formed virions to escape the infected host cell, inhibited by drugs such as zanamivir and oseltamivir [89-93]. While efficacious, these drugs may raise selectivity concerns due to the potential functional overlap between viral enzymes and some cellular enzymes. The emergence of techniques that allow monitoring of the viral capsid assembly [94, 95] has opened up avenues to investigate drugs targeting viral capsid proteins, which are functionally unique to viruses and thus may serve as a more specific drug target. A few pioneering studies have led to the discovery of molecules/compounds that interfere with capsid assembly or capsid uncoating. Pleconaril, for instance, is known to bind to a hydrophobic pocket on the capsid protein VP1 of enteroviruses, which rigidifies the capsid shell and prevents the capsid proteins from undergoing conformational changes necessary for disassembly [96-98]. A peptide, CA-I, inserts in between the dimer interface of HIV capsid proteins, while another allosteric small molecule inhibitor of HIV, CAP-1, alters the interface geometry to disrupt the viral capsid assembly [99-103]. Given the different structural aspects one can target on viral capsid proteins, be it the structural core or the protein-protein interface, we should be cautious about drug selectivity. It is thus helpful to evaluate which structural features of viral capsid proteins are exclusive in the virosphere and therefore pathogen-specific, so as to rationalize antiviral drug design that minimally interferes with normal cellular activities.

Lastly, viral capsid proteins provide a great example of protein plasticity in accommodating varied modes of protein-protein interactions. As introduced earlier, quasi-equivalence of inter-subunit interfaces in viral capsids provides a general framework for large icosahedral viruses to maximize their genetic economy. Different modes of interaction between genetically identical capsid proteins can be achieved via an array of mechanisms; in small plant viruses this is frequently in the form of the switch between order/disorder of a terminal arm of the capsid proteins, or RNA segments, or a combination of both [104-106], and in larger viruses often helper proteins are required to control the size of the assembled complex, such as the tape protein P30 that defines the vertex-to-vertex distance between two pentamers in PRD1 viruses [107]. Structural data of viral capsids thus provide a wealth of information on fine-tuning protein-protein interactions with subtle conformational switching among the constituent monomers, which can lend insights into fundamental principles of molecular recognition. Before we apply such knowledge gained to for instance improve scoring functions for protein-protein docking studies, we need to establish its generalizability, by assessing whether the inter-subunit interfaces in viral capsids span the structural interface space of all proteins including cellular ones.

1.3. Overview of thesis chapters:

Motivated by these open questions, I performed large scale comparative analysis between available structural models of all capsid proteins and non-capsid proteins, focusing on two primary structural features of viral capsid proteins, namely the folded topology and the protein-protein interfaces. Chapter II discusses my findings on the structural overlap in protein fold space between viral capsid proteins and cellular proteins, which informs the geometric constraints of individual building blocks for tiling into a closed shell. Chapter III describes a computational tool I developed, PCalign, to quantify physical and chemical similarities between a given pair of protein-protein interfaces. Because no current method prior to this work deals sufficiently with low-resolution structural data that is typical with viral capsids, this tool (PCalign) is specifically designed for assessing the degree to which capsid inter-subunit interfaces are representative of all interfacial patterns present in cellular macromolecular complexes, as explored in detail in Chapter V. In Chapter IV, I apply the same tool to designing protein inhibitors of the influenza virus hemagglutinin protein, and predict a number of new candidate binders that can further undergo experimental validation, illustrating the usefulness of the tool in setting up a general framework for search of novel protein-based drug leads.

In Chapter II, I mainly explore the question of whether the prevalent jellyroll fold of capsid proteins is exclusively found in the virosphere, to which end I examine representative capsid protein chains in the VIrus Particle ExploreR database (VIPERdb) [10] as well as a non-redundant subset of protein domains in the Structural Classification Of Proteins (SCOP) database [108], not including the viral capsid entries. In this work, I design a novel statistical framework for analyzing the degree of connectivity between two mutually exclusive sets of data, where the partition is made based on functional annotation while the connectivity is drawn based on structural similarity. This same framework is also applied in Chapter V when studying protein-protein interfaces.

Unlike the case of comparing individual protein structures, for which a state-of-the-art tool (TMalign) is available [109], a metric for quantifying similarity between protein-protein interfaces that accounts for patterns of interest to us is lacking. One major limitation with currently available implementations for methods that compare a given pair of protein-protein interfaces is that they all require atomic details of structural models as input to define what constitutes an interface. However, large macromolecular complexes such as viral capsids are often times not amenable to X-ray crystallography due to the challenge in crystal formation, and lower-resolution techniques such as cryo-Electron Microscopy (cryo-EM) serve as a complement that provides structural models typically with fewer details [110, 111]. In the VIPERdb database [10], for instance, there are 29 cryo-EM structural models (10 of which have homologues determined by X-ray crystallography) for icosahedral viral capsids in addition to 392 X-ray crystal structures, which is a small but non-negligible proportion. Chapter III of my thesis covers specifically a method that I developed, PCalign, to address this limitation, taking into account the extent to which the spatial and chemical arrangement of residues lining two protein-protein interfaces overlap with each other. This method facilitates comparison across structural models with different levels of detail by relying on a hierarchical definition of interfacial residues, using a distance criteria optimized by a correlation study that I carried out through mining the Protein Data Bank (PDB) [112].

Chapter IV of my thesis illustrates an application of the method PCalign other than the intended large scale comparative studies detailed in Chapter V, i.e. template-based protein inhibitor design for a given therapeutic target, which is the hemagglutinin of influenza viruses in this study. Motivated by recent works of David Baker and coworkers [113, 114], which used an *ab initio* approach to successfully design protein inhibitors for the hemagglutinin protein, I attempted to address the same problem by searching the existing library of protein-protein interfaces to identify putative binders for the viral protein, which interact with their respective native partners in similar fashions as the viral protein complexes with a known antibody, based on interface similarity recognized by the PCalign program. These putative binders, bearing a binding site that resembles the paratope, can then serve as a starting point for redesign for affinity for the target protein, using an array of computational modelling tools. Candidate protein inhibitors identified as such can then undergo further experimental validation to verify the feasibility of our design protocol.

Knowing that the geometric shape of capsid proteins plays a role in determining the assembled architecture, I was next interested in whether the glue that holds these building blocks together is also subject to evolutionary constraints. Chapter V addresses this question by testing the hypothesis that the protein-protein interfaces in viral capsids are representative of generic protein-protein interfaces using a similar statistical framework as with the folded topology. The comparative analysis of a non-redundant set of inter-subunit interfaces in viral capsids from VIPERdb and protein-protein interfaces in non-capsid protein complexes from the PDB allows us to assess if the overlap in structural space of protein-protein interfaces of the two sets is significantly small. Together with my earlier study of the folded topology of viral capsid proteins, we can delineate important structural features in viral capsid proteins in directing the final complex formation of nanoscale containers.

In summary, I have developed a general statistical framework for testing the uniqueness of structural features for a given functional class of proteins, as well as a computational tool, available as a software program, to quantify physicochemical similarity between protein-protein interfaces. These methodology developments led to novel biological insights into the design principles of large biological assemblies, with broad applications in biomedical and materials science.

CHAPTER II

Viral capsid proteins are segregated in structural fold space

2.1 Introduction:

Viral capsid proteins protect the viral genome by forming a closed protein shell around it. Most of currently found viral shells with known structure are spherical in shape and observe icosahedral symmetry [115]. Comprised of a large number of proteins, such large, symmetrical complexes assume a geometrically sophisticated architecture not seen in other biological assemblies. Here we make a distinction between protein cages in viral capsid shells that have sizes ranging from about 10 nm to about 90 nm in radius (Figure 2.1(A)), and other oligomeric containers of a much smaller scale, such as ferritins and chaperones. In the simplest form, 60 identical copies of an icosahedral asymmetric unit (IAU) are assembled with 5:3:2 symmetry, by positioning three IAUSs on each of the 20 triangular faces of the icosahedron [116]. The triangulation-number, or T-number, can be used to describe the number of proteins in each icosahedral asymmetric unit and therefore the size of the virus. Thus the number of capsid proteins in each shell is a multiple of 60, such as 180 proteins for a T=3 virus and 240 proteins for a T=4 virus. While T=1 viruses can place each protein in an identical environment, other viruses having multiple proteins per IAU achieve the symmetry by following the ‘quasi-equivalence’ principle proposed by Caspar and Klug [11]. Also worth noting is that large viruses, such as double-stranded

RNA (dsRNA) viruses, deviate from this principle, while preserving a rigid icosahedral symmetry nonetheless [116].

Geometry of the complex architecture aside, another striking feature of viral capsid proteins lies in the folded topology of the monomers, with the canonical jelly-roll β barrel appearing most prevalent (but not unique) as a core structural motif among capsid proteins that make up these viral shells of varying sizes [117]. Traditionally, this fold has also been termed as a wedge shape [118], an RNA virus capsid domain [119], a β-barrel [120], a β-sandwich [121], and an eight-stranded antiparallel β-barrel fold with a β-roll topology [122], all of which are consistent with the overall morphological characteristic of the fold (Figure 2.1(B)). Remarkable diversity in the loop regions connecting the β strands has been observed across different viruses, with variations in length and in inserted segments ranging from secondary structural elements to complete domains [123]. This signature fold of capsid proteins has been extensively studied [124, 125], and has also been compared with non-viral proteins in many separate works, most of which aimed to investigate the evolutionary relationship between viruses and their hosts. Other than the jelly-roll β barrel, there are also the Greek key β barrel with six strands [126], the helix bundle [127] and the immunoglobulin-like fold [128].

Given the unique geometry of the complex formed by viral capsid proteins, one interesting question arises as to whether the structural folds of capsid proteins that assemble into this distinct architecture are also unique to viruses. By comparing the structural topology of capsid proteins that form the icosahedral shells and generic proteins that interact to form other types of complexes, we can potentially establish a link between capsid fold and capsid architecture, or the lack thereof. The answer to this

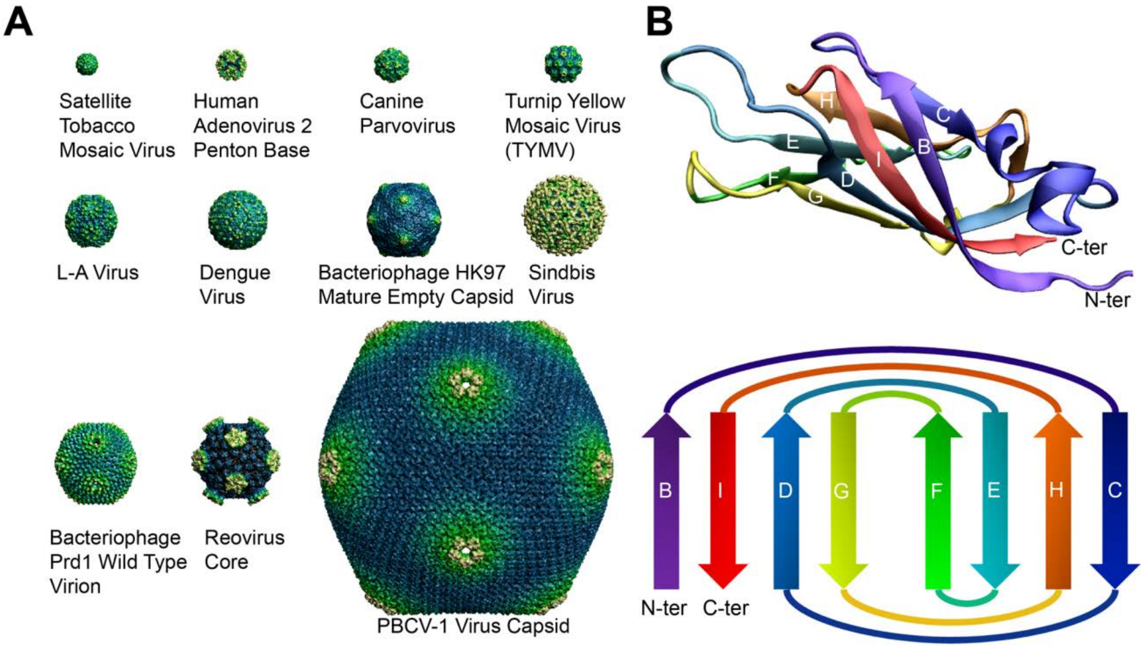
A
Satellite Tobacco Mosaic Virus
Human Adenovirus 2 Penton Base
Canine Parvovirus
Turnip Yellow Mosaic Virus (TYMV)
L-A Virus
Dengue Virus
Bacteriophage HK97 Mature Empty Capsid
Sindbis Virus
Bacteriophage Prd1 Wild Type Virion
Reovirus Core
PBCV-1 Virus Capsid
B
H
C
E
B
I
D
F
G
C-ter
N-ter
B
I
D
G
F
E
H
C
N-ter
C-ter

Figure 2.1. Capsid shells and the folded topology of a typical capsid protein. (A) Representative icosahedral viral capsid structures with varying sizes. The Satellite Tobacco Mosaic Virus which is a T=1 virus has a radius of 8.8 nm, and the Paramecium bursaria Chlorella *virus* 1 (*PBCV*-1) which is a pT=169 virus has a radius of 92.9 nm. Here pT stands for ‘pseudo T number’, which simply means the subunits are not chemically identical (the primary sequences are different). These protein shells are large in that they are assembled from tens of up to hundreds of protein monomers, and they are highly symmetrical. (B) The signature jelly-roll of viral capsid proteins, with 8 β-strands forming two antiparallel sheets, exemplified by a satellite tobacco mosaic virus protein subunit here (PDB code: 1a34). The wedge or trapezoidal shape of this particular fold immediately reveals six flat surfaces for monomer-monomer interaction; the sides, the two loop ends and the top and the bottom. The prevalence of the jelly-roll fold among capsid proteins might be related to their relative ease for tiling.

question can lend novel insights to protein-protein interactions, in terms of how folds of protein monomers, as opposed to their surface chemistry, might be related to the assembled multimer complex architecture. Furthermore, the ability of many viral capsid proteins to self-assemble spontaneously makes them an attractive platform for synthetic manipulation across the fields of biomedical applications and nanosciences [2]. Understanding how much influence viral capsid folds place on the assembled architecture is likely to provide guiding principles in the design of drug delivery systems and nanomaterials.

In this work, we present, to the best of our knowledge, the first attempt to examine whether the structural folds of viral capsid proteins set them apart from generic proteins, and with how much statistical significance. We recognize that a general assumption is that any class of proteins with a unique function is expected to be found in exclusive folds, which may or may not hold, given that folded topology is a coarse description of structural characteristics. Thus in addition to testing our hypothesis in the specific case of viral capsid proteins, we perform similar analysis for a few representative classes of proteins with diverse functions. At a finer level of granularity, i.e., the superfamily level, Abroi and Gough also surveyed the classification of all viral proteins and the other superkingdoms to study their genetic interactions in evolutionary history [3]. We distinguish our work by restricting our analysis to viral capsid proteins, which are functionally unique in viruses, in order to establish the link between topology of the building block and the assembled complex architecture. In another related work, Janin and coworkers provided an extensive analysis of physicochemical characteristics of protein-protein interfaces in icosahedral viruses, and compared them with generic protein-protein interfaces [129]. Rather than adopting the same approach of enumerating what's similar and what's different between the two classes, we will employ a direct comparison metric to evaluate whether there is significant statistical evidence supporting our conjecture that viral capsid proteins are structurally unique.

2.2 Materials and methods:

To test our hypothesis that viral capsid folds are not commonly found in generic proteins, we proceed to evaluate if the proportion of non-viral capsid proteins that share similar structural folds with viral capsid proteins is significantly small (Figure 2.2), based on a well-defined quantitative measure.

2.2.1 Comparison metric

We chose the Template Modeling-score (TM-score) [109] as our structural comparison metric, for the following reasons. This structure-alignment-based scoring function using the fr-TM-align algorithm [130] is very fast to compute and suits our large-scale comparison; it is normalized, or protein size independent, making the comparison between pairs of domains with complex topology and pairs with simpler ones fair; it has been established in large scale benchmark studies that most of the pairs of proteins with a TM-score of more than 0.5 have the same fold classification, and most of those with a TM-score of less than 0.5 are in different fold classes [131]. In addition, a TM-score of 0.4 has also been extensively used as a criterion to decide if a pair of structures are similar or not [132]. Given that many proteins within the same SCOP fold can have a TM-score of 0.4 and higher, we chose the TM-score of 0.4 as the threshold to validate our hypothesis.

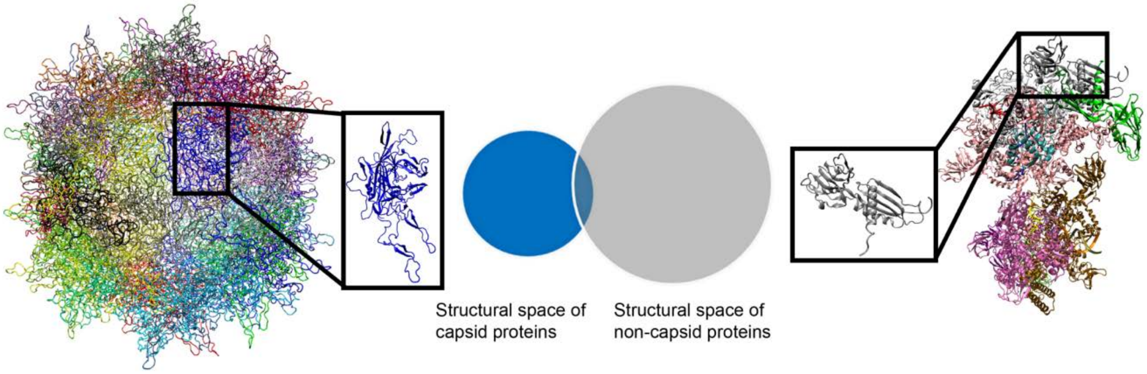
Structural space of capsid proteins
Structural space of non-capsid proteins

Figure 2.2. Comparison in structural fold space of capsid proteins and non-capsid ones. Capsid proteins form large, highly symmetric protein shells (left, PDB code: 3kic), while generic proteins form other types of complexes (right), exemplified here by an RNA polymerase elongation complex (PDB code: 2o5i). Overlap between the structural space of viral capsid proteins and that of generic proteins signifies the set of non-capsid 'relatives' of capsid proteins. Figure is for illustration purposes and not drawn to scale.

Briefly, the structural alignment score is defined as

$$\text{TM-score} = \boldsymbol{Max}\left[\frac{\mathbf{1}}{\boldsymbol{L_N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{L_T}}\frac{\mathbf{1}}{\mathbf{1}+\left(\frac{\boldsymbol{d_i}}{\boldsymbol{d_0}}\right)^{\mathbf{2}}}\right], \tag{2.1}$$

where L_N and L_T are the lengths of the two peptides being compared, d_i is the distance between the Cα atoms of the structurally equivalent residues, and d_0 is a normalization score to make the alignment length-independent. The term *Max* stands for an optimal superimposition between the two structures to minimize distances between structurally-equivalent residues. We define structural distance between a pair of proteins by (1−TM-score), which ranges from zero to one.

2.2.2 Data collection

In our work, we included all of capsid, nucleocapsid and envelope proteins for analysis, which we collectively call capsid proteins, because of their common structural role in forming the viral shell despite differentiated functions in a few cases. We collected the viral capsid protein set from the VIrus Particle ExploreR (VIPERdb) [10], which is a database of icosahedral virus capsid structures, with 319 entries in total. Altogether 1174 protein chains having at least 80 residues were extracted from these entries, as short peptides are known to assume very simple topologies. These 1174 were further cut into domains; while 452 proteins have domain annotations in SCOP, 637 proteins have homologues (sharing a sequence identity of at least 40%) that are well-annotated by SCOP. The remaining 85 were examined visually and dissected into individual domains. Lastly, the non-compact domains (extended structure with little secondary structure content) are removed, leaving 1447 domains in total.

We used the non-redundant set of 10569 proteins covering 1195 folds from the database Structural Classification Of Proteins (SCOP) 1.75 [133] filtered at 40% sequence identity, available from the ASTRAL compendium [134], to constitute our total protein set. This set was further reduced to 8921 proteins covering 1047 folds after removal of short peptides with fewer than 80 residues. The viral capsid protein set was then subtracted from the total protein set to yield the non-capsid protein set. In addition, 24 capsid proteins in the total protein set that were originally not deposited in VIPERdb were added to the capsid set and removed from the non-capsid set (Appendix A, Table A.1). A sequence filter of 40% identity was then applied to the domains of the capsid set, which resulted in 151 domains that are sequence-wise non-redundant.

As viruses across the same family are known to share limited sequence identity despite remarkable structural resemblance, a further structural filter was applied to the capsid set of 151 domains by clustering analysis. We performed hierarchical clustering via the average linkage method, and selected the cluster medoids of the resulting N clusters as our structurally non-redundant capsid set. Optimal partitioning of the data from hierarchical clustering was obtained by choosing the minimal number of clusters such that all intra-cluster distances are less than 0.6, using our structural distance measure. This criterion is based on the rationale that we would like to sort out the most representative capsid structures, without their repeating one another resulting in unfair comparison with the permutation test that we will describe shortly.

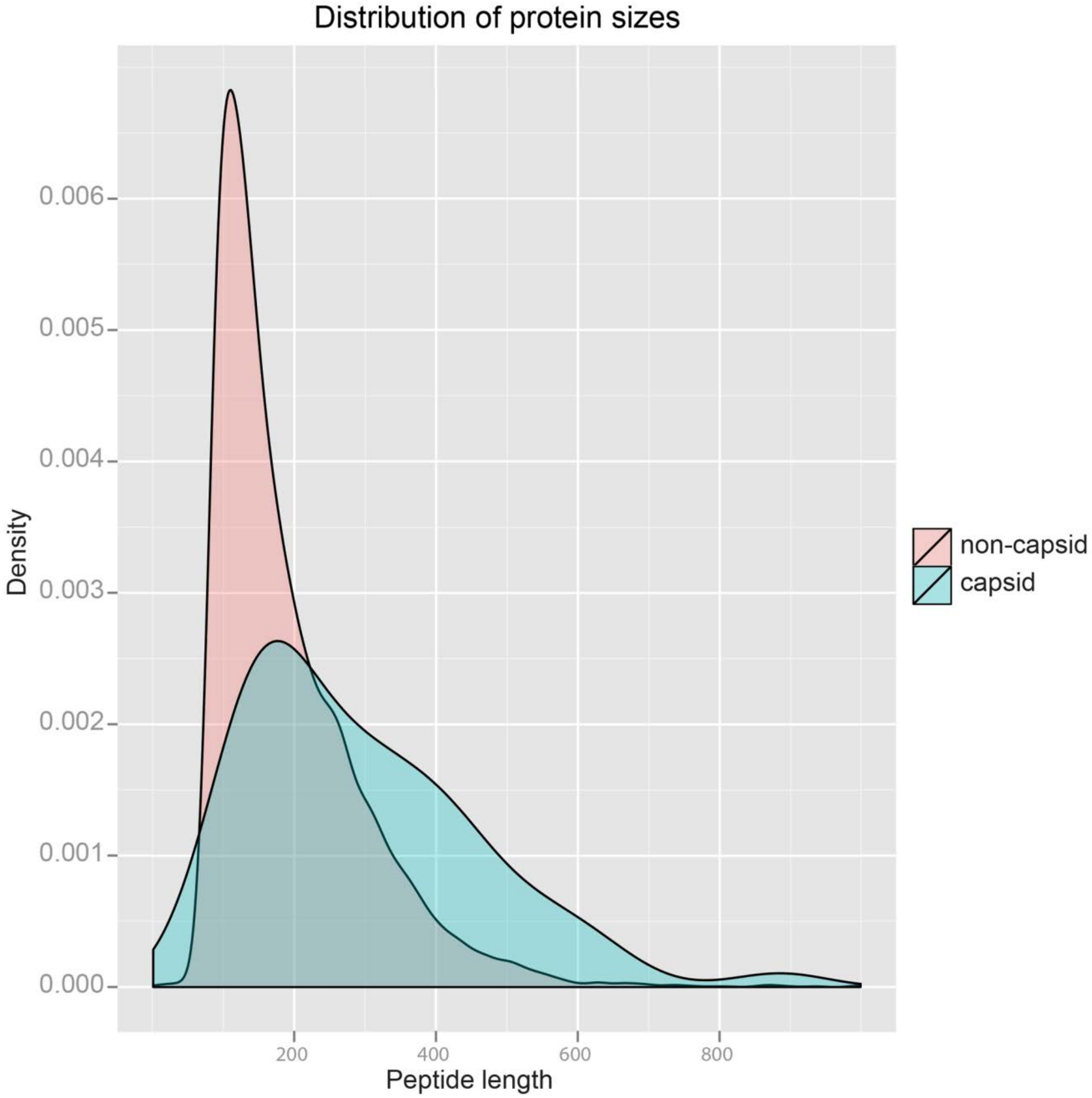

Figure 2.3. Domain size distribution. Shown in pink is the density distribution of the lengths of non-capsid proteins, and that of capsid proteins is shown in blue. Viral capsid proteins appear to have overall larger domains compared to their cellular counterparts, with a few exceptionally complex domains having more than 600 residues. 600 was later used as a size cutoff in order to examine the two sets that are of comparable sizes.

A preliminary survey of the two sets revealed differences in the sizes of domains. As shown in Figure 2.3, a typical capsid domain (blue) has approximately 180 residues, compared to about 150 residues for a typical non-capsid domain (pink). This size comparison is purely based on existing structural data of viral capsid proteins, but we do see a larger proportion of complex topologies in certain capsid domains, as opposed to the under-representation of longer folds in generic proteins. In order to preclude the possibility of concluding that capsid and non-capsid proteins have different folds that are in fact largely a result of the difference in length, we performed an additional separate analysis by removing domains having longer than 600 residues in both datasets.

2.2.3 Shared folds as the test statistic

After obtaining the non-redundant viral capsid set and the non-capsid set, we quantify the extent to which the structural space of the non-capsid set overlaps with that of the capsid set in the following manner. We performed an all-against-all structural comparison between the non-capsid set and the capsid set. For each member in the non-capsid set, we select its nearest neighbor in the capsid set, and use the distance between the two to represent how far structurally this particular non-capsid protein is to viral capsid proteins. With the structural distances between all non-capsid proteins and their nearest neighbors in capsids in hand, we then filter the non-capsid set by retaining only proteins that are less than 0.6 away from capsid proteins. We thus obtain the final distribution of distances between viral capsid proteins and those non-capsid proteins that structurally resemble capsid proteins. Among these 'relatives' of viral capsid proteins, we count the number of folds covered by them, following the fold classification in SCOP. This defines our test statistic, which we term as 'shared folds' in the rest of the chapter.

2.2.4 Statistical significance of the test statistic

To estimate the statistical significance of the number of shared folds between capsids and non-capsid proteins, we calculated the probability of observing at most the same number of shared folds by random chances by running a permutation test on the total protein set. The total set of proteins was randomly partitioned into set A and set B, with set A consisting of an equal number of proteins as that in the capsid set, and set B being their complement in the total set. The same procedure as described above was carried out to obtain the number of shared folds between this particular set A and their non-self counterparts. To avoid finding 'relatives' in set B that are evolutionarily closely related to (i.e. belonging to the same family) the proteins in set A, we further excluded 'self folds' from the shared folds found, as an approximation to, or a lower bound of, folds shared with non-self proteins. Here 'self fold' is defined as the fold annotation by SCOP of a particular structural analogue found in the large protein set that is already covered by any protein in the small set of proteins. Altogether 10,000 independent permutations were done to give rise to the estimated distribution of shared folds, based on which the p-value of our test statistic can be evaluated.

2.2.5 Cross-checking with other functional classes of proteins

To examine if unique function generally implies unique folds, we chose a few functional classes of proteins to perform the same analysis described above for capsid proteins. Seven classes were chosen, namely kinases, globins, dehydrogenases, DNA/RNA polymerases, chaperones, antigens and muscle proteins, with functions ranging from catalysis, to transport to signal transduction. The total protein set which is filtered at 40% sequence identity level was partitioned into two sets based on SCOP annotations at the domain level; one being the functional class and the other being the complementary set, and the statistical significance of shared folds is again estimated by permutation tests.

2.3 Results:

2.3.1 Representative folds adopted by viral capsid proteins

We found 56 clusters for the viral capsid set, using the criterion described in the Materials and Methods section. These clusters are fairly compact, with all members within each cluster being less than 0.6 apart from one another. Furthermore, the clusters are maximally separated, with only 26 pairs of proteins (0.24%) from two different clusters being closer than 0.4. In Figure 2.4, we show the statistics demonstrating a good separation between clusters that are reasonably homogeneous. The resulting 56 cluster medoids thus represent the distinct domain architecture adopted by capsid proteins.

Figure 2.5 illustrates these 56 clusters with all members in each cluster superimposed on one another. The alignment shows high structural similarity across the same cluster, while different clusters display mostly different folding topologies, in agreement with our quantitative assessment. There are a fairly large number of singlet clusters that are unlike one another, mostly because the structural data for these few viral families are lacking. The few most populated clusters correspond to the canonical jelly-roll fold, with variations in the terminal ends.

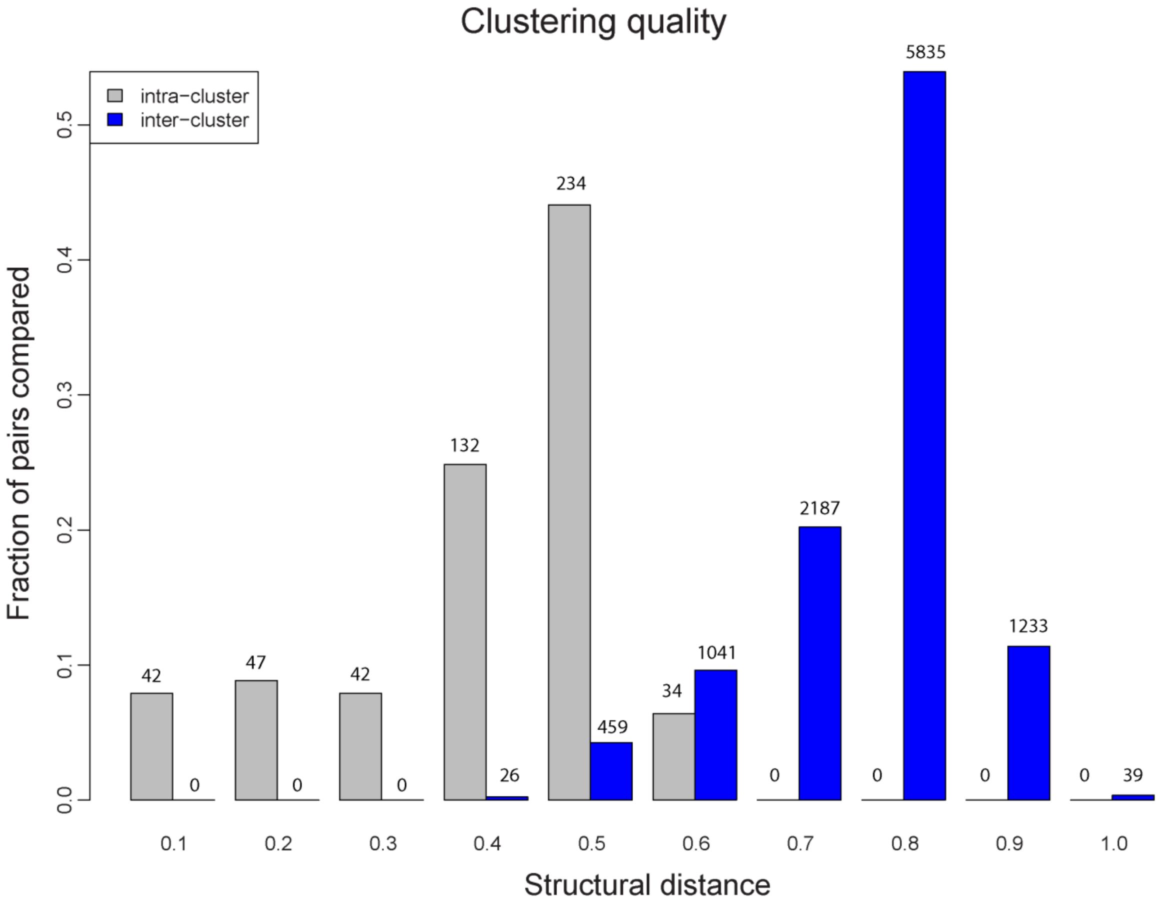

Figure 2.4. Clustering to find representative capsid folds. Shown here are all pairwise distances between members from the same cluster (grey) and between members from different clusters (blue). The numbers on top of each bar indicate the number of pairs that fall into that bin. Partitioning was chosen such that each cluster is maximally homogeneous, with no members within the same cluster being farther than 0.6 apart.

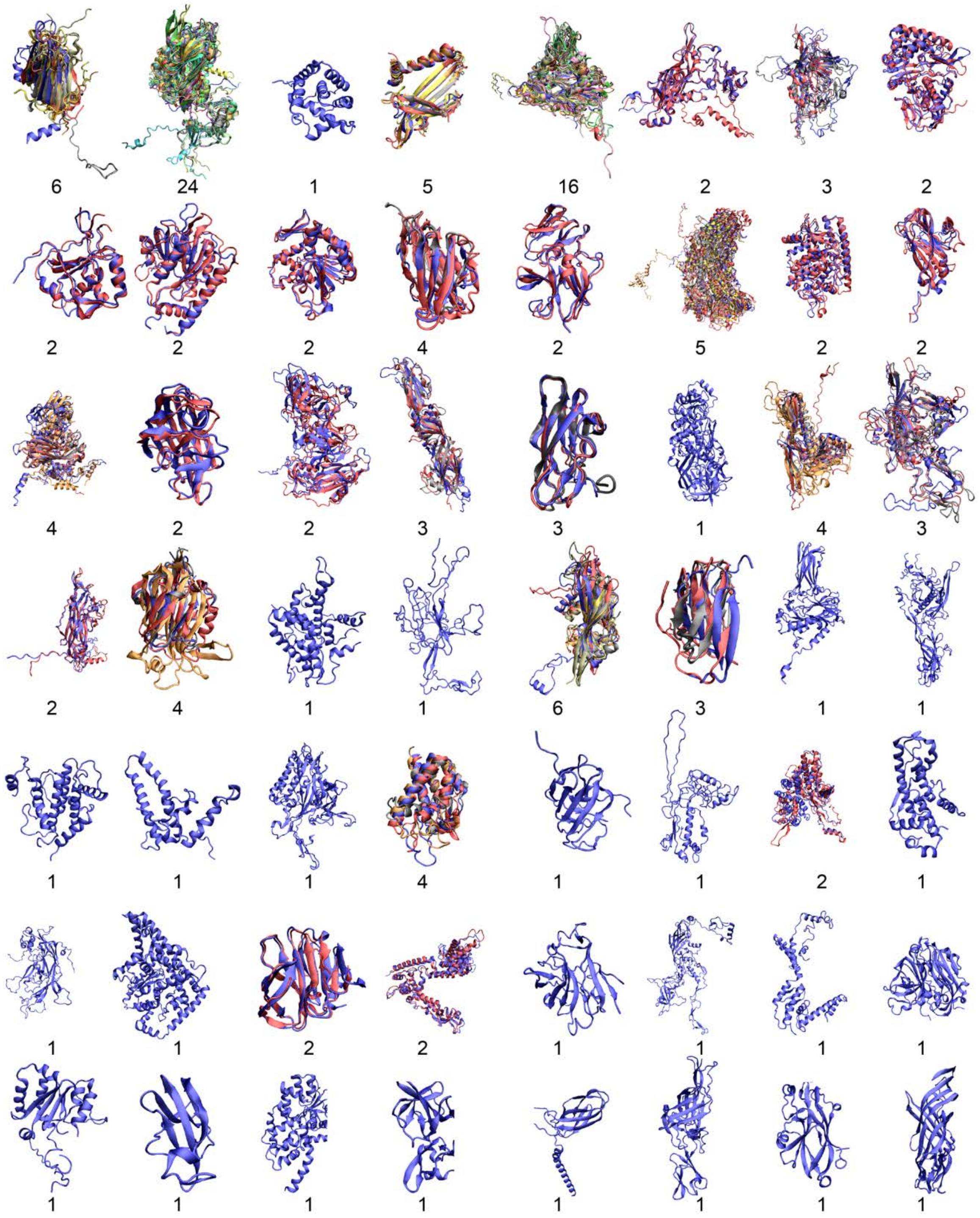

Figure 2.5. The 56 representative capsid folds. Domains within one cluster are superimposed on one another to show good structural alignment, with number of members in each cluster indicated. The prevalence of singlet clusters reflects the scarcity of structural data for many viral families.

2.3.2 Viral capsid proteins are segregated in structural fold space from generic proteins

By comparing the viral capsid set and the non-capsid set, we found altogether 2078 generic proteins sharing similar topology with viral capsid proteins, based on a distance cutoff of 0.6. These 2078 proteins cover 210 folds in total. If we disregard marginally similar capsid-like proteins by looking at those within a distance 0.5 of capsid proteins only, we find altogether 600 proteins covering 21 folds (Table 2.1). A further inspection of the distribution of shared folds for randomly sampled sets of 56 proteins and their non-self counterparts immediately reveals that viral capsid proteins are structurally separated from generic proteins. Referring to Figure 2.6, the cumulative fraction of non-self proteins across the entire structural distance spectrum from viral capsid proteins is clearly shifted to the right compared to those of the 10,000 permutation tests. Through this plot, we expect to arrive at the answer that capsid proteins are different from generic proteins regardless of the distance cutoff used in defining similar folds.

2.3.3 Estimation of statistical significance

The distribution of shared folds, estimated from the 10,000 permutation tests, is plotted in Figure 2.7. The number of capsid-like folds shared by non-capsid proteins hence lies on the extreme left tail of the distribution, demonstrating that viral capsid folds are far less populated in structural fold space compared to generic proteins (Figure 2.7). The one-tailed p-value of our test statistic is less than 0.0001, and we thus conclude that there is significant statistical evidence against the null hypothesis that viral capsid folds

Table 2.1. The 21 folds covered by structural relatives of capsid proteins. 14 out of these 21 folds are either greek-key or jelly-roll (the latter fold being a specific variation of the former). Remarkably, 17 folds are specific to non-capsid proteins, and are only marginally similar to capsid proteins in structure.

fold (as in SCOP)	**name of fold**	**description of fold**	**whether contains capsid proteins**	**example of non-capsid relatives**	**SCOP ID of example**
b.1	Immunoglobulin-like beta-sandwich	*sandwich; 7 strands in 2 sheets; greek-key. some members of the fold have additional strands*	Yes	Titin, I27	d1tiua_
b.2	Common fold of diphtheria toxin/transcription factors/cytochrome f	*sandwich; 9 strands in 2 sheet; greek-key; subclass of immunoglobin-like fold*	No	Runt-related transcription factor 1	d1eaqa_
b.6	Cupredoxin-like	*sandwich; 7 strands in 2 sheets, greek-key* *variations: some members have additional 1-2 strands*	No	Auracyanin	d1qhqa_
b.7	C2 domain-like	*sandwich; 8 strands in 2 sheets; greek-key*	No	Chaperone protein Caf1M	d1p5va2
b.14	Calpain large subunit, middle domain (domain III)	*sandwich; 8 strands in 2 sheets; jelly-roll*	No	M-Calpain	d1df0a2
b.18	Galactose-binding domain-like	*sandwich; 9 strands in 2 sheets; jelly-roll*	No	Xyn10B carbohydrate-binding module	d1h6ya_
b.22	TNF-like	*sandwich, 10 strands in 2 sheets; jelly-roll*	No	Tumor necrosis factor superfamily member 4	d2hewf1
b.23	CUB-like	*sandwich, 10 strands in 2 sheets; jelly-roll*	No	Acidic seminal fluid protein (spermadhesin)	d1sfpa_
b.29	Concanavalin A-like lectins/glucanases	*sandwich; 12-14 strands in 2 sheets; complex topology*	Yes	Sugar binding protein	d1is3a_
b.47	Trypsin-like serine proteases	*barrel, closed; n=6, S=8; greek-key* *duplication: consists of two domains of the same fold*	Yes	human alpha-thrombin	d1h8d.1
b.71	Glycosyl hydrolase domain	*folded sheet; greek-key*	No	alpha-galactosidase	d1uasa1
b.82	Double-stranded beta-helix	*one turn of helix is made by two pairs of antiparallel strands linked with short turns has appearance of a sandwich of distinct architecture and jelly-roll topology*	No	transcriptional regulator, HTH_3 family	d1y9qa2
b.121	Nucleoplasmin-like/VP (viral coat and capsid proteins)	*sandwich; 8 strands in 2 sheets; jelly-roll; some members can have additional 1-2 strands characteristic interaction between the domains of this fold allows the formation of five-fold and pseudo six-fold assemblies*	Yes	Nucleoplasmin-like protein (histone chaperone)	d1nlqa_
b.132	Supernatant protein factor (SPF), C-terminal domain	*sandwich; 8 strands in 2 sheets; jelly-roll; similarity to the Nucleoplasmin-like/VP fold*	No	Lipid Binding Protein	d1olma2
b.135	Superantigen (mitogen) Ypm	*sandwich; 9 strands in 2 sheets; jelly-roll*	No	superantigen from Yersinia pseudotuberculosis	d1pm4a_
c.2	NAD(P)-binding Rossmann-fold domains	*core: 3 layers, a/b/a; parallel beta-sheet of 6 strands, order 321456*	No	Shikimate dehydrogenase	d1nyta1

c.16	Lumazine synthase	*3 layers, a/b/a; core: parallel beta-sheet of 4 strands, order 2134*	No	lumazine synthase	d1ejba_
c.23	Flavodoxin-like	*3 layers, a/b/a; parallel beta-sheet of 5 strand, order 21345*	No	Lysine aminomutase	d1xrsb1
c.37	P-loop containing nucleoside triphosphate hydrolases	*3 layers: a/b/a, parallel or mixed beta-sheets of variable sizes*	No	elongation factor Sel B	d1wb1a4
c.44	Phosphotyrosine protein phosphatases I-like	*3 layers: a/b/a; parallel beta-sheet of 4 strands, order 2134*	No	IIBcellobiose	d1iiba_
c.66	S-adenosyl-L-methionine-dependent methyltransferases	*core: 3 layers, a/b/a; mixed beta-sheet of 7 strands, order 3214576; strand 7 is antiparallel to the rest*	No	salicylic acid carbox yl methyltransferase	d1m6ex_

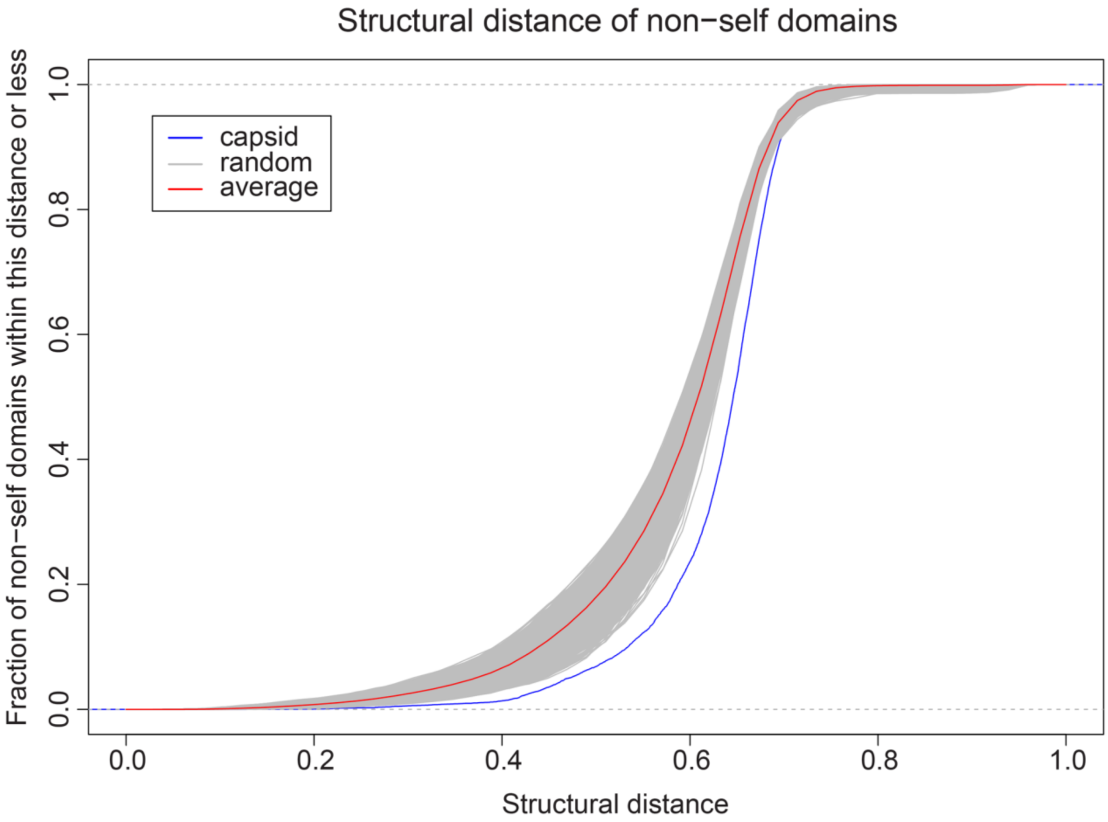

Figure 2.6. Capsid proteins are structurally distant from generic proteins. Each curve plots the empirical cumulative fraction distribution of distances between one set of 56 proteins and their nearest neighbor in the complementary set. The comparison between the capsid set and the non-capsid proteins is colored in blue, while those from the 10,000 permutation tests are colored in grey. The average empirical cumulative fraction distribution of the 10,000 permutation tests is colored in red. The capsid set is clearly further away from its non-self set compared to what happens with random chances.

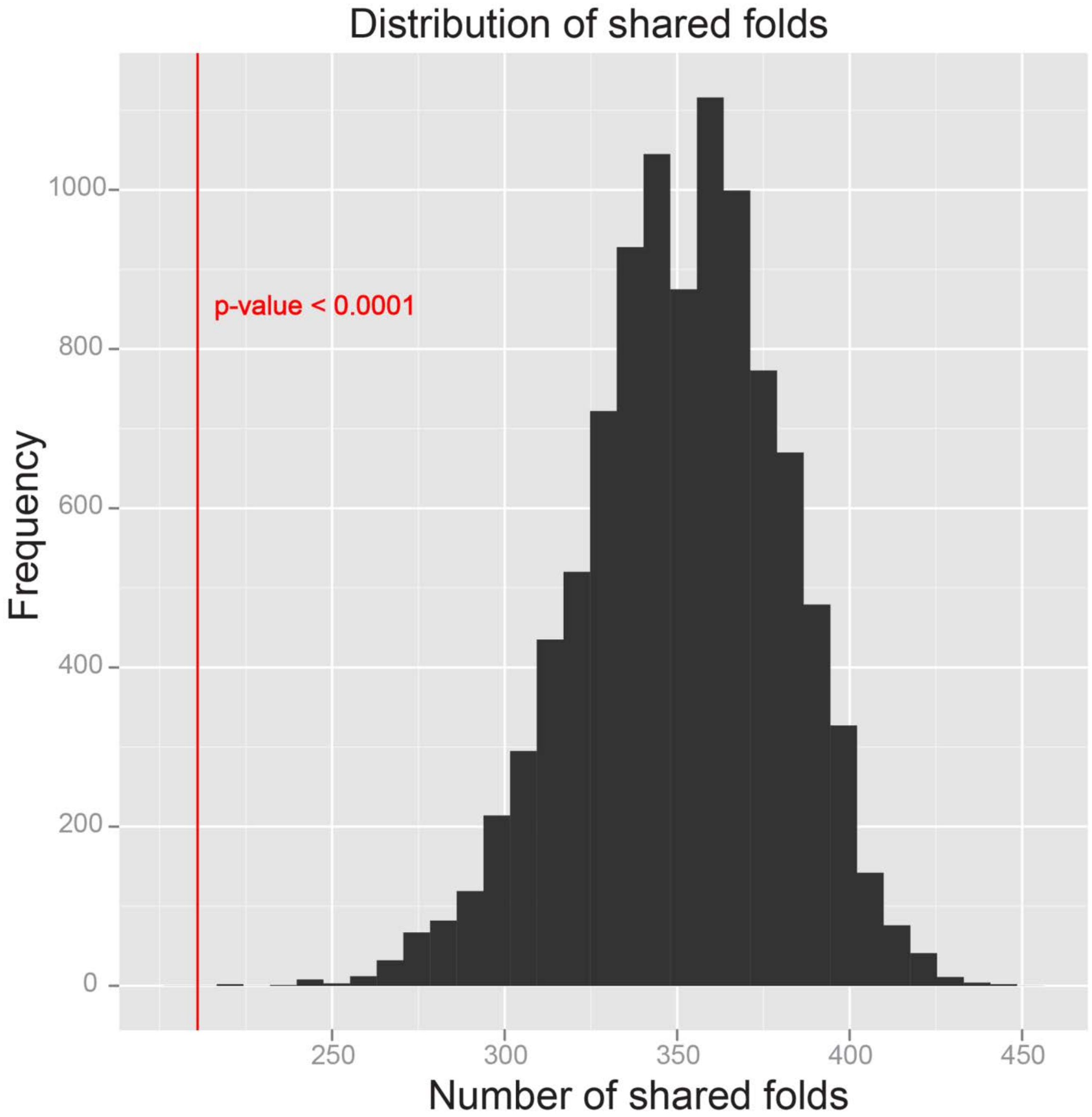

Figure 2.7. Statistical significance of test statistic. No single case in the 10,000 permutations has resulted in 210 or fewer shared folds between the set of 56 protein domains and their complement set, which makes the p-value of our test statistic less than 0.0001, as an upper bound for the statistical significance.

span the protein fold space. We also show in Figure A.1 (Appendix A) that the p-value of our test statistic, based on the datasets containing domains of comparable sizes only, is 0.0002, therefore excluding size as a compounding factor contributing to the difference in fold. In conclusion, viral capsid folds are unique to viruses.

2.3.4 Other functional classes

The seven other functional classes of proteins we examined range in size from 18 to 297 in the total set of 8921 proteins. When compared with their complementary set, the number of shared folds with non-self proteins is found to be statistically insignificant, with a one-tailed p-value greater than 0.05 in all cases (Table 2.2). This is not surprising, given that cellular proteins have evolved over a relatively shorter period of time compared to viral proteins, and therefore their folds are more similar to one another as compared to viral ones, similar being defined by having a TM-score of greater than 0.4. We thus showed that it is not always true that unique function implies unique structural folds. Without making this assumption, we further proved that viral capsid proteins are segregated in structural fold space, which is remarkable.

2.4 Discussion:

2.4.1 Possible differences in domain definition

In this work, our major interest is to compare the independently folded domains of capsid proteins with generic protein domains, so as to reveal their relationship with the higher order of structural organization. Domains defined in this work therefore refer to integral structural units that are connected by single peptide to neighboring domains,

Table 2.2. Seven additional functional classes of proteins studied. These are found to be not significantly distinguished in their folded topology. The shared folds between each functional class of proteins and their complement are not significantly small compared to what happens with random chances, with a one-tailed p-value greater than 0.05 in every case, suggesting that these cellular proteins are highly connected in structural fold space.

Functional class	**Size of class**	**Subgroups, if any, included**	**One-tail p-value**
Kinase	213	-	0.1449
Globin	32	*Myoglobin and hemoglobin*	0.4154
Dehydrogenase	297	-	0.3461
Polymerase	67	*DNA/RNA polymerase*	0.0572
Chaperone	33	-	0.2925
Antigen	49	-	0.4411
Muscle	18	*Actin, myosin, titin, nebulin*	0.1972

although in a few cases these criteria are not fully met. We followed strictly the definition of domains in SCOP to make fair comparison with generic proteins collected from the same database. Our work does not focus on a finer granularity of structure such as subdomains, or motifs, which might have been called ‘domains’ in certain literature for the interpretation of their evolutionary origin. While our choice of domain definition addresses our question of interest adequately, we also note that the question of whether viral folds and generic proteins are evolutionarily segregated can be answered by comparing subdomains or structural motifs, which is outside the scope of discussion here.

2.4.2 What are the capsid-like proteins and why they do not form shells

Prior to our work, several studies have reported that certain classes of cellular proteins also share similar topologies or structural cores with certain capsid proteins. These include the tumor necrosis factor superfamily [121], the serine proteases [126], the superantigen class [135], the concavalin A class [124], and the CUB-like domains [136]. All of the above classes of proteins were among the generic proteins that we found to share similar folds as capsid proteins, as expected. In addition, analysis of our set of 600 non-viral relatives of capsid proteins revealed that many virus proteases, certain hydrolases, transcription regulators and histone chaperones also shared close topological characteristics with viral capsid proteins (Table 2.1).

We first examined the structural relatives that are highly similar to capsid proteins (within a distance 0.4 or less). Many of these structural relatives possess the typical jelly-roll topology, with some variations in each case. The tumor necrosis factor superfamily is characterized by 10 strands in two sheets, with the core eight strands having identical connectivity as that of a standard capsid jelly-roll. Truncation in one strand and addition of two extra strands make them slightly different in shape compared to capsid proteins. The CUB-like domains in spermadhesins display a particular variation of the jelly-roll topology in terms of connectivity, including reversed β-strands, two disulphide bridges and two additional β-strands. They thus share a minimal structural core with capsid proteins (specifically the bean pod mottle virus capsid protein), but have shorter β-strands and overall smaller shape as a distinction. Superantigen Ypm is yet another class that overlaps significantly in structure with capsid proteins, especially satellite tobacco necrosis virus capsid proteins. Other than an additional disulphide bond connecting the C terminus with one β-strand that differentiates itself, superantigen Ypm also has a much more compact structure compared to capsid proteins, owing to its shorter loops connecting the β-strands. The supernatant protein factor protein consists of two domains, and the C-terminal domain also follows the jelly-roll topology that resembles satellite tobacco necrosis virus most, with minute differences in the concavity of the two β-sheets. The histone chaperone proteins are characterized by the same topology as capsid proteins, with some of them having one or two additional strands. Remarkably, all of these proteins discussed occur naturally (as opposed to crystal packing) as heterodimers (the monomers having identical topology), trimers, pentamers or hexamers, although their modes of interaction differ from that of capsid proteins in many cases. This suggests that the β-sandwich formed by proteins with varying connectivity generally facilitates aggregation, presumably because of the greasy, flat surfaces presented by their wedge-like shapes to promote monomer association.

In addition to these structural analogues found naturally in oligomeric states, we also identified quite a few proteins in the immunoglobulin fold and the methyltransferase fold that are highly similar to capsid proteins; however, they typically occur as part of some multi-domain proteins, such as the N-terminal binding fragment of the human polymeric immunoglobulin receptor. It thus might not be feasible to simultaneously arrange all domains on a shell in such cases, which may explain why we are not observing multimeric complexes for these proteins. We omit here discussion on the remaining types of protein domains, mainly for the reason of their limited structural similarity to capsid proteins (distance-wise more than 0.4 apart). These proteins typically either appear smaller in size or are tightly coupled with other domains, and consequently significantly different in shape, and have not been observed to form symmetric complexes in general.

Given the above interesting observations, we need to highlight that the structural relatives of capsid proteins only marginally resemble capsid proteins to the extent of their common structural core, as evident from the large structural distances (majority are greater than 0.4) between the two classes. Decorations on top of this level of similarity directly differentiate the exposed edges of the proteins, such that geometrical complementarities along multiple symmetry axes are easily satisfied by repeating units of the same monomers in the case of capsid proteins but not in the other. In other words, the positions in which monomers interact with one another are also fine-tuned by geometric and physicochemical factors of protein-protein interfaces. We thus do not observe any protein cages assembled from these cellular proteins despite their sharing similar structural topologies with capsid proteins. Lastly, we speculate that the structural but not functional close relationship between these few classes of proteins and capsid proteins resulted from ancient genetic interactions between viruses and their hosts, although further investigation is needed to support this view.

2.4.3 Scarcity and possible bias in the data

An important aspect that cannot be overlooked is that we have drawn our conclusion in this work based solely on existing structural data of capsid proteins taken from icosahedral viruses. We cannot exclude possibilities of identifying novel viral capsid folds that span a larger subspace of protein folds in future, as predicted in several recent publications [1-3] given the diversity of the virosphere. This is especially so when we take into account the current challenges in determining the structure of viral proteins embedded in lipid membranes for enveloped viruses. In addition, experimental limitations in determining the structure of large assemblies place a heavy bias in highly symmetrical viral particles, and thus statistics for irregularly shaped viruses such as HIV are missing in our analysis. Given all structural data available up to this date, we have derived our conclusion with rigor and confidence, but we remain open to potential changes should abundant novel discoveries be made.

2.4.4 Implications on protein-protein interaction and other applications

Our study provided support for the hypothesis that viral capsid proteins, which are functionally unique in viruses in constructing protein shells, are also structurally unique in terms of their folding topology. This implies that protein-protein interactions, in the case of viral capsids at least, confer evolutionary constraints on capsid proteins, specifically on their folds. Bhadur and Janin [137] found that residues making up capsid cores are more conserved than interface residues and surface residues, which highlights a greater selective pressure on capsid structural core. Interpreted together, the characteristic folds (and therefore fundamental shapes) of capsid proteins are most likely a consequence of geometric requirements of the building block so as to form the cage-like macromolecular assembly, which corroborates the theory proposed by Mannige and Brooks that demonstrates a trapezoid as the only shape available to capsid proteins for monohedral tiling into an icosahedron [16]. From a more general point of view, core residues of cellular proteins have also been shown to evolve at a slower rate compared to interface and surface residues [138], with a 25%-35% higher conservation score compared to surface residues. Most studies that investigated the degree to which proteins are subject to constraints due to their interactions with other proteins mainly focused on interface residues [138-140], and it remains to be established whether the greater conservation of structural cores of generic proteins is similarly affected by the interaction with their partners during evolution. Our work sheds light on this missing link by studying the particular case of viral capsid proteins, and it will be interesting to verify whether this evolutionary constraint is true in general.

Additionally, virus-like particles (VLPs), which are self-assembling capsid shells without the infectious viral genetic materials encapsulated, are already a popular choice among a variety of nanoparticle platforms for a wide range of applications both in the biomedical arena and in materials science [141-147]. For a comprehensive review, readers may refer to this paper [148]. Compared to other nanoparticle materials, VLPs offer several advantages, including the full range of protein templates they provide that adapt to diverse environmental conditions including extreme thermal environments [149], their proteinaceous nature which makes them biodegradable [52], and their plasticity to a wide range of synthetic manipulations [150-152]. For biomedical applications, VLP design has been formulated for targeted delivery of drug molecules [53], tissue-specific imaging reagents [151], as well as novel vaccine development [153]. VLPs have also been extensively explored as nanocontainers [75] and nanotubes [154] in materials science. In order to fulfill their desired purposes, VLPs are introduced into new functional modules to facilitate specific interactions with the intended biological sites or nonbiological surfaces, to alter the overall architecture and stability [155], and to package various cargos as well as directing the cage assembly [156]. Our work laid out the fundamental principle in such tailored design of VLP platforms; in order to preserve the assembled architecture of viral capsid shells, it is important for the newly formulated protein subunits to adhere to the library of viral capsid folds. In other words, significant adaptations that result in unfolding or misfolding of capsid proteins are undesirable. Where human creativity has no bound in exploring all synthetic possibilities, feasibility has its bound; decorations on VLPs should minimally disrupt the folded topology and geometry of the building block to make it work.

2.4.5 Are viral capsid interfaces also unique to viruses?

Having established that viral capsid proteins possess distinct folds, we would like to take one step further by examining whether the protein-protein interfaces in viral capsid assemblies are also unique to viruses. Because differences in monomer structure do not imply differences in protein-protein interfaces [132], our conclusion of the uniqueness of capsid fold cannot be directly extended to capsid interfaces. The results of this second comparison will again have interesting implications. Should capsid interfaces resemble those of generic ones, the mode of capsid-capsid interation is then governed purely by physicochemical laws, and evolution merely plays a part in dictating the building block structure for their proper tiling. If, on the other hand, we learn that viral capsid interfaces are quantitatively different from interfaces formed by their cellular counterparts, we can then tap on this difference and design pathogen-specific antiviral drugs targeted at disintegrating the protection shells, without disrupting normal cellular activities. This question is fully addressed in Chapter V.

In summary, our comprehensive analysis of the viral capsid proteins and their cellular counterparts revealed the segregation of capsid proteins in structural fold space. This provides important clues to requirements of the building blocks for the distinctive viral shell architecture; the unique folds of viral capsid proteins present favorable geometry to allow effective packing and assembly into the right complex architecture. With this in mind, the design of gene therapy delivery agents as well as nanoparticles, both targeted at making packing tools, can be tailored to satisfy geometric constraints by following closely the viral capsid templates nature has created for us.

Appendix A

Additional information for CHAPTER II

Table A.1. Capsid proteins added from SCOP that are not deposited in VIPERdb.

SCOP id	Description of protein
d1wyka	Sindbis virus capsid protein
d2df7a1	structural polyprotein VP2
d1vpsa	Polyomavirus vp1 pentamer
d1hx6a2	major capsid protein
d1p2za1	Hexon protein
d1p2za2	Hexon protein
d1ahsa	African horse sickness virus (serotype 4) vp7
d1bvp11	bluetongue virus coat protein vp7
d1bvp12	bluetongue virus coat protein vp7
d1qhda1	viral capsid vp6
d1qhda2	viral capsid vp6
d1jmu.1	protein mu-1
d1u7ka	Gag polyprotein
d2eiaa2	Eiav capsid protein p26
d1em9a	Gag polyprotein capsid protein p27
d1tx9a1	Scaffolding protein D
d1kqra	VP4
d1svba2	Tick-borne encephalitis virus glycoprotein
d2alaa2	Structural polyprotein (P130)
d1slqa	VP4
d1m3ya1	The Major capsid protein of PBCV-1, Vp54
d1m3ya2	The Major capsid protein of PBCV-1, Vp54
d2pxrc1	Gag-Pol polyprotein (Pr160Gag-Pol)
d2v33a1	E1 envelope glycoprotein

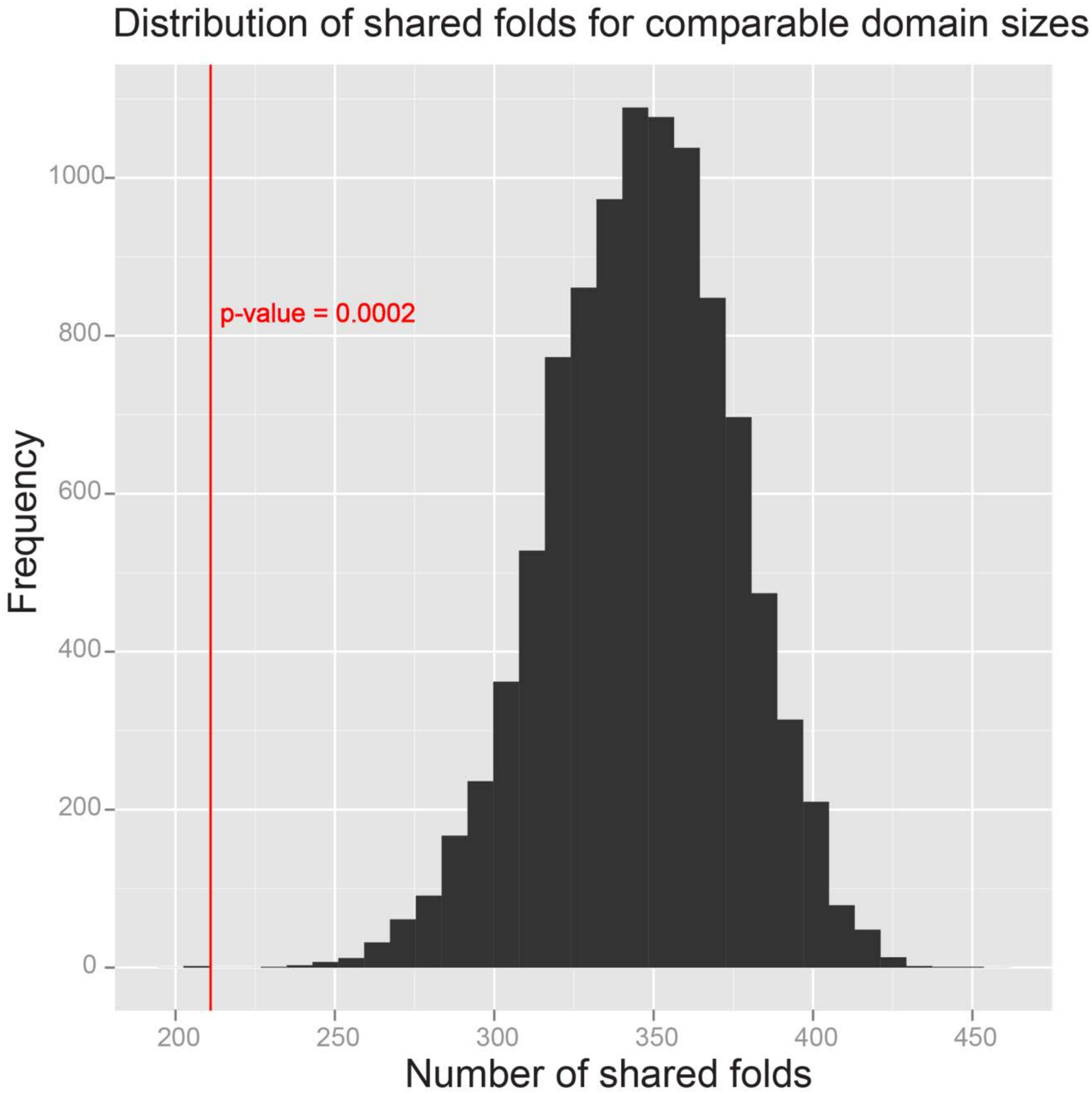

Figure A.1. Statistical significance of test statistic for small domains. For permutation analysis performed on protein domains of 600 residues or fewer, only 2 out of the 10,000 permutations showed 210 or fewer shared folds between the set of 53 proteins and their complement set, which makes the p-value of our test statistic 0.0002 and thus still significant.

CHAPTER III

PCalign: a method to quantify physicochemical similarity of protein-protein interfaces

3.1. Introduction:

Protein-protein interactions play important functional roles in almost all biological activities, including, but not restricted to, signal transduction, gene regulation, catalytic enzymatic activities and maintenance of structure [157]. Characterization and classification of protein-protein interactions would allow us to organize information in protein-protein interaction networks, to make predictions on their function, as well as to facilitate drug design targeted at interfering with those disease-associated protein-protein interactions. While many high-throughput studies are directed toward mapping out the entire protein interactome, providing information on whether two given proteins interact or not, the question of how two proteins interact still requires three-dimensional structural models of protein complexes. Advances in experimental techniques in recent years have led to exponential growth in structural data available for protein complexes [112], and the rise of low-resolution alternative techniques such as cryo-Electron Microscopy (cryo-EM) have made it possible to visualize even large macromolecular complexes that were previously not amenable to crystallization [110]. With increasing efforts from structural genomics initiatives to populate the structural space of protein complexes [158], we expect the gap between protein-protein interaction networks and the matching structural details to close quickly.

At the center of protein-protein interactions are the binding surfaces, or interfacial residues which form contacts between binding partners and stabilize protein complexes. Characteristics of residues lining interfaces have been extensively studied, some focusing on their collective statistics such as hydrophobicity, buried surface area, depth index and planarity [159-161], others focusing on hot spot residues which contribute significantly to the free energy of binding [162, 163]. While these approaches provide insights into the mechanism of protein-protein recognition, they are not suitable for measuring similarities between a given pair of protein-protein interfaces. The latter is useful for revealing potential biological relationships between different complexes, for instance, to discover evolutionarily closely-related protein complexes, or to recognize competitive binding partners for the same protein, or to search potential off-targets for drug molecules that target specific protein-protein interfaces, or to identify structural templates for protein-protein docking studies. Therefore a suitable method to directly compare protein-protein interfaces across randomly selected protein complexes and to quantitatively assess their pairwise similarities is highly desirable.

Depending on the specific biological question being asked, methods for protein-protein interface comparison with different focuses have been developed. Most of the methods are based on structural alignment of protein-protein interfaces, as this provides the most straightforward way to both visualize and quantitatively measure interface similarities. Gao and Skolnick developed the dynamic programming-based algorithm Ialign [164] to detect protein-protein interfaces with shared geometric patterns. The currently available implementation for this method initially defines interfacial residues using atomic details, and then uses positions of the Cα atoms for structural alignment and scoring, and includes a sequence-order dependent version [164] and a sequence-order independent version [165]. Shulman-Peleg *et al.* developed I2I-SiteEngine [166] to compare the physicochemical properties of the functional groups forming protein-protein interfaces, which uses an algorithm similar to pharmacophore mapping. Both of these methods work well for accurately predicting functional relationships between protein complexes determined at atomic resolution. Despite different levels of details involved in the scoring stage (one with Cα atoms and one with side chain atoms and molecular surface), both of these methods require high resolution of the structural models in order to determine what constitutes an interface. As a result, they are not applicable in cases where the data quality of structural models is relatively poor and only backbone atoms are traceable.

The current statistics in the Protein Data Bank (PDB) [112] shows that a very small proportion of the deposited entries were determined at low-resolution, with only 1379 entries having a resolution of 3.5 Å or lower compared to 78511 atomic resolution models (statistics for proteins only). Many of these structures are macromolecular complexes, such as viral capsid shells, that are typically solved by cryo-EM, and may capture protein-protein interaction patterns that occur exclusively in large oligomeric complexes. The current under-representation of these macromolecular complexes in the database merely reflects the limitation of experiments, and does not justify complete neglect of their contribution to our knowledge of protein-protein interactions. In addition, with increasing efforts from the Structural Genomics project to target high-value, novel protein structures, we expect homology modeling techniques to be valuable tools to fill up the structural space. The modeled structures generated as such are unlikely to be high in quality, but nonetheless should contain useful information. In other words, there is significant added value in exploring these coarse-grained structural models when studying protein-protein interfaces, calling for an interface comparison method that can traverse through different resolutions of structural models.

In this work, we develop a method that combines the advantages of existing methods to quantify the similarity of any given pair of protein-protein interfaces in terms of their physicochemical properties. This method not only disregards the sequence-order of interface fragments in performing the structural alignment, but also takes into account the mapping of different chemical types of amino acid residues. More importantly, our method facilitates comparison of structural models determined at different resolutions, greatly expanding the structural space of protein-protein interactions that can be studied systematically.

3.2. Methods:

3.2.1. Extract interfacial residues

While many structural models for protein monomers and homodimers can be determined at atomic resolution by X-ray crystallography or NMR techniques, the structures of large macromolecular complexes are typically solved by cryo-EM and are hence low in resolution. In spite of their lower resolution, these coarser structural models are nonetheless informative and can aid our understanding of protein-protein interactions. Statistics from PDB show over 600 structural models only contain information on the Cα atoms, with many of them populating the lower end of the resolution spectrum (Figure B.1 in Appendix B). To facilitate comparison of interfaces across models with different levels of details, we apply a hierarchical approach in defining interfacial residues with a distance criterion, as described below.

Given the structural model of a protein dimer determined at atomic resolution, we define two residues to be in contact if at least two heavy atoms, one from each residue, are within 4.5Å. The collection of all residues that are in contact with at least one other residue in the binding partner is considered the set of interfacial residues. When side chain information is not available, we use a Cα-Cα distance cutoff criterion to determine if two residues are in contact. Traditionally, a common distance cutoff is used for all types of amino acids. Considering the fact that different amino acids have side chains that vary in size, the Cα-Cα distance between different types of pairs of amino acids that make a contact via their side chains may differ by a non-negligible amount. To account for the side chain size factor, we examined the statistics of Cα-Cα distances for different pairs of amino acids (e.g., a Ser-Lys pair) that are in contact from the PDB. Similar to what Kolinski and Skolnick did in parameterizing pairwise interactions between side chain groups of different amino acids [167], our distance cutoff for Cα-Cα distances is determined by the following,

$$\mathrm{cutoff}_{i,j} = \mathrm{mean}_{i,j} + \xi \times \mathrm{sd}_{i,j} \tag{3.1}$$

where *i*,*j* represent a given pair of residues of amino acid type *i* and of amino acid type *j* respectively (*i*, *j* can be the same type). Mean$_{i,j}$ and sd$_{i,j}$ represent the average values of the Cα-Cα distances in the set of high-resolution (37474) structures in the PDB and their standard deviation respectively. The statistics are listed in Appendix B, Table B.1 and B.2. ξ represents a multiplication factor that is of a fixed value across different residue types, and its optimal value was determined to be 0.5 based on our correlation study (see Appendix B, Figure B.2).

After extraction of the interfacial residues, only the coordinates of the Cα atoms are retained for use in computing the pairwise interface similarity score, so as to allow comparison of interfaces with different levels of structural details. This bare bones criterion sufficiently captures the skeleton architecture that hosts amino acid residues of various chemical types at an interface, without adding noise to the data representation that arises from fluctuations of the side chain orientations.

3.2.2. Identify initial alignments

In order to quantitatively assess the degree to which two sets of interfacial residues resemble each other spatially and chemically, we first superimpose the two sets of interfacial residues. As interfacial residues are fragments that are clustered at the binding site without these fragments necessarily following the peptide sequence order, we chose the sequence order-independent comparison technique, geometric hashing [168-171], to find the transformation needed to superimpose one interface onto the other. This algorithm treats each interface as a set of color-labeled points (of the Cα atoms) scattered in the three-dimensional space, where the color corresponds to the chemical type of each residue. The goal is to find a transformation (i.e. translation and rotation) applied to one point cloud to be overlaid with the other so as to maximize the number of points that match spatially and chemically. Details of the algorithm dealing with simply the geometric properties of the interfacial residues are explained in Appendix B (B.2). In short, the algorithm uses a voting procedure to count the number of points that can be matched between two sets of points for a given superposition, where a point *i* is described by a feature defined by the Cartesian coordinates of the point, (x_i,y_i,z_i). The superposition that receives a high vote corresponds to one having many points that can be matched between the two sets of interfacial residues.

There are two additional factors to be taken into account in this problem; first is that an interface is not a single entity (a set of points) but rather consists of two binding fragments, A and B. Thus in aligning one interface to another, one needs to simultaneously align fragment A in interface 1 to its counterpart, fragment A’ in interface2, and B to B’, and not allow crossing over. To achieve this, we add to the feature of each residue (xyz-coordinates originally) an additional attribute of fragment label (either binding site A or binding site B), and count the votes only when both the coordinates and the fragment labels match. Without knowing the correspondence of binding fragments prior to alignment, we attempt both ways by swapping the fragment labels.

In addition to the fragment label, another factor to be considered is the chemical label that is associated with each residue. Given the coarse-grained nature of our method (using only Cα atoms), we applied a reductive method to classify the 20 amino acids based on the prominent functional group in each side chain. The assignment of the individual functional groups in the side chain is based on the definition in earlier work by Schmitt *et al* [172]. Depending on what functional groups are present in a specific amino acid, we classified them to one of the following six categories: donor (K, R), acceptor (E, D), mixed donor/acceptor (N, Q, S, T), aromatic (F, W), Aliphatic (C, A, I, L, M, P, V, G) and mixed donor/acceptor or aromatic (H,Y). This classification scheme largely agrees with previous assignment of the residue type [173]. Since the six classes are not mutually exclusive, we allow matches across two different classes as long as they share at least one common functional group. For instance, asparagine has both a donor group, ND2-HD21, and an acceptor group, OD1, and thus classified as “mixed donor/acceptor”. Therefore asparagine can be matched with either arginine due to the shared functional group of a donor, or aspartate due to the shared functional group of an acceptor. In summary, we consider features representing residues as equal if their binned xyz-coordinates and fragment labels are the same and their chemical labels match (not necessarily identical).

Lastly, different pairs of orthogonal bases typically yield degenerate transformation matrices, and thus those receiving sufficiently high votes are further clustered to retain representative transformations as a last step, with the 100 top-ranking transformations processed for further refinement.

3.2.3. Iterative refinement

The previous step proposed candidate transformations to superimpose the two interfaces initially, and in the iterative refinement step we aim to further improve the structural alignment in order to maximize the final similarity score. Based on each proposed initial alignment, a list of structurally equivalent pairs of residues between the two interfaces can be identified. Here we used maximum weight matching in bipartite graphs [174] to identify structural equivalence, which is the problem of optimizing one-to-one mapping between two sets of nodes based on the weight of the edge that connects two nodes (one from each set). We implemented the Hungarian algorithm [175], which is explained in detail in Appendix B (B.3).

In our particular problem, we'd like the weight to reflect spatially how close two residues are and also how well their chemical types match, and hence we've chosen the following scheme to quantify equivalence between i^{th} residue in interface 1 and j^{th} residue in interface 2,

$$\text{equivalence-score}_{ij} = \frac{1}{1+0.25\times\left(1-I_{ij}(\text{same chem type})\right)+\frac{d_{ij}^{2}}{16}} \tag{3.2}$$

where I_{ij}(same chem type) is the indicator function that takes the value of 1 when the pair of residues (i, j) share the same chemical type and 0 when they don't. d_{ij} is the Euclidean distance between the Cα atoms of the two residues in Å after structural superposition. After obtaining the list of equivalent residues, we then apply the Kabsch algorithm [176] to translate and rotate the second interface so as to minimize the sum of squared errors between all the equivalent pairs of residues. Based on this new structural superposition, we obtain a new updated list of structurally equivalent residues using maximum weight matching in bipartite graphs, which will be submitted to the same procedure for refinement. This process is iterated until no further improvement is possible.

The overall scoring function, PC-score$_{\text{raw}}$, based on the converged alignment is given by

$$\text{PC-score}_{\text{raw}} = \frac{\boldsymbol{f_c}}{\boldsymbol{L_{ave}}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{L_{ali}}}\frac{\mathbf{1}}{\mathbf{1+0.25}\times\left(\mathbf{1}-\boldsymbol{I_{ii}}(\text{same chem type})\right)+\frac{\boldsymbol{d_{ii}}^{\mathbf{2}}}{\mathbf{16}}} \tag{3.3}$$

where L_{ave} is the average number of interfacial residues for the pair of interfaces compared, and L_{ali} is the number of all aligned residues identified by the aforementioned algorithm that have an equivalence-score of 0.20 or higher. f_c is the ratio of common contacts between the two sets of aligned interfacial residues, and is calculated as

$$\boldsymbol{f_c} = \frac{\overleftrightarrow{\boldsymbol{N}}^{\boldsymbol{1}} \cdot \overleftrightarrow{\boldsymbol{N}}^{\boldsymbol{2}}}{(\overleftrightarrow{\boldsymbol{N}}^{\boldsymbol{1}} \cdot \overleftrightarrow{\boldsymbol{N}}^{\boldsymbol{1}} + \overleftrightarrow{\boldsymbol{N}}^{\boldsymbol{2}} \cdot \overleftrightarrow{\boldsymbol{N}}^{\boldsymbol{2}})/\boldsymbol{2}} \tag{3.4}$$

where $\overleftrightarrow{N}^1$ and $\overleftrightarrow{N}^2$ are $L_{ali} \times L_{ali}$ matrices representing the contact maps of the aligned interfacial residues in interface 1 and those in interface 2 respectively. The dot operation represents the inner product. This scoring function is largely adapted from the scoring function of IS-score for the program Ialign [164], given its demonstrated excellent performance in the original study, with the modifications here to specifically address our question of interest.

Finally, the raw PC-score is further scaled by the following equation to remove the dependency of the score on the interface size to derive our final scoring function:

$$\text{PC-score} = \text{PC-score}_{\mathbf{raw}}{}^{\boldsymbol{ln}\mathbf{0.3}/\boldsymbol{ln}(\mathbf{0.14}+\mathbf{0.29}\times\mathbf{0.97}^{\boldsymbol{L_{ave}}})} \tag{3.5}$$

This scaling function was derived by fitting the curve of the raw PC-score as a function of the size of randomly selected pairs of interfaces being compared (see Appendix B, Figure B.3).

The candidate alignment that receives the highest PC-score represents the optimal alignment solution, and its associated PC-score gives the measurement of physicochemical similarity between the two interfaces being compared. This scoring function is normalized between 0 and 1, and takes the value of 1 when comparing two identical interfaces. The associated statistical significance with a PC-score is derived empirically from the distribution of PC-scores for random interface alignments (Appendix B, Figure B.4).

3.3. Results:

3.3.1. Validation of the scoring function

Although the quantification of protein-protein interface similarity has no corresponding experimental observables to benchmark against, we can nonetheless evaluate whether the scoring function is reasonable by comparing it with other physically sound metrics. This is only a proof of concept, but still provides useful information in terms of judging the performance of the method. Specifically, we tested our scoring function against the Q-score in quasi-equivalent viral capsid protein-protein interfaces.

The Q-score is a normalized score based on equivalent residue contacts in interfaces formed by viral capsid proteins [177]. Viral capsid proteins are special structural proteins; repeating units of the same capsid protein assemble into large, symmetric shells that embed the viral genetic materials inside. An inter-subunit interface within a capsid is thus formed by two monomers with the same peptide sequence. In the smallest icosahedral viruses, 60 copies of the same protein tile the icosahedral shell, where each protein is placed in the same environment. Correspondingly, all the interfaces with the same dimerization states are chemically identical. For larger viruses, however, multiples (with the multiplicity denoted by the Triangulation number or T-number) of 60 copies of the same protein assemble into macromolecular complexes that also obey icosahedral symmetry. Based on the theory proposed by Caspar and Klug [11], this can be achieved by allowing slightly varied modes of interaction in the proteins such that those protein-

protein interfaces following strict 2-fold, 3-fold or 5-fold symmetry (as in an icosahedron) and those which do not are quasi-equivalent to each other, but not identical.

The Q-score was developed to specifically quantify the level to which two quasi-equivalent interfaces resemble each other. First the contact map represented by an N×N matrix of 1's and 0's between the two binding partners of each interface is calculated, where N is the number of amino acids in the capsid protein. The Q-score is computed by taking the normalized inner product of the two contact maps of the interfaces:

$$\text{Q-score} = \frac{2\overleftrightarrow{N}^{a}\cdot\overleftrightarrow{N}^{b}}{\overleftrightarrow{N}^{a}\cdot\overleftrightarrow{N}^{a}+\overleftrightarrow{N}^{b}\cdot\overleftrightarrow{N}^{b}} \quad (3.6)$$

$$= \frac{\boldsymbol{2\sum_i\sum_j N^{a}_{i,j}N^{b}_{i,j}}}{\boldsymbol{\sum_i\sum_j\left[(N^{a}_{i,j})^2+(N^{b}_{i,j})^2\right]}} \quad (3.7)$$

where $\overleftrightarrow{N}^{a}$ and $\overleftrightarrow{N}^{b}$ are matrices representing the contact maps of interface a and interface b respectively.

Hence the Q-score reflects the ratio of common contacts between two interfaces, and is equal to 1 for identical interfaces and 0 for two interfaces with no common contacts. This quantification metric is thus a reasonable one with straightforward physical interpretation in the case of capsid protein-protein interactions. We therefore attempted to compare our interface similarity score with the published Q-scores for 18 T=3 viruses in [177].

Our results show that our interface similarity score largely agrees with the Q-score in viral capsid protein-protein interfaces, with a high overall correlation coefficient of 0.93 (Appendix B, Table B.5). Although the Q-score only measures geometric properties, we showed that with the chemical type taken into consideration in our method, the two agree well, suggesting that our scoring metric is reasonably accurate in capturing the physical properties of interfaces.

3.3.2. Comparison of performance with existing methods

The performance of PCalign is compared with two existing methods, Ialign with its sequence-order independent version and I2I-SiteEngine [166], by testing how well these methods can distinguish highly similar interfaces from less similar ones. We manually collected from the database, Structural Classification Of Protein-Protein Interfaces (SCOPPI) [178], a set of interfaces which are grouped into several different families based on their evolutionary relationship, and evaluate if our scoring function, as well as the two existing methods, can provide a reasonable cutoff value to separate interfaces that are highly related from those that are not, where the "relatedness" label is given by pairs of interfaces annotated to be in the same group. We should highlight here that the goal of our method is to quantify the physical and chemical properties of protein-protein interfaces, regardless of the familial relationship between the monomers forming the interfaces. Thus benchmarking similarity of interfaces against evolutionary relatedness of the monomers forming the interfaces does not accurately reflect the performance of the three methods. Nonetheless, it gives us a crude measurement of how confident we can be in applying our method to study interfaces in general.

Together we collected 609 dimers from 124 pairs of protein families in SCOPPI, where dimers within the same group are obtained after applying a 50% sequence redundancy filter in the database, and are also selected to have similar interaction modes (i.e., they present the same "faces" at the binding sites, based on the database classification) [178]. The chosen families cover a wide range of structural folds, as annotated by the first letter and the succeeding number in their SCOP ID (Appendix B, Table B.6). We then performed an all-against-all comparison for these 609 interfaces using the three methods, and obtained their respective scores for each pair compared. For I2I-SiteEngine, we included all three scores reported by the program, including the match score, the total score, and the t-score [166]. In addition to applying our method to the original data, in order to demonstrate the robustness of our method in application to noisy low-resolution models, we applied the same analysis described above to the "backbone" set, which is the same data set reported here except that all structural models were first "corrupted" to retain their Cα atoms only, and then had the positions of the Cα atom perturbed in a random direction by a magnitude that follows a Gaussian distribution centered at 0 with a standard deviation of 1 Å. This creates an artificial low-resolution dataset for testing the robustness of our method. The two existing methods do not deal with low-resolution data and are thus not applied to this backbone dataset. For any given cutoff value of each score, we tabulated the counts of true positives, false negatives, false positives and true negatives for all pairs of interfaces, where real positives correspond to pairs of interfaces belonging to the same pair of protein families, and predicted positives correspond to pairs of interfaces that have a similarity score higher than the given cutoff value. We thus derived the receiver operating characteristic (ROC) curve of the three methods so as to evaluate if the method serves as a good classifier of related/unrelated interfaces.

Our results show that our method performs comparably to the two exsiting methods in capturing most of the interface similarity reflected in the evolutionary relationship. As shown in Figure 3.1, the area under the curve (AUC) value of predictions derived from IS-score (in orange), which corresponds to the alignment program Ialign, ranks the highest at 0.980. Following Ialign, our PCalign applied to the original data (in red) and to the backbone data (in magenta) resulted in AUC values of 0.970 and 0.955 respectively. The three scores reported by I2I-SiteEngine, the match score (in blue), the total score (in green), and the t-score (in cyan), gave an AUC value of 0.831, 0.884 and 0.909 respectively. From this result we see that Ialign is a very useful tool in predicting highly related interfaces. Although our method may appear inferior when applied to this dataset, it should be noted that our method is developed to detect similar interfaces, which may result not only from evolutionary relatedness but also arise from nature's recycling her limited choices of interface design. What were reported by PC-score to be FPs could well be putative positives based on interface and not monomer structural similarity, which can be captured by our program and may be dismissed by Ialign. To verify if indeed this is the case, we carried out further analysis as follows.

Because we are interested in knowing if PCalign does better than Ialign in recognizing interface similarity across unrelated protein dimers, we select one representative structure from each of the 124 clusters of protein dimers, and compare all-against-all. For each pair compared, we tabulate the fraction of aligned interfacial residues (i.e. coverage) as well as the RMSD between the aligned interfacial residues. Ideally, a good structural alignment program should find high coverage and low RMSD values. Note that this criterion is purely geometric. As shown in Figure 3.2(A), PCalign aligns slightly more interfacial residues on average compared to Ialign, with lower

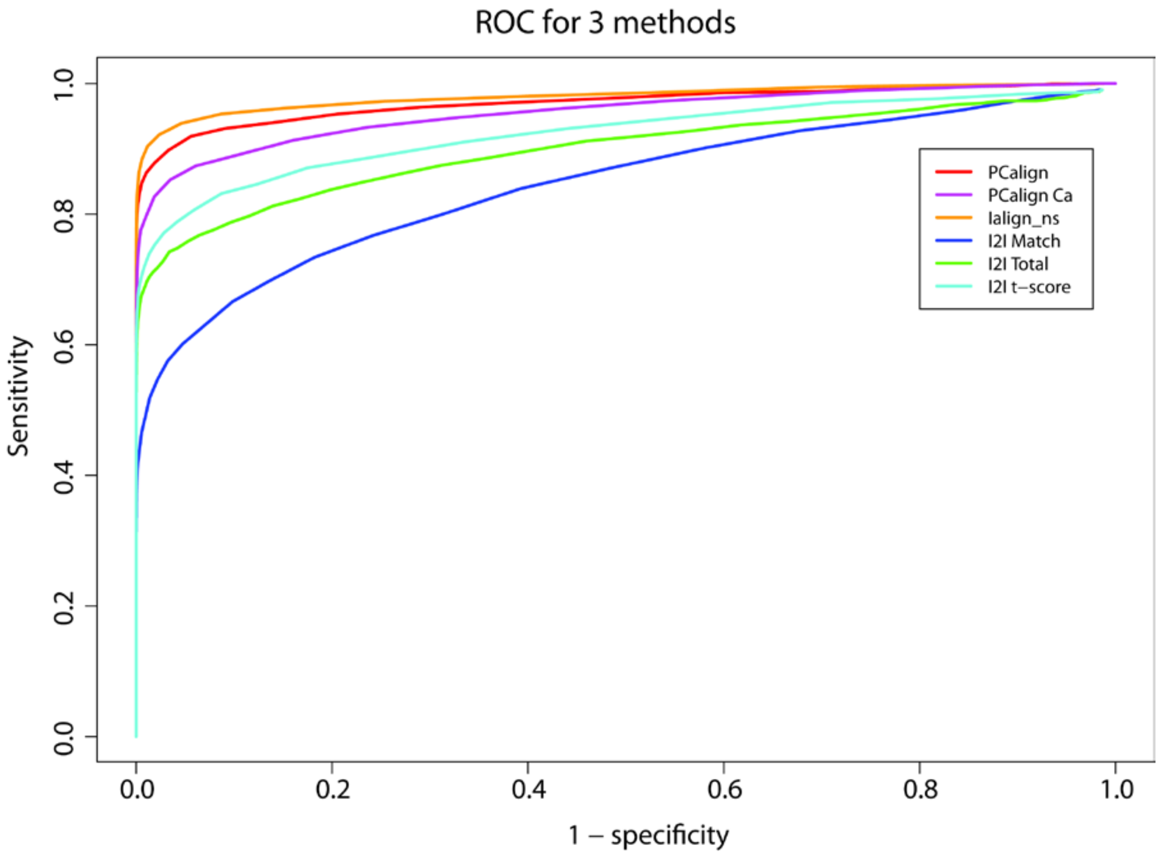

Figure 3.1. The ROC curves for predicting highly related interfaces using three methods, PCalign, Ialign, and I2I-SiteEngine. As shown by the red and magenta curves, our method PCalign gives an AUC value of 0.970, and for the backbone set 0.955. In comparison, Ialign gives an AUC of 0.980. I2I-SiteEngine performs slightly worse, with those predicted by match score, total score and t-score having AUC values of 0.831, 0.884 and 0.909 respectively.

RMSD as well, even though PCalign considers both geometric and chemical aspects of protein-protein interfaces. When the chemical term in our scoring function (Equations 3.2 and 3.3) is turned off to only compare structural alignment from a geometric point of view, the advantage of PCalign over Ialign becomes more pronounced (Figure 3.2(B)). In addition, we find for each of the 124 interfaces its nearest neighbor in the non-redundant set based on PC-score and IS-score respectively, and perform the same analysis. We again observe the same trend for the closest, unrelated match identified by the two methods, with PCalign marginally outperforming Ialign with and without the chemical term considered (Figure 3.2(C), (D)).

As such, our method may detect more often than other methods highly similar but non-related interfaces that are counted as FPs in this benchmark test, which is the question of interest that our method aims to address and which cannot be accurately assessed by the classification analysis in Figure 3.1. To further check the odds of our method outperforming Ialign versus the other way round in terms of getting a better structural alignment, we tabulate the statistics of each scenario among all 185136 pairs compared, again using the same geometric criterion as in Figure 3.2. With the chemical term switched on, PCalign outperforms Ialign in 50838 pairs, with an average coverage of 0.623 and an average RMSD of 3.64 Å, as compared to 0.483 and 4.10 Å for Ialign respectively. Ialign is found to outperform PCalign in 34087 cases, with an average coverage of 0.619 and an average RMSD of 3.60 Å, as compared to 0.486 and 4.04 Å for PCalign. If we remove the compounding factor of chemical types for fairer comparison, PCalign outperforms Ialign in 57639 cases, with an average coverage of 0.629 and an average RMSD of 3.63 Å, as compared to 0.486 and 4.09 Å for Ialign respectively. In

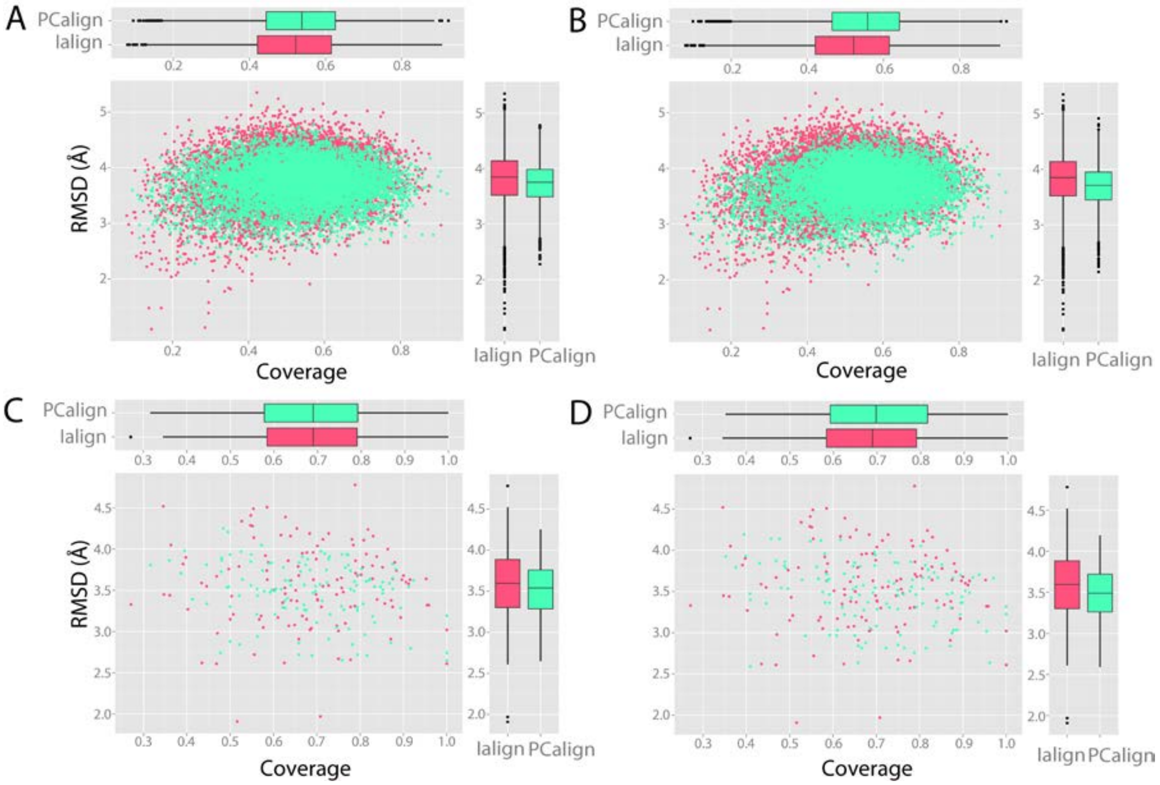

Figure 3.2. Recognition of interface similarity across unrelated interfaces by PCalign and Ialign. The comparison is based on two geometric criteria; fraction of aligned residues (coverage) and RMSD of aligned residues. (A) All-against-all pairwise comparison, with PCalign (Ialign) aligning on average 53.1±13.3% (51.4±13.8%) of residues with RMSD of 3.725±0.371Å (3.810±0.473Å). (B) All-against-all pairwise comparison, with PCalign (Ialign) aligning on average 54.8±13.2% (51.4±13.8%) of residues with RMSD of 3.686±0.378Å (3.810±0.473Å), where the chemical term in PCalign is turned off to capture geometric similarity only. (C) Closest unrelated interface in the set of 124 dimers, with PCalign (Ialign) aligning on average 68.4±14.5% (68.3±15.6%) of residues with RMSD of 3.483±0.366Å (3.563±0.502Å). (D) Closest unrelated interface in the set of 124 dimers, with PCalign (Ialign) aligning on average 70.1±15.4% (68.3±15.6%) of residues with RMSD of 3.466±0.371Å (3.563±0.502Å), considering the geometric part of the scoring function in PCalign only. In all scenarios, PCalign does slightly better than Ialign in recognizing geometric similarities across unrelated interfaces, and using a scoring function that considers both chemical and geometric properties in PC-score performs less well compared to using one that considers purely geometric properties in PC-score, due to the fact that this analysis uses purely geometric criteria.

contrast, Ialign does better than PCalign in only 27790 cases, with an average coverage of 0.630 and an average RMSD of 3.56Å, as compared to 0.504 and 3.99 Å for PCalign. In summary, we have an odds ratio of 1.5 for PCalign doing better than Ialign with the original scoring function of PCalign, and an odds ratio of 2.1 for PCalign performing better when we only consider the physical environment of protein-protein interfaces. Figure B.5 in Appendix B gives an anecdotal illustration of a scenario where PCalign recognizes significant structural similarity between two unrelated interfaces that is missed by Ialign.

In terms of computational time, our method is slower than Ialign and faster than I2I-SiteEngine (Appendix B, Figure B.6). This higher computational cost arises from the algorithm complexity, which samples a larger initial alignment space by disregarding the peptide topology. Nonetheless, such cost is sufficiently low for our method to be applied to large-scale comparison studies. We would therefore argue that the performance of our method parallels those of existing methods. Additionally, our method tackles structural models spanning the resolution spectra, which existing methods fail to do, and is able to detect spatial and chemical patterns shared by interfaces regardless of their sequence similarity in the constituent monomers. We thus expect our method to be a handy tool in exploring the repertoire of protein-protein interfaces and understanding their structural relationships.

3.3.3. Application of the method in detecting convergently evolved similar interfaces

Our method quantifies interface similarity based on the spatial and chemical organization of discontinuous interface fragments. This method therefore accounts for sequence order-independent patterns shared between protein-protein interfaces that arise not necessarily from divergent evolution, but potentially from convergent evolution, which leads to identification of functional relationships masked by apparent lack of structural resemblance.

One such interesting example is viral mimicry. Over the long-standing history of pathogen-host interaction, viruses have evolved various strategies to evade detection by the host immune system [179], to manipulate the cellular signaling network to their advantage [180], and to hijack the cellular transcription and translation machinery for self-replication [181]. Among these strategies is molecular mimicry, which can arise sometimes from viruses capturing host genes followed by deriving their homologues via divergent evolution, and more frequently from viruses independently evolving similar binding sites without any sequence or structural similarity to the endogenous protein they compete with [182]. The latter is especially of interest, as being able to identify which endogenous proteins are displaced by these viral proteins when such mimicry is masked by the lack of sequence and structural similarity can significantly enhance our understanding of how viruses interfere with the cellular pathways for their purposes.

To illustrate the usefulness of our tool in detecting interface mimicry in virus-host interaction, we show here three examples of viral mimicry that are well understood (Table 3.1). The first example concerning immune evasion is that of the Murid herpesvirus 4 M3 protein, which binds strongly to the CC chemokine ligand 2/monocyte chemoattractant protein 1 (CCL2/MCP-1) [183]. It is known that chemokines play a

Table 3.1. Three examples of viral mimicry achieved via convergent evolution. The first seven columns describe the identity of the proteins involved as well as their representative PDB IDs and corresponding chains used for our analysis. The eighth column gives the sequence identity between each viral protein and the endogenous binding partner it displaces, using the program ClustalW [201]. The ninth column gives the template modeling score (TM-score) between each viral protein and the endogenous binding partner it mimics, where TM-score measures the structural similarity between two proteins [109], and a TM-score of 0.4 or higher typically implies high structural similarity. The last column corresponds to the PC-score of the two interfaces measured by our program, and a PC-score of 0.4 or higher implies highly similar interfaces. In all three examples, the cognate binding partner of the target protein shares no sequence or structural similarity with the viral protein, as evident from the low sequence identity and TM-scores. In all three cases, PCalign computes a significantly high score for the two interfaces, recognizing the interfacial patterns that the viral proteins evolved to mimic over time. We also compute the interface similarity scores returned by the two existing measures. For Ialign, the non-sequential version and not the sequential version recognizes the convergently evolved interfaces as significantly similar with IS-scores of 0.393, 0.436 and 0.299 for the three cases respectively. For I2I-SiteEngine, we had no success generating the input files for the first case of M3 protein for which the program reported an error. For the remaining two cases, the match scores are 27 and 22, the total scores 1339 and 648, and the t-scores 47 and 44 respectively. While it is not clear whether these scores are statistically significant, a comparison with the large scale study performed on the SCOPPI benchmark dataset suggests that these scores are not sufficient (the latter suggest a match score of at least 31 to be classified as similar interfaces).

Viral protein	PDB ID	Human target protein	PDB ID (complexed with viral protein)	PDB ID (complexed with cognate partner)	Displaced human binding partner	PDB ID	Seq identity	TM-score	PC-score
Murid herpesvirus 4, M3	2nz1A	C-C motif chemokine 2	2nz1D	1dokA	C-C motif chemokine 2	1dokB	7%	0.17940	0.445
Simian virus 5, nonstructural protein V	2b5lC	DNA damage-binding protein 1	2b5lA	3ei4A	DNA damage-binding protein 2	3ei4B	4%	0.21976	0.546
Nipah virus, glycoprotein G	2vskA	Ephrin-B2	2vskB	2hleB	Ephrin type-B receptor 4	2hleA	12%	0.24913	0.430

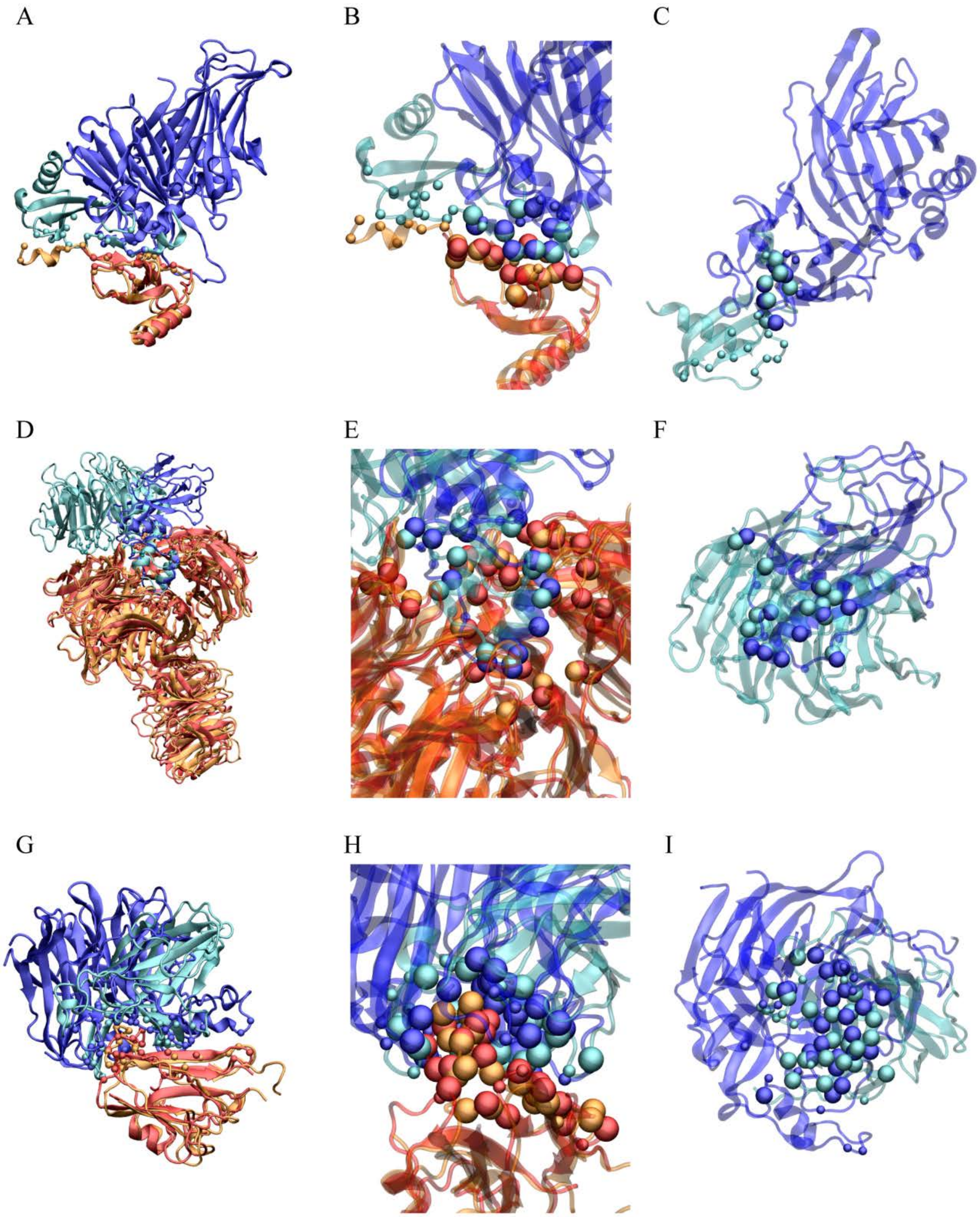

Figure 3.3. Three examples of viral mimicry resulting from convergent evolution. The first example is that of the M3 protein mimicking CCL2 in complexing with another CCL2 monomer (A,B,C), the second being the V protein competing with DDB2 in binding with DDB1 (D,E,F), and the third case being the G protein targeting the ephrin B2 ligand in similar ways with its native ephrin type-B receptor 4 (G,H,I). They are shown with the two complexes superimposed (A,D,G), with a focused view of the matched interfacial residues (B,E,H), and with just one binding site on the viral protein and that on the host protein it mimics (C,F,I). In all illustrations the viral protein is colored in blue, and the host protein it displaces is colored in cyan.

The human target protein is colored red when bound with the viral protein, and orange when complexed with its cognate binding partner. The small spheres represent the Cα positions of all the interfacial residues present in the original complex, while the large spheres represent those which are structurally equivalent in the virus-host protein complex and in the endogenous complex. Figures are generated by the VMD software [184].

crucial role in inducing directed chemotaxis for trafficking of nearby leukocytes [185], which is part of the host immune response. Studies have shown that oligomerization of the CC chemokines, among other types of chemokines, is critical for recruiting cells in vivo [186]. Herpesvirus thus evolved the M3 protein as a decoy receptor for CCL2 as shown in the experimental structure [183], which binds strongly to the chemokine at the same site where it forms a homodimer with another chemokine [187], therefore inhibiting oligomerization of chemokines that is necessary for its recruitment of leukocytes. Using our method PCalign, we found that despite the complete lack of sequence and structural similarity between the viral protein and the one it displaces, the M3-CCL2 interface indeed overlaps extensively with that of the CCL2 homodimer interface, with a high PC-score of 0.445 (Figure 3.3(A), (B), (C)).

In a second case pertaining to viral pathogenesis, the Simian virus 5 V proteins target the DNA damage-binding protein 1 (DDB1), a protein involved in the ubiquitin-proteasome pathway, leading to degradation of the STAT1 protein [188]. The latter results in the type I interferon signaling pathway being blocked, effectively preventing the establishment of a cellular antiviral environment [189]. In achieving this function, the viral protein has adopted a similar binding site as that of the DNA damage-binding protein 2 (DDB2), which is known to form a complex with DDB1 to participate in UV-induced nucleotide excision repair [190], as well as in stimulating E2F1-activated transcription [191]. Experimental evidence exists that the V protein and DDB2 bind to DDB1 in a mutually exclusive manner [192]. Thus through sequestering DDB1 and inhibiting its association with DDB2, V proteins are expected to disrupt the normal function of the UV-DDB complex in DNA repair and cell cycle regulation, which are associated with the viral pathogenesis [193]. Through analyzing the structural models of the V protein-DDB1 complex [194] and the UV-DDB complex [190], we again found significant interface similarity with a PC-score of 0.546, and it is clear from Figure 3.3(D) that such mimicry is established from the viruses' rapid mutation leading to the converged interface, rather than from divergent evolution.

The last example of viral mimicry involves the mechanism of viral entry into host cells. Nipah viruses employ their attachment glycoprotein G (NiV-G) for anchoring to the cell surface before initiating membrane fusion, specifically via binding of the glycoprotein G to ephrin-B2 [195, 196], which is a transmembrane ligand for the ephrin B class of receptor tyrosine kinases. Comparison of the NiV-G-ephrin-B2 complex and the cognate ephrin-B2-ephrin-B4 receptor complex reveals striking similarity in their structures [197, 198], preserving key interactions at the G-H binding loop in both cases [199]. Although experimental evidence for the viral protein's competitive binding to ephrin-B2 with the target protein's cognate receptor remains to be established, the observed interface structural similarity has already spurred propositions of therapeutic schemes that target the anchor site of the viral protein while avoiding disrupting the endogenous ephrin receptor interactions [199]. Unsurprisingly, this interface mimicry is also captured by our method with a high PC-score of 0.430, demonstrating the power of our method in identifying shared patterns correlated with biological significance.

We also performed the same analysis using the two existing methods. While the non-sequential version of Ialign also detected significant similarity in all three cases, I2I-SiteEngine assigns scores that do not quantify these mimicked interfaces as sufficiently similar (Table 3.1). Based on these anecdotal analyses of protein-protein interfaces bearing biological significance with limited overall structural similarity in the constituent proteins, Ialign and PCalign appear to be the recommended methods for detecting the interface similarity when all structural details are available.

3.4. Discussion and conclusions:

Characterizing, classifying and annotating protein-protein interactions are fundamental to understanding the structural or functional relationship between proteins, and to provide additional insights into what can be revealed by studying individual proteins alone. Central to protein-protein interactions from a structural point of view is protein-protein interfaces, which may be dissimilar for similar monomers, and similar for dissimilar monomers. Structural comparison of protein-protein interfaces is thus expected to aid in organizing information hidden in the protein-protein interaction network, and enable predictions for novel biological functions undisclosed by protein monomer structures.

This work presents PCalign, a method to quantitatively measure interface similarity for a given pair of protein-protein interfaces, taking into account the chemical and spatial patterns of residues lining the interfaces. It primarily uses a geometric hashing algorithm to identify the optimal superimposition of two sets of discontinuous fragments of interfacial residues while disregarding their connectivity. Based on the optimal superimposition, a normalized scoring function, PC-score, is calculated to reflect the extent to which the two sets of interfacial residues overlap with each other in terms of their physicochemical properties. A major contribution of this new method is that PCalign adopts a coarse-grained approach in representing interfaces, aligning interfaces and scoring the alignment, therefore it is able to accommodate input data across different resolutions. This is expected to gain advantage over existing methods in the next era of structural bioinformatics, given the rate at which large macromolecular complexes solved at nano-resolutions continue to populate the pool of structural data. Performance-wise, we demonstrated that our method is comparable to existing methods in terms of accuracy and computational complexity, and is able to detect significant structural similarities at protein-protein interfaces, which are sometimes missed by existing methods.

As PCalign is aimed at capturing the overall degree of equivalence between protein-protein interfaces, a necessary limitation with such a design is the lack of sensitivity towards local structural motifs shared among interfaces that are globally dissimilar (‘globally’ here refers to the entire interface, rather than the entire monomer structure). In such situations the local signal becomes diluted out upon normalization against the average interface size. In this regard, caution needs to be taken when screening for small signature motifs embedded in large interfaces using our method that was designed for measuring global similarity among interfaces.

With this new tool for protein-protein interface comparison, we would now like to expand our investigation of structural properties of protein-protein interfaces by analyzing large macromolecular assemblies, such as viral capsids. This class of proteins presents unique structural and functional characteristics unseen in cellular protein complexes [200], and the wealth of information contained in their structural data may enrich our knowledge of protein-protein interfaces in general. We expect PCalign to be a useful tool in exploring some interesting questions pertaining to the higher-order organization of these assemblies.

Appendix B

Additional information for CHAPTER III

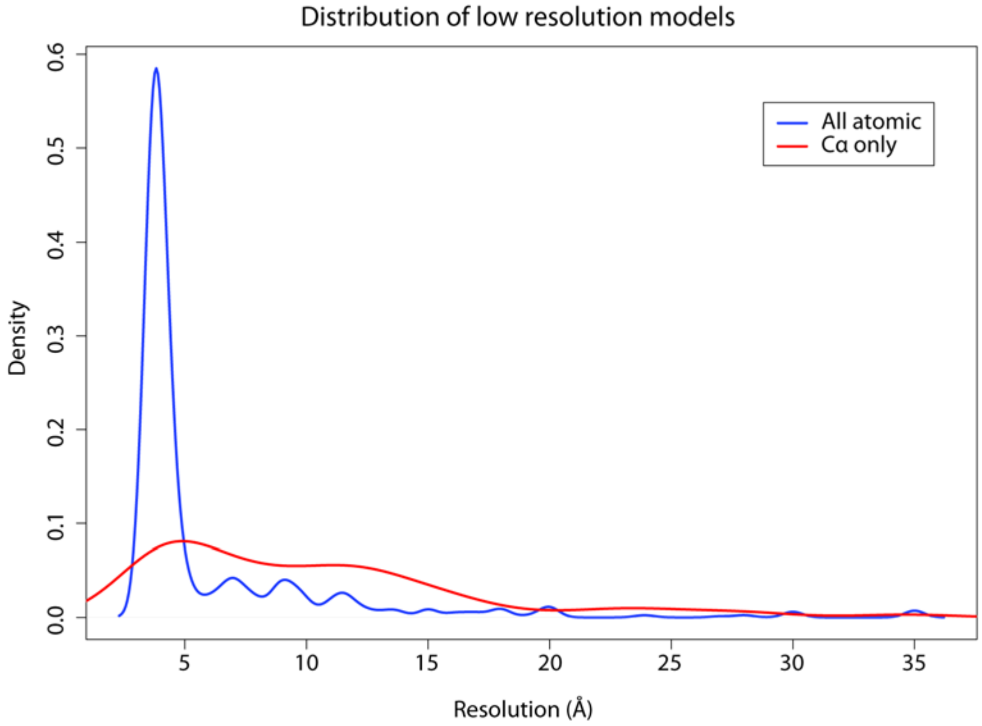

Figure B.1. Distribution of structural models in the PDB with resolution lower than 3.5 Å. While models with all atomic details (shown in blue) are mostly clustered on the higher end of the resolution spectrum, structural models with coordinates of Cα atoms only, in comparison, are more likely to be populated on a wide range of resolutions, especially towards the lower end.

B.1 Hierarchical definition of interfacial residues:

To identify the optimal ξ value, we took a non-redundant set of 4248 protein dimers determined at atomic resolution, and selected all residues that had their Cα atom within 15Å of at least another Cα atom of a residue belonging to the binding partner. All pairs of residues, one from each chain, in each dimer then comprise our total set. We can compute the Matthews correlation coefficient (MCC) for each given ξ value,

$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{B.1}$$

where true positives (TP) represent the number of residues that are in contact based on the heavy atoms within 4.5Å definition and "predicted" to be in contact based on the residue-specific Cα-Cα distance cutoff criterion with a given ξ value. False positives (FP) represent the number of residues that are not in contact based on the side chain definition but "predicted" to be in contact based on the backbone criterion. True negatives (TN) represent the number of residues that are not in contact based on the side chain criterion and also "predicted" to be not in contact based on the backbone criterion. False negatives (FN) represent the number of residues that are not in contact based on the side chain criterion but are "predicted" to be in contact based on the backbone criterion. Thus MCC gives a quantitative measure of how well the two definitions of contacting residues match for a given ξ value. We varied the ξ value, which stands for the fraction of the standard deviation to be added to the mean value, from -1.0 to 2.0, and chose the ξ value that maximizes the MCC with the side chain distance criterion. In addition, we also tested a range of generic Cα-Cα distance cutoffs (from 6Å to 12Å) that are invariant for the residue type and computed their respective MCC with the side chain distance criterion, in order to compare with the performance of our amino acid type-specific criterion.

In the range of ξ values we calculated in determining a good type-specific Cα-Cα distance cutoff, we found a ξ value of 0.5 gave the highest MCC value of about 0.48 (Figure B.2 in blue). In comparison, the optimal general Cα-Cα distance cutoff that is invariant to residue type, shown to be 8 Å, only yielded a MCC value of about 0.42 (Figure B.2 in red). A potential explanation for the fairly low MCC value in our residue type-specific criterion could be that, the rich repertoire of side-chain rotamers at the interface region results in large fluctuations in their Cα-Cα distances that almost cancel out any significant differences resulting from different residue types. This can be inferred from the comparable magnitudes in the standard deviation values of Cα-Cα distances across various types of residue pairs and in the difference between their mean values (Table B.1 and Table B.2). Despite the small improvement, given that the residue-type-specific Cα-Cα distance cutoff criterion matches better with the heavy-atom based criterion, we have incorporated this criterion with a ξ value of 0.5 into our hierarchical definition of contacting residues.

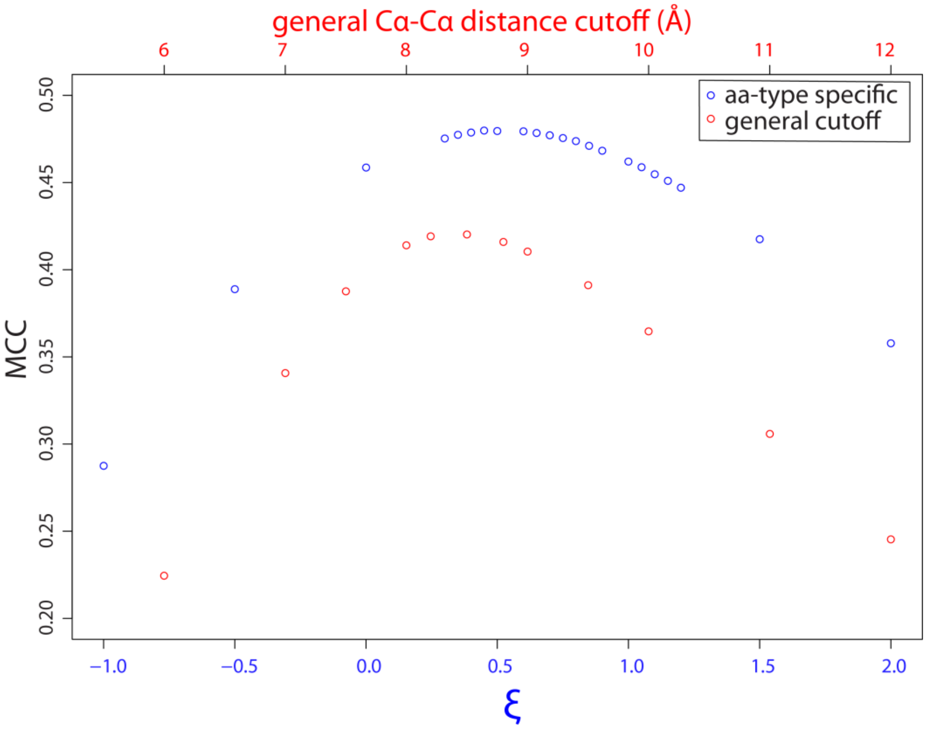

Figure B.2. Matthews Correlation Coefficient for the amino acid type-specific distance criterion and a general Cα-Cα distance cutoff criterion. The horizontal axis on top (red) gives the range of the distance cutoffs in Å we tested for the general Cα-Cα distance cutoff criterion, which resulted in a peak MCC value of 0.42 when the cutoff is chosen to be 8 Å (shown in red circles). The horizontal axis at the bottom (blue) shows the range of ξ values, which are the multiplicity factor of the standard deviation to be added to the mean value for each residue-residue type, and we obtained the highest MCC value of 0.48 when ξ = 0.5.

Table B.1. Mean Cα-Cα distances. Statistics are for a given pair of contacting residues of specific types based on statistics in Protein Data Bank, where contacting residues are defined by having at least two heavy atoms, one from each residue, that are less than 4.5 Å apart.

	A	**R**	**N**	**D**	**C**	**E**	**Q**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
A	5.53	7.99	6.42	6.24	6.03	6.76	6.97	6.91	6.74	6.97	6.96	7.14	7.59	6.10	5.70	6.27	7.89	7.68	6.36
R	7.99	10.31	8.80	9.32	8.02	9.87	9.21	9.29	8.84	8.73	9.88	8.95	9.32	8.40	8.41	8.57	9.30	9.75	8.40
N	6.42	8.80	7.31	7.46	6.65	7.95	7.79	8.12	7.33	7.61	8.37	8.02	8.23	6.83	6.82	7.01	8.78	8.71	7.09
D	6.24	9.32	7.46	7.53	6.64	8.06	8.06	8.34	7.32	7.27	8.77	7.66	8.00	6.73	6.68	6.81	8.72	8.98	6.87
C	6.03	8.02	6.65	6.64	5.93	7.09	7.69	7.97	7.27	7.43	7.58	7.59	8.08	6.64	6.37	6.54	8.48	7.99	6.74
E	6.76	9.87	7.95	8.06	7.09	8.63	8.49	8.90	7.61	7.86	9.40	8.07	8.39	7.40	7.36	7.47	9.11	9.31	7.35
Q	6.97	9.21	7.79	8.06	7.69	8.49	8.74	8.66	7.88	8.13	8.75	8.28	8.65	7.49	7.35	7.60	8.92	9.25	7.60
H	6.91	9.29	8.12	8.34	7.97	8.90	8.66	8.56	8.00	8.26	8.74	8.38	8.70	7.31	7.50	7.70	9.32	9.15	7.67
I	6.74	8.84	7.33	7.32	7.27	7.61	7.88	8.00	7.93	8.14	7.91	8.31	8.61	7.27	6.82	7.41	9.23	8.58	7.53
L	6.97	8.73	7.61	7.27	7.43	7.86	8.13	8.26	8.14	8.23	7.99	8.38	8.77	7.49	7.16	7.61	9.23	8.76	7.77
K	6.96	9.88	8.37	8.77	7.58	9.40	8.75	8.74	7.91	7.99	9.54	8.44	8.47	7.77	7.97	8.02	8.84	9.30	7.54
M	7.14	8.95	8.02	7.66	7.59	8.07	8.28	8.38	8.31	8.38	8.44	8.49	8.89	7.71	7.13	7.70	9.39	8.84	7.94
F	7.59	9.32	8.23	8.00	8.08	8.39	8.65	8.70	8.61	8.77	8.47	8.89	9.21	7.72	7.58	8.19	9.97	9.29	8.45
P	6.10	8.40	6.83	6.73	6.64	7.40	7.49	7.31	7.27	7.49	7.77	7.71	7.72	6.71	6.42	6.85	8.10	8.30	6.99
S	5.70	8.41	6.82	6.68	6.37	7.36	7.35	7.50	6.82	7.16	7.97	7.13	7.58	6.42	6.15	6.41	8.13	8.14	6.60
T	6.27	8.57	7.01	6.81	6.54	7.47	7.60	7.70	7.41	7.61	8.02	7.70	8.19	6.85	6.41	6.71	8.43	8.47	6.93
W	7.89	9.30	8.78	8.72	8.48	9.11	8.92	9.32	9.23	9.23	8.84	9.39	9.97	8.10	8.13	8.43	9.96	9.90	8.80
Y	7.68	9.75	8.71	8.98	7.99	9.31	9.25	9.15	8.58	8.76	9.30	8.84	9.29	8.30	8.14	8.47	9.90	9.37	8.40
V	6.36	8.40	7.09	6.87	6.74	7.35	7.60	7.67	7.53	7.77	7.54	7.94	8.45	6.99	6.60	6.93	8.80	8.40	7.11

Table B.2. Standard deviation of Cα-Cα distances. Statistics are for a given pair of contacting residues of specific types based on statistics in Protein Data Bank, where contacting residues are defined by having at least two heavy atoms, one from each residue, that are less than 4.5 Å apart.

	A	**R**	**N**	**D**	**C**	**E**	**Q**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
A	0.84	1.94	1.04	1.06	0.92	1.28	1.32	1.37	1.07	1.13	1.65	1.43	1.48	0.81	0.87	0.88	1.78	1.73	0.84
R	1.94	2.70	2.14	2.02	1.94	2.23	2.21	2.19	2.03	1.95	2.52	2.08	2.14	1.95	2.02	2.02	2.24	2.47	1.92
N	1.04	2.14	1.45	1.44	1.31	1.58	1.70	1.49	1.37	1.43	1.96	1.62	1.79	1.18	1.25	1.32	1.85	2.00	1.18
D	1.06	2.02	1.44	1.67	1.24	1.80	1.65	1.61	1.27	1.42	1.89	1.52	1.73	1.19	1.08	1.22	1.95	1.90	1.12
C	0.92	1.94	1.31	1.24	1.19	1.44	1.42	1.35	1.30	1.18	1.81	1.48	2.05	1.04	1.04	1.11	2.10	1.89	1.19
E	1.28	2.23	1.58	1.80	1.44	2.10	1.79	1.85	1.48	1.41	2.09	1.59	1.77	1.32	1.38	1.40	1.97	2.13	1.28
Q	1.32	2.21	1.70	1.65	1.42	1.79	1.95	1.87	1.51	1.50	2.10	1.73	1.87	1.41	1.44	1.41	1.99	2.24	1.40
H	1.37	2.19	1.49	1.61	1.35	1.85	1.87	2.09	1.63	1.49	2.04	1.81	1.75	1.40	1.47	1.50	2.14	2.19	1.42
I	1.07	2.03	1.37	1.27	1.30	1.48	1.51	1.63	1.47	1.40	1.67	1.59	1.70	1.14	1.19	1.29	1.90	1.88	1.28
L	1.13	1.95	1.43	1.42	1.18	1.41	1.50	1.49	1.40	1.51	1.63	1.62	1.83	1.24	1.28	1.25	1.96	1.81	1.23
K	1.65	2.52	1.96	1.89	1.81	2.09	2.10	2.04	1.67	1.63	2.95	1.92	1.89	1.77	1.87	1.94	2.10	2.32	1.57
M	1.43	2.08	1.62	1.52	1.48	1.59	1.73	1.81	1.59	1.62	1.92	1.92	1.93	1.45	1.59	1.56	2.15	2.06	1.57
F	1.48	2.14	1.79	1.73	2.05	1.77	1.87	1.75	1.70	1.83	1.89	1.93	2.16	1.56	1.55	1.74	2.45	2.16	1.66
P	0.81	1.95	1.18	1.19	1.04	1.32	1.41	1.40	1.14	1.24	1.77	1.45	1.56	1.17	0.98	1.01	1.61	1.77	0.99
S	0.87	2.02	1.25	1.08	1.04	1.38	1.44	1.47	1.19	1.28	1.87	1.59	1.55	0.98	1.10	1.10	1.81	1.91	1.04
T	0.88	2.02	1.32	1.22	1.11	1.40	1.41	1.50	1.29	1.25	1.94	1.56	1.74	1.01	1.10	1.25	1.92	1.93	1.10
W	1.78	2.24	1.85	1.95	2.10	1.97	1.99	2.14	1.90	1.96	2.10	2.15	2.45	1.61	1.81	1.92	2.69	2.34	1.87
Y	1.73	2.47	2.00	1.90	1.89	2.13	2.24	2.19	1.88	1.81	2.32	2.06	2.16	1.77	1.91	1.93	2.34	2.67	1.84
V	0.84	1.92	1.18	1.12	1.19	1.28	1.40	1.42	1.28	1.23	1.57	1.57	1.66	0.99	1.04	1.10	1.87	1.84	1.15

B.2 Geometric hashing:

Geometric hashing is divided into two phases; the construction phase and the voting phase. The construction phase constructs a lookup table for each structure, and is computed only once. The voting phase retrieves the two tables corresponding to the two structures being compared and finds the best transformations to apply to one of the interfaces to be overlaid with the other.

In the construction phase, a lookup table is built for an interface, which describes the projected coordinates of each point (residue) based on different reference bases. Each orthogonal basis ($\vec{x}, \vec{y}, \vec{z}$) is defined by an ordered triplet of points (residues i, j, k, which are interfacial residues that occur consecutively in their sequence order) in the following manner,

$$\vec{x} = \overrightarrow{v_i} - \overrightarrow{v_j}, \tag{B.2}$$

$$\overrightarrow{y'} = \overrightarrow{v_k} - \overrightarrow{v_j}, \tag{B.3}$$

$$\vec{z} = \vec{x} \times \overrightarrow{y'}, \tag{B.4}$$

$$\vec{y} = \vec{z} \times \vec{x}. \tag{B.5}$$

where $\overrightarrow{v_i}, \overrightarrow{v_j}$ and $\overrightarrow{v_k}$ are the coordinates of residues i, j, k respectively. The × operation refers to the cross product. Each point is then projected onto this particular reference basis to obtain its new coordinates, which are then discretized based on a given grid size (we used 4 Å in our case). The new, discretized coordinates of the point (a feature) is then recorded as a hash key pointing to the basis that leads to this specific transformation.

All possible ordered triplets are computed and registered into the lookup table based on their transformed coordinates (the features).

The voting procedure browses the two lookup tables that have been pre-computed and finds pairs of orthogonal bases that generate a sufficiently large number of features that match. As we look up each hash key (feature) in the first table, we find the matching key in the second table, and give one vote to all the pairs of orthogonal bases pointed by this feature in the two tables. Consequently pairs of orthogonal bases that receive a high vote count correspond to bases to re-orient the two structures so that they have many points that overlap. These bases are then used to compute candidate initial alignments. Our extensive tests show that this approach yields the final optimal score with sufficient sampling (Table B.3).

Table B.3. Percentage of correctly mapped interfacial residues by PCalign in quasi-equivalent protein-protein interfaces within the same capsid. In quasi-equivalent inter-subunit interfaces within a capsid, structurally equivalent residues have an exact one-to-one correspondence due to the monomers being sequence-wise identical. It is evident from these numbers that our sampling is sufficient in correctly assigning equivalence to pairs of residues for which the correspondence is known. There are, however, a few cases where not all matched pairs have the same residue ID, and they mostly occur in the comparison between A1B5 and C1C6, and between A1A2 and B1C6. These comparisons are between a “bent” interface on a five-fold symmetry axis and a “flat” interface on a six-fold symmetry axis, which are therefore less equivalent, as reflected by the lower PC-scores in the brackets. In these cases, the “mismatches” identified by the program are the consequence of maximizing the overlap of physicochemical patterns between less equivalent interfaces, rather than having resulted from inefficient sampling.

Virus name	**PDB ID**	**A1B1: B1C1**	**A1C1: B1C1**	**A1B1: A1C1**	**A1B5: C1-C6**	**A1A2: B1C2**	**A1A2: B1C6**	**B1C2: B1C6**
TNV	1c8n	100%	100%	100%	100%	100%	100%	97% (0.483)
SMV	1smv	100%	100%	100%	100%	100%	100%	100%
SBMV	4sbv	100%	100%	100%	97% (0.722)	100%	100%	100%
RYMV	1f2n	100%	100%	100%	100%	100%	100%	100%
BBV	2bbv	100%	100%	100%	91% (0.595)	100%	95% (0.597)	100%
NOV	1nov	100%	100%	100%	96% (0.558)	100%	100%	85% (0.495)
PAV	1f8v	100%	100%	100%	88% (0.433)	100%	93% (0.400)	100%
TBSV	2tbv	100%	100%	100%	100%	100%	100%	100%
NV	1ihm	100%	100%	100%	100%	100%	98% (0.737)	100%
CCMV	1cwp	100%	100%	100%	100%	100%	96% (0.545)	100%
CMV	1f15	100%	100%	100%	100%	-	-	100%
TYMV	1auy	97% (0.722)	100%	100%	100%	100%	92% (0.715)	100%
DYMV	1ddl	100%	94% (0.620)	100%	100%	100%	100%	100%
PhMV	1qjz	100%	100%	100%	100%	94% (0.695)	97% (0.569)	97% (0.698)
						B1B2: A1C6	**B1B2: C1A2**	**A1C6: C1A2**
GA	1gav	100%	94% (0.748)	100%	100%	100%	100%	100%
FR	1frs	100%	100%	100%	100%	100%	-	91% (0.700)
MS2	2ms2	100%	100%	100%	100%	100%	-	95% (0.693)
QB	1qbe	94% (0.772)	92% (0.680)	100%	100%	100%	100%	87% (0.755)

B.3 Hungarian algorithm:

The Hungarian algorithm is used for finding maximal weight matching in bipartite graphs [175]. Briefly, the algorithm starts with some initial state, and iteratively improves the objective function until it is optimized.

Key definitions included in this algorithm are as follows:

A bipartite graph, $G(V,E)$, consists of two sets of notes, X and Y, that satisfy $V= X \cup Y$ and $X \cap Y = \emptyset$, and the set of edges $E \subseteq X \times Y$. The weights of the edges are given by $w(x,y)$ for $x \in X$, $y \in Y$. A matching is a subset of edges, $M \subseteq E$, such that $\forall v \in V$, at most one edge in M is incident upon v. The neighborhood of v, $N(v)$, is all vertices that share an edge with v, and the neighborhood of a set of vertices S, $N(S)$, is all vertices that share an edge with a vertix in S. The vertex labelling function is defined by l: $V \rightarrow R$, which is feasible if $l(x) + l(y) \geq w(x,y)$ $\forall$ $x \in X$, $y \in Y$. For $G_l(V,E_l)$, if E_l satisfies $E_l = \{(x,y)$: $l(x)+l(y) = w(x,y)\}$, then $G_l(V,E_l)$ is an equality subgraph.

Based on Kuhn-Munkres Theorem, if l is feasible and M is a perfect matching in E_l, then M is a maximum weight matching, which is what the algorithm aims to find.

The Hungarian algorithm works as such:

Initialization: $\forall y \in Y$, $l(y)=0$; $\forall x \in X$, $l(x) = \max_{y \in Y}\{w(x,y)\}$. Set $M=\emptyset$.

Iteration: while M is not perfect,

1.Choose an unmatched vertex $u \in X$, and set $S=\{u\}, T=\emptyset$.

2.If the $N_l(S)$ = T, define a slack variable, $\boldsymbol{\delta} = \mathbf{min}_{\boldsymbol{u}\in\boldsymbol{S},\boldsymbol{y}\in\boldsymbol{Y}\setminus\boldsymbol{T}}(\boldsymbol{l}(\boldsymbol{x}) + \boldsymbol{l}(\boldsymbol{y}) - \boldsymbol{w}(\boldsymbol{x},\boldsymbol{y}))$, and update labelling as such:

$$l'(v) = \begin{cases} l(v) - \delta, & v\epsilon S \\ l(v) + \delta, & v\epsilon T \\ l(v), & \text{otherwise} \end{cases} \qquad \text{(B.6)}$$

3.If the $N_l(S) \neq T$, pick $y \in N_l(S) - T$,

If y is free, augment M and go back to while loop.

If y is matched to some vertex z, extend the alternating tree by setting $S = S \cup \{\text{z}\}$, $T = T \cup \{y\}$. Go back to 2.

When the iteration stops, M is perfect, and it is the maximum weight matching we look for. In the search for structurally equivalent residues across different interfaces, we define the weight function, $w(x,y)$, by the equivalence-score (Equation 3.2) in the Methods section (3.2).

B.4 Correcting for length-dependency of raw PC-score:

Earlier studies have reported the dependency of raw alignment scores on the sizes of interfaces being compared [164]. We therefore applied a similar approach to make the scoring function independent of the interface size, by considering the mean PC-score$_{\text{raw}}$ of all random pairs whose interface sizes are ±5% of a given size. When performed on a non-redundant set of 1797 interfaces, we observed that the raw score showed exponential dependence on the interface size, as shown in Figure B.3.

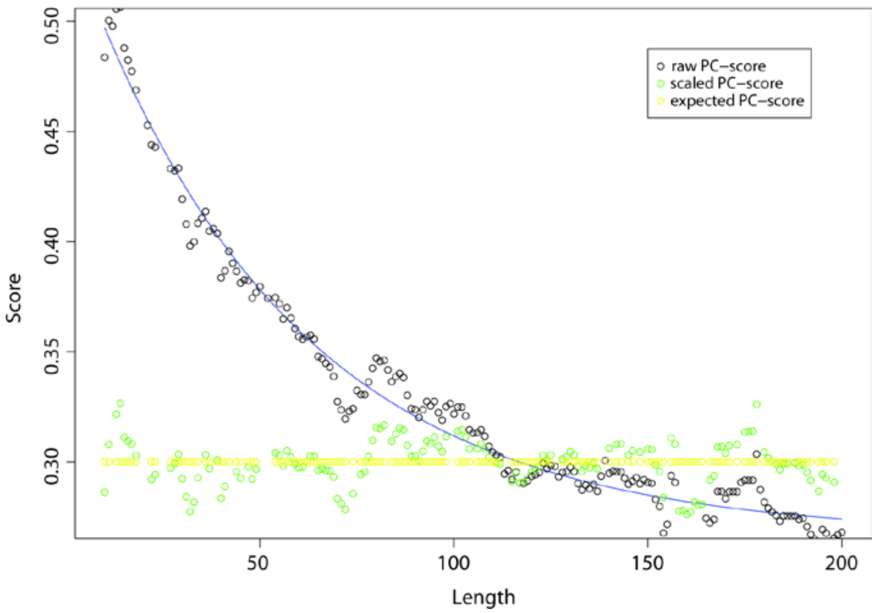

Figure B.3. The mean of PC-score$_{\text{raw}}$ across random, unrelated pairs whose interface sizes are between 95% and 105% of a given interface size. The length-corrected PC-score based on Equation 3.5 of the main text, derived from the curve fitting here, becomes 0.3 for randomly chosen interface pairs regardless of their size.

B.5 Statistical significance of PC-scores:

To derive the p-value of any PC-score, we generated the distribution of PC-scores for comparing 1,613,706 pairs of interfaces, using the same dataset as in Figure B.3. Based on the distribution of PC-scores for random alignments, shown in Figure B.4, we can estimate the p-value of a given PC-score *s* empirically by obtaining the area under the curve to the right of *s*. A few representative p-values and their corresponding PC-scores are provided in Table B.4.

B.6 Example of PCalign outperforming Ialign:

Figure B.5 gives an example where PCalign outperforms Ialign in finding better structural alignment between two unrelated interfaces. For this pair of unrelated dimers, one being a chorismate mutase and the other being a hypothetical protein with unknown function, the monomers share an overall low structural similarity, measured by a low Template Modeling-score (TM-score) [109] of 0.29132. However, PCalign recognizes significant interface similarity that is missed by Ialign, by finding structural alignment that has higher coverage and lower RMSD.

B.7 Computational costs of different interface comparison methods:

For comparison of computational time among the different methods, we collected the statistics of the running time reported by each program for the 185136 pairs we compared, excluding the one-time cost for preparing the dataset in the cases of PCalign and I2I-SiteEngine. Our performance lies between that of Ialign and I2I-SiteEngine, as shown in the boxplot of Figure B.6.

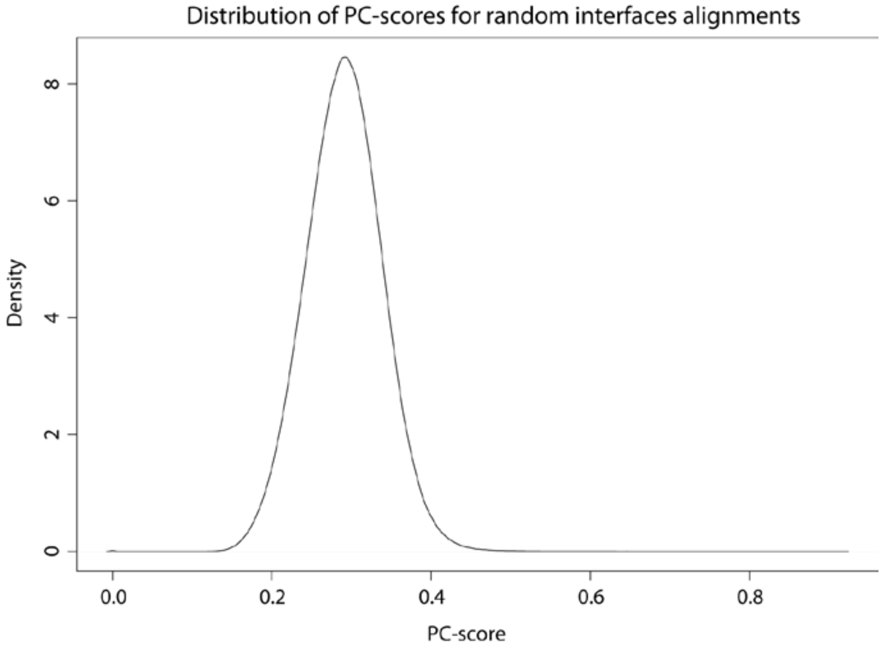

Figure B.4. The distribution of PC-scores. This is obtained by tabulating the all-against-all pairwise interface comparison for a non-redundant set of 1797 interfaces.

Table B.4. Statistical significance of PC-scores.

p-value	0.05	0.01	0.005	0.001	0.0005	0.0001	0.00001
PC-score	0.370	0.407	0.424	0.483	0.541	0.699	0.849

Table B.5. The interface equivalence measured by Q-score and PC-score for different T=3 viruses. The "/" separates the Q-score reported in [177] and the PC-score computed using our method. The two scoring functions match well with an overall correlation coefficient of 0.93, and show better agreement in some viruses than in others (last column).

Virus name*	**PDB ID**	**A1B1:B1C1**	**A1C1:B1C1**	**A1B1:A1C1**	**A1B5:C1C6**	**A1A2:B1C2**	**A1A2:B1C6**	**B1C2:B1C6**	**c.c. within virus**
TNV	1c8n	0.93/0.97	0.95/0.96	0.98/0.97	0.47/0.53	0.81/0.84	0.30/0.58	0.29/0.52	0.97
SMV	1smv	0.97/0.93	0.97/0.94	0.97/0.97	0.38/0.57	0.76/0.78	0.37/0.58	0.38/0.55	0.99
SBMV	4sbv	0.92/0.93	0.90/0.94	0.93/0.95	0.36/0.52	0.74/0.80	0.28/0.52	0.26/0.53	0.99
RYMV	1f2n	0.70/0.75	0.73/0.76	0.94/0.95	0.15/0.33	0.93/0.92	0.53/0.47	0.52/0.49	0.96
BBV	2bbv	0.90/0.88	0.93/0.94	0.90/0.90	0.33/0.42	0.88/0.90	0.54/0.45	0.53/0.47	0.97
NOV	1nov	0.82/0.93	0.91/0.93	0.86/0.96	0.40/0.41	0.87/0.86	0.41/0.47	0.44/0.47	0.98
PAV	1f8v	0.70/0.80	0.78/0.86	0.72/0.79	0.29/0.37	0.74/0.81	0.33/0.41	0.42/0.48	1.00
TBSV	2tbv	0.87/0.84	0.90/0.83	0.85/0.91	0.77/0.77	0.86/0.83	0.44/0.55	0.42/0.53	0.97
NV	1ihm	0.77/0.79	0.74/0.82	0.84/0.87	0.96/0.93	0.63/0.70	0.69/0.70	0.58/0.66	0.97
CCMV	1cwp	0.77/0.66	0.80/0.67	0.74/0.68	0.90/0.83	0.46/0.58	0.47/0.56	0.96/0.96	0.88
CMV	1f15	0.92/0.74	0.71/0.7	0.77/0.87	0.78/0.83	-	-	0.92/0.92	0.36
TYMV	1auy	0.59/0.73	0.80/0.83	0.68/0.78	0.88/0.90	0.54/0.66	0.57/0.70	0.85/0.89	0.99
DYMV	1ddl	0.79/0.81	0.56/0.65	0.71/0.76	0.93/0.92	0.72/0.82	0.78/0.79	0.90/0.89	0.97
PhMV	1qjz	0.51/0.62	0.66/0.67	0.51/0.59	0.49/0.63	0.42/0.59	0.44/0.52	0.42/0.62	0.66
						B1B2:A1C6	**B1B2:C1A2**	**A1C6:C1A2**	
GA	1gav	0.55/0.62	0.62/0.60	0.74/0.84	0.88/0.92	0.55/0.53	0.34/0.51	0.79/0.79	0.92
FR	1frs	0.52/0.62	0.57/0.62	0.74/0.73	0.86/0.92	0.52/0.61	-	0.45/0.55	0.97
MS2	2ms2	0.39/0.56	0.52/0.56	0.71/0.78	0.92/0.94	0.50/0.57	-	0.45/0.64	0.95
QB	1qbe	0.44/0.70	0.69/0.66	0.64/0.68	0.92/0.94	0.58/0.57	0.36/0.47	0.60/0.65	0.84

*Tobacco necrosis virus (TNV), Sesbania mosaic virus (SMV), Southern bean mosaic virus (SBMV), Rice yellow mottle virus (RYMV), Black beetle virus (BBV), Nodamura virus (NOV), Pariacoto virus (PAV), Tomato bushy stunt virus (TBSV), Norwalk virus (NV), Cowpea chlorotic mottle virus (CCMV), Cucumber mosaic virus (CMV), Turnip yellow mosaic virus (TYMV), Desmodium yellow mottle virus (DYMV), Physalis mottle virus (PhMV), Bacteriophage GA (GA), Bacteriophage FR (FR), Bacteriophage MS2 (MS2), Bacteriophage Q beta (QB).

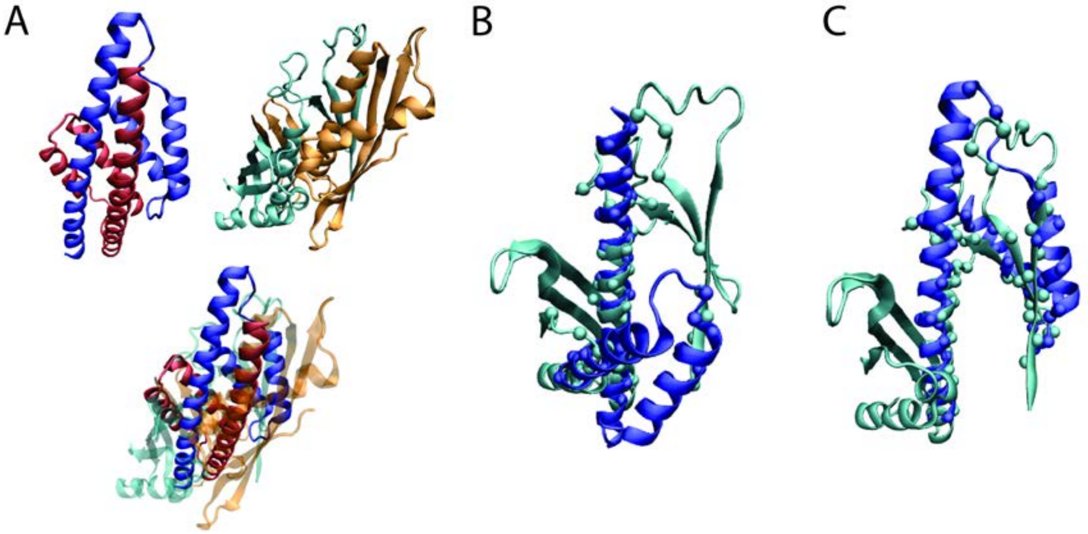

Figure B.5. An example of similar interfaces across unrelated dimers missed by existing methods but captured by PCalign. (A) The homodimer [PDB code: 2d8d] in blue and red and the homodimer [PDB: 2d7v] in cyan and orange show no global structural similarity in their monomers, but both dimers formed intertwined interfaces in a fashion that resembles two hairpins clipped to each other. (B) Ialign aligns 74 residues out of 103 interfacial residues of the smaller interface with an RMSD of 3.87Å, and the IS-score of 0.220 has a p-value > 0.05 which makes the interface similarity insignificant. (C) PCalign aligns 93 out of 103 interfacial residues of the smaller interface with an RMSD of 3.47 Å, and the PC-score of 0.418 has a p-value of 0.007, which regards the interface similarity as significantly high. This is illustrated by the interface alignment showing just one binding site, given the interface is symmetric. Figures are generated by the VMD software.

Table B.6. Protein dimers grouped by their familial annotation by SCOPPI. All dimers within the same group are less than 50% identical sequence-wise, and have the same dimerization states. The first alphabet and the first number of the SCOP classification describe the folded topology of the monomer structure, the third number representing its superfamily and the last its family. As shown in the table, the selected groups of protein dimers cover a diverse range of folds, and there are no “shared” monomers (in terms of their familial annotation) across different groups.

Grp	**PDB ID**	**Family of chain 1**	**SCOP classification**	**Family of chain 2**	**SCOP classification**
1	1all_AB	Phycocyanin-like phycobilisome proteins	a.1.1.3	Phycocyanin-like phycobilisome proteins	a.1.1.3
	1eyx_AB				
	1ha7_AB				
2	3bro_AB	MarR-like transcriptional regulators	a.4.5.28	MarR-like transcriptional regulators	a.4.5.28
	1lj9_AB				
	2a61_AB				
	2hr3_AB				
	2fxa_AB				
	1s3j_AB				
	1hsj_AB				
	1z9c_AB				
	1lnw_AB				
	2frh_AB				
	2eth_AB				
	2fbk_AB				
3	1nek_AB	Succinate dehydrogenase/ fumarate reductase flavoprotein C-terminal domain	a.7.3.1	2Fe-2S ferredoxin domains from multidomain proteins	d.15.4.2
	1e7p_AB				
	1kfy_AB				
4	2f8n_DK	Nucleosome core histones	a.22.1.1	Nucleosome core histones	a.22.1.1
	2aro_AB				
	2fj7_AB				
5	2yw7_AE	Ferritin	a.25.1.1	Ferritin	a.25.1.1

	1ji4_AD				
	2bkc_AC				
	2ux1_AC				
	1eum_AF				
	1sq3_AB				
	1vlg_AF				
	1n1q_AB				
	1nf4_AB				
	2fjc_AD				
	2htn_AB				
	1mfr_AG				
	1jre_AB				
	1zuj_AB				
	1z6o_AM				
	1tk6_AB				
6	1fg9_AC	Interferons/interleukin-10 (IL-10)	a.26.1.3	Fibronectin type III	b.1.2.1
	1j7v_LR				
	1lqs_LR				
7	2cx9_AB	Medium chain acyl-CoA dehydrogenase-like, C-terminal domain	a.29.3.1	Medium chain acyl-CoA dehydrogenase, NM (N-terminal and middle) domains)	e.6.1.1
	2reh_AB				
	1t9g_AB				
	1ivh_AB				
	1jqi_AB				
	1rx0_AC				
8	2gzu_AB	SinR domain-like	a.35.1.3	SinR domain-like	a.35.1.3
	1zzc_AB				
	1y7y_AB				
	2b5a_AB				
	2ofy_AB				

9	2mys_AC	Calmodulin-like	a.39.1.5	Motor proteins	c.37.1.9
	1w7i_AB				
	1br1_AB				
	1l2o_AC				
10	1ea4_AB	CopG-like	a.43.1.3	CopG-like	a.43.1.3
	1x93_AB				
	1p94_AB				
11	1v2a_AB	Glutathione S-transferase (GST), C-terminal domain	a.45.1.1	Glutathione S-transferase (GST), N-terminal domain	c.47.1.5
	1tu7_AB				
	1pn9_AB				
	1tw9_AB				
	2aaw_AC				
	1y6e_AB				
	2c8u_AB				
	1pl2_AB				
	1zgn_AB				
	1m0u_AB				
	2ab6_AB				
	1gul_AB				
	1gwc_BC				
	1oyj_AB				
	1k0c_AB				
	1v40_AD				
	1axd_AB				
	2pmt_AB				
	1n2a_AB				
	1ljr_AB				
12	2elc_AD	Nucleoside phosphorylase/phosphoribosyltra	a.46.2.1	(Nucleoside phosphorylase/phosphoribosyl	a.46.2.1
	1kgz_AB				

	2gvq_AD	nsferase N-terminal domain		transferase N-terminal domain	
	1brw_AB				
	1azy_AB				
13	1p51_AB	Prokaryotic DNA-bending protein)	a.55.1.1	Prokaryotic DNA-bending protein	a.55.1.1
	1b8z_AB				
	1exe_AB				
	2o97_AB				
	1ihf_AB				
14	2a1j_AB	Hef domain-like	a.60.2.5	Hef domain-like	a.60.2.5
	1x2i_AB				
	2aq0_AB				
	1z00_AB				
15	1f5q_AB	Cyclin	a.74.1.1	Protein kinases, catalytic subunit	d.144.1.7
	1g3n_AC				
	1xo2_AB				
	2f2c_AB				
	1fvv_AB				
	1w98_AB				
16	2nz8_AB	DBL homology domain (DH-domain)	a.87.1.1	G proteins	c.37.1.8
	1foe_AB				
	1ki1_AB				
	1lb1_AB				
	2dfk_AB				
	1x86_AB				
17	2pv7_AB	TyrA dimerization domain-like	a.100.1.12	TyrA dimerization domain-like	a.100.1.12
	2g5c_AC				
	2f1k_AB				
18	1tnb_AB	Protein prenyltransferases	a.102.4.3	Protein prenylyltransferase	a.118.6.1
	1ltx_AB				

	1sa5_AB				
19	1aj8_AB	Citrate synthase	a.103.1.1	Citrate synthase	a.103.1.1
	1ixe_AB				
	1o7x_AB				
	4cts_AB				
	1nxg_AB				
20	1dd4_AB	Ribosomal protein L7/12, oligomerisation (N-terminal) domain	a.108.1.1	Ribosomal protein L7/12, oligomerisation (N-terminal) domain	a.108.1.1
	1dd3_AB				
21	1s9d_AE	Sec7 domain	a.118.3.1	G proteins	c.37.1.8
	1re0_AB				
	1r8q_AE				
22	1zk8_AB	Tetracyclin repressor-like, C-terminal domain	a.121.1.1	Tetracyclin repressor-like, C-terminal domain	a.121.1.1
	2gfn_AB				
	2hku_AB				
	1z0x_AB				
	1vi0_AB				
	2g3b_AB				
	1rpw_AB				
	2zoz_AB				
	1rkt_AB				
	1t33_AB				
	3loc_AB				
	1bjy_AB				
23	2gpp_AB	Nuclear receptor ligand-binding domain	a.123.1.1	Nuclear receptor ligand-binding domain	a.123.1.1
	1yy4_AB				
	1h9u_AB				
	1uhl_AB				
	1m7w_AB				
	1xls_AE				

	1fm6_AD				
	1r1k_AD				
	1xdk_AB				
	1zeo_AB				
	1ot7_AB				
	1p8d_AB				
24	1gk2_AD	HAL/PAL-like	a.127.1.2	HAL/PAL-like	a.127.1.2
	1w27_AB				
	1t6p_AC				
25	2q80_AB	Isoprenyldiphosphate synthases	a.128.1.1	Isoprenyldiphosphate synthases)	a.128.1.1
	1v4e_AB				
	1rtr_AB				
	1rqi_AB				
26	2d8d_AB	Dimericchorismatemutase	a.130.1.1	Dimericchorismatemutase	a.130.1.1
	1ecm_AB				
	2h9c_AB				
27	1wwm_AB	TENA/THI-4	a.132.1.3	TENA/THI-4	a.132.1.3
	1rtw_AB				
	2gm8_AB				
	1yaf_AC				
	1udd_AB				
	1z72_AB				
28	1cl5_AB	Vertebrate phospholipase A2	a.133.1.2	Vertebrate phospholipase A2	a.133.1.2
	1y38_AB				
	1oyf_AB				
29	2ijc_AF	Atu0492-like	a.152.1.3	Atu0492-like	a.152.1.3
	2gmy_AB				
	2oyo_AB				
	2prr_AB				

	2pfx_AB				
30	1fqv_AB	Skp1 dimerisation domain-like	a.157.1.1	F-box domain	a.158.1.1
	1nex_AB				
	1p22_AB				
	2ovr_AB				
31	1y74_AD	L27 domain	a.194.1.1	L27 domain	a.194.1.1
	1vf6_AC				
	1y76_AD				
	1rso_AB				
	1zl8_AB				
32	2qgs_AB	HD domain	a.211.1.1	(HD domain	a.211.1.1
	2pjq_BC				
	3dto_AB				
	3djb_AB				
33	2jr2_AB	YejL-like	a.284.1.1	YejL-like	a.284.1.1
	2juz_AB				
	2jrx_AB				
34	1n0l_AB	Pilus chaperone	b.1.11.1	Pilus subunits	b.2.3.2
	2uy7_AB				
	1pdk_AB				
	2j2z_AB				
	1ze3_CH				
	1p5u_AB				
	2co6_AB				
35	1fft_AB	Periplasmic domain of cytochrome c oxidase subunit II	b.6.1.2	Cytochrome c oxidase subunit I-like	f.24.1.1
	1ehk_AB				
	3dtu_AB				
	2dys_AB				
36	1fi8_AC	Ecotin, trypsin inhibitor	b.16.1.1	Eukaryotic proteases	b.47.1.2

	1id5_HI 1azz_AC 1n8o_BE 1xx9_AC 1p0s_HE 1ezu_BC				
37	1uxe_AB 1qiu_AB 1nob_AB	Adenovirus fiber protein "knob" domain	b.21.1.1	Adenovirus fiber protein "knob" domain	b.21.1.1
38	1xu2_AR 1oqd_AK 1oqe_AK 1xu1_AR	TNF-like	b.22.1.1	BAFF receptor-like	g.24.1.2
39	1ukg_AB 2dvg_AB 1qmo_AE 1qot_AB 1avb_AB 1fat_AB 2ltn_AC	Legume lectins	b.29.1.1	Legume lectins	b.29.1.1
40	1n9e_AB 2oov_AB 1ivu_AB 1qal_AB 1ksi_AB	Amine oxidase catalytic domain	b.30.2.1	Amine oxidase catalytic domain	b.30.2.1
41	2b24_AB 1uli_AB 1ndo_AB	Ring hydroxylating alpha subunit ISP domain	b.33.1.2	Ring hydroxylating alpha subunit catalytic domain	d.129.3.3
42	2cz0_AB	Nitrile hydratase beta chain	b.34.4.4	Nitrile hydratase alpha chain	d.149.1.1

	1v29_AB				
	1ugr_AB				
43	1hx5_AB	GroES	b.35.1.1	GroES	b.35.1.1
	1wnr_AB				
	1pf9_OP				
44	1n9s_AB	Sm motif of small nuclear ribonucleoproteins, SNRNP	b.38.1.1	Sm motif of small nuclear ribonucleoproteins, SNRNP	b.38.1.1
	1i8f_AB				
	1d3b_AB				
	1b34_AB				
	1m8v_AB				
	1i5l_AB				
	1i4k_AB				
45	1lt5_DE	Bacterial AB5 toxins, B-subunits	b.40.2.1	Bacterial AB5 toxins, B-subunits	b.40.2.1
	1qb5_DE				
46	1ggp_AB	Ricin B-like	b.42.2.1	Plant cytotoxins	d.165.1.1
	2q3n_AB				
	2mll_AB				
	1hwn_AB				
	2aai_AB				
47	1usc_AB	NADH:FMN oxidoreductase-like	b.45.1.2	NADH:FMN oxidoreductase-like	b.45.1.2
	1rz1_AB				
	1i0r_AB				
48	1hkw_AB	Eukaryotic ODC-like	b.49.2.3	Alanine racemase-like, N-terminal domain	c.1.6.1
	1qu4_AB				
	1twi_AB				
49	2rsp_AB	Retroviral protease (retropepsin)	b.50.1.1	Retroviral protease (retropepsin)	b.50.1.1
	1ivp_AB				
	2p3b_AB				
	1ec2_AB				

	3fiv_AB				
50	1ytf_CD	Transcription factor IIA (TFIIA), beta-barrel domain	b.56.1.1	Transcription factor IIA (TFIIA), beta-barrel domain	b.56.1.1
	1nvp_CD				
	1nh2_CD				
51	2pbk_AB	Herpes virus serine proteinase, assemblin	b.57.1.1	Herpes virus serine proteinase, assemblin	b.57.1.1
	1o6e_AB				
	1at3_AB				
	1id4_AB				
	1nju_AB				
52	2pa7_AB	dTDP-sugar isomerase	b.82.1.1	dTDP-sugar isomerase	b.82.1.1
	1pm7_AB				
	1ofn_AB				
	1dzr_AB				
	1wlt_AB				
	2ixl_AB				
53	1f7p_AB	dUTPase-like	b.85.4.1	dUTPase-like	b.85.4.1
	1q5h_AB				
	1smc_AB				
54	1gkq_AD	Hydantoinase (dihydropyrimidinase)	b.92.1.3	Hydantoinase (dihydropyrimidinase)	b.92.1.3
	1k1d_AB				
	1gkr_AB				
	2fvm_AC				
55	2vq0_AB	Tombusviridae-like VP	b.121.4.7	Tombusviridae-like VP	b.121.4.7
	1c8n_AB				
	4sbv_AB				
	1f2n_AB				
56	2nad_AB	(Formate/glycerate dehydrogenases, NAD-domain	c.2.1.4	Formate/glycerate dehydrogenases, NAD-domain	c.2.1.4
	1j49_AB				
	1qp8_AB				

	1ygy_AB				
	1psd_AB				
	1gdh_AB				
57	1z7x_WX	28-residue LRR	c.10.1.1	Ribonuclease A-like	d.5.1.1
	1a4y_AB				
	2bex_AC				
58	1yg8_AG	Clp protease, ClpP subunit	c.14.1.1	Clp protease, ClpP subunit	c.14.1.1
	2ce3_AG				
	1y7o_AG				
	1tg6_AG				
	2f6i_AG				
59	1lqm_AB	Uracil-DNA glycosylase	c.18.1.1	(Uracil-DNA glycosylase inhibitor protein	d.17.5.1
	2j8x_AB				
	1udi_EI				
60	2c57_AB	Type II 3-dehydroquinate dehydratase	c.23.13.1	Type II 3-dehydroquinate dehydratase	c.23.13.1
	1gqo_AB				
	1v1j_AB				
61	1ufv_AB	Pantothenatesynthetase (Pantoate-beta-alanine ligase, PanC))	c.26.1.4	Pantothenatesynthetase (Pantoate-beta-alanine ligase, PanC)	c.26.1.4
	1iho_AB				
	3coy_AB				
62	1zpd_AB	Pyruvate oxidase and decarboxylase Pyr module	c.36.1.5	Pyruvate oxidase and decarboxylase PP module	c.36.1.9
	1pyd_AB				
	1ozh_AB				
	1ovm_AB				
	2c31_AB				
	1upb_AB				
	1n0h_AB				
	1mcz_AB				
63	1hoo_AB	Nitrogenase iron protein-like	c.37.1.10	Nitrogenase iron protein-like	c.37.1.10

	1lny_AB				
	1dj3_AB				
64	2tec_EI	Subtilases	c.41.1.1	CI-2 family of serine protease inhibitors	d.40.1.1
	1mee_AI				
	1y3f_EI				
65	1nw2_AD	Thioltransferase	c.47.1.1	Thioltransferase	c.47.1.1
	1xwb_CD				
	1ep8_AB				
	1f9m_AB				
66	1a49_AB	Pyruvate kinase, C-terminal domain	c.49.1.1	Pyruvate kinase, C-terminal domain	c.49.1.1
	1pky_AB				
	1pkl_AB				
67	1g5c_AB	beta-carbonic anhydrase, cab	c.53.2.1	beta-carbonic anhydrase, cab	c.53.2.1
	2esf_AB				
	1ekj_AB				
68	2nrh_AB	CoaX-like	c.55.1.13	CoaX-like	c.55.1.13
	2f9w_AB				
	3bf1_AB				
69	1je1_AD	Purine and uridinephosphorylases	c.56.2.1	Purine and uridinephosphorylases	c.56.2.1
	1k9s_AD				
	1odi_AB				
	1nw4_AB				
	1tgv_AB				
	1ybf_AB				
	1t8w_AB				
70	1bxg_AB	Aminoacid dehydrogenases	c.58.1.1	Aminoacid dehydrogenases	c.58.1.1
	1leh_AB				
	1bvu_AF				
	1v9l_AF				

	1nr7_AE				
71	1bq3_AB	Cofactor-dependent phosphoglycerate mutase	c.60.1.1	Cofactor-dependent phosphoglycerate mutase	c.60.1.1
	2a9j_AB				
	1rii_AB				
72	1a96_AB	Phosphoribosyltransferases (PRTases)	c.61.1.1	Phosphoribosyltransferases (PRTases)	c.61.1.1
	1vdm_AB				
	1pzm_AB				
	1hgx_AB				
	1grv_AB				
	1tc1_AB				
	1hmp_AB				
	1cjb_AB				
	1qk4_AB				
73	1iy9_AD	Spermidine synthase	c.66.1.17	Spermidine synthase	c.66.1.17
	2o0l_AB				
	1mjf_AB				
	1uir_AB				
	2b2c_AB				
	2q41_AD				
	1jq3_AD				
74	1cs1_AD	Cystathionine synthase-like	c.67.1.3	Cystathionine synthase-like	c.67.1.3
	1gc0_AB				
	1n8p_AD				
	1i43_AD				
75	1h3m_AB	Cytidylytransferase	c.68.1.13	Cytidylytransferase	c.68.1.13
	1vpa_AB				
	1vgw_AB				
76	2abq_AB	Ribokinase-like	c.72.1.1	Ribokinase-like	c.72.1.1
	2awd_AB				

	2ajr_AB				
	2dcn_AB				
	1tz3_AB				
	1gqt_AB				
	1v19_AB				
	2afb_AB				
	1vm7_AB				
77	2bri_AB	PyrH-like	c.73.1.3	PyrH-like	c.73.1.3
	2brx_AB				
	2v4y_AF				
78	1gc8_AB	Dimericisocitrate&isopropylmalate dehydrogenases	c.77.1.1	Dimericisocitrate&isopropylmalate dehydrogenases	c.77.1.1
	1wpw_AB				
	1a05_AB				
	1t0l_AB				
	2g4o_AB				
	1hqs_AB				
79	2bht_AB	(Tryptophan synthase beta subunit-like PLP-dependent enzymes	c.79.1.1	Tryptophan synthase beta subunit-like PLP-dependent enzymes	c.79.1.1
	1o58_AB				
	1oas_AB				
	1pwh_AB				
	1tzm_AB				
	1ve5_AD				
	1j0b_AB				
	1v7c_AB				
	1m54_AB				
	1wdw_BD				
	1wkv_AB				
	1e5x_AB				
80	1viv_AB	mono-SIS domain	c.80.1.3	mono-SIS domain	c.80.1.3

	1vim_AB 3bjz_AB 1x94_AB 1tk9_AD				
81	1vle_MN 1kqg_AB 1h0h_AB	Formate dehydrogenase/DMSO reductase, domains 1-3	c.81.1.1	Ferredoxin domains from multidomain proteins	d.58.1.5
82	1ad3_AB 1wnd_AD 1euh_AB 1ez0_AD 1o02_AB 1uzb_AB 1bpw_AB	ALDH-like	c.82.1.1	ALDH-like	c.82.1.1
83	1vgv_AB 1v4v_AB 1o6c_AB	UDP-N-acetylglucosamine 2-epimerase	c.87.1.3	UDP-N-acetylglucosamine 2-epimerase	c.87.1.3
84	1jja_AC 1hfw_AC 2ocd_AB 1djo_AB 2d6f_AB	Glutaminase/Asparaginase	c.88.1.1	Glutaminase/Asparaginase	c.88.1.1
85	1jhz_AB 1lbi_AB 1sxi_AD 1byk_AB	L-arabinose binding protein-like	c.93.1.1	L-arabinose binding protein-like	c.93.1.1
86	1mm6_AB 2a5t_AB 1y1m_AB	Phosphate binding protein-like)	c.94.1.1	Phosphate binding protein-like	c.94.1.1

87	1u0m_AB	Chalcone synthase-like	c.95.1.2	Chalcone synthase-like	c.95.1.2
	1ub7_AB				
	1xpk_AB				
	2eft_AB				
	1tee_AB				
	1mzj_AB				
	1u0u_AB				
	2qnz_AB				
88	1ux1_AD	Cytidine deaminase	c.97.1.1	Cytidine deaminase	c.97.1.1
	2z3h_AD				
	2fr6_AD				
	1r5t_AC				
89	1vhy_AB	YggJ C-terminal domain-like	c.116.1.5	YggJ C-terminal domain-like	c.116.1.5
	1v6z_AB				
	1vhk_AB				
90	2h1o_AB	PIN domain	c.120.1.1	PIN domain	c.120.1.1
	1v8o_AB				
	1v96_AB				
91	1vgq_AB	CoA-transferase family III (CaiB/BaiF)	c.123.1.1	CoA-transferase family III (CaiB/BaiF)	c.123.1.1
	1x74_AB				
	1xa3_AB				
92	1poi_AC	CoA transferase alpha subunit-like	c.124.1.2	CoA transferase alpha subunit-like	c.124.1.2
	1ope_AB				
	1k6d_AB				
	2ahu_AD				
93	2g4d_AB	Adenain-like	d.3.1.7	Ubiquitin-related	d.15.1.1
	2bkr_AB				
	2ckh_AB				
	1euv_AB				

94	2z7a_AB	Ketosteroidisomerase-like	d.17.4.3	Ketosteroidisomerase-like	d.17.4.3
	1ohs_AB				
	1cqs_AB				
95	1qy9_AB	PhzC/PhzF-like	d.21.1.2	PhzC/PhzF-like	d.21.1.2
	1u1w_AB				
	1u0k_AB				
96	2i7r_AB	Antibiotic resistance proteins	d.32.1.2	Antibiotic resistance proteins	d.32.1.2
	2a4x_AB				
	1ecs_AB				
	2pjs_AB				
	1xrk_AB				
	1r9c_AB				
	1nki_AB				
97	2h4u_AB	PaaI/YdiI-like	d.38.1.5	PaaI/YdiI-like	d.38.1.5
	1wm6_AC				
	2f3x_AB				
	1t82_AB				
	1yoc_AB				
	2fs2_AB				
	1sh8_AB				
	1sbk_AC				
	1q4s_AB				
	2ov9_AB				
98	1ues_AB	Fe,Mn superoxide dismutase (SOD), C-terminal domain	d.44.1.1	Fe,Mn superoxide dismutase (SOD), C-terminal domain	d.44.1.1
	2nyb_AB				
	1bsm_AB				
	1kkc_AB				
	1wb8_AB				
99	1tkk_AF	Enolase N-terminal domain-like	d.54.1.1	Enolase N-terminal domain-	d.54.1.1

	2dw6_AB			like	
	3muc_AB				
	2gl5_AB				
	1sjd_AB				
	1wue_AB				
	1wuf_AB				
	1yey_AB				
	1r0m_AC				
100	1r3n_AB	Bacterial exopeptidase dimerisation domain	d.58.19.1	Bacterial exopeptidase dimerisation domain	d.58.19.1
	1vgy_AB				
	1ysj_AB				
	1vix_AB				
	1cg2_AD				
	1z2l_AB				
101	2p92_AB	TM1457-like	d.64.2.1	TM1457-like	d.64.2.1
	1s12_AD				
	2g0j_AB				
	2idl_AB				
102	1nfh_AB	DNA-binding protein AlbA	d.68.6.1	DNA-binding protein AlbA	d.68.6.1
	2bky_AX				
	1udv_AB				
103	1xho_AB	Chorismate mutase	d.79.1.2	Chorismate mutase	d.79.1.2
	1com_AB				
	1ode_AB				
104	1u1i_AB	Dihydrodipicolinatereductase-like	d.81.1.3	Dihydrodipicolinatereductase-like	d.81.1.3
	1arz_AD				
	1jkf_AB				
	1p1k_AB				
	1yl7_AB				

	1vm6_AD				
	1r0l_AB				
	3dap_AB				
105	1rm6_AB	CO dehydrogenase flavoprotein C-terminal domain-like	d.87.2.1	Molybdenum cofactor-binding domain	d.133.1.1
	1t3q_BC				
	1jrp_AB				
	1ffv_BC				
106	2b67_AB	NADH oxidase/flavinreductase	d.90.1.1	NADH oxidase/flavinreductase	d.90.1.1
	1oon_AB				
	2ifa_AB				
	1f5v_AB				
	1v5y_AB				
	2fre_AB				
107	1j2g_AD	Urate oxidase (uricase)	d.96.1.4	Urate oxidase (uricase)	d.96.1.4
	1xxj_AC				
	2yzd_AB				
	1vax_AB				
	1ws3_AC				
108	1y4o_AB	Roadblock/LC7 domain	d.110.7.1	Roadblock/LC7 domain	d.110.7.1
	1veu_AB				
	1j3w_AB				
109	1vhg_AB	MutT-like	d.113.1.1	MutT-like	d.113.1.1
	1viu_AB				
	1viq_BC				
110	1syn_AB	Thymidylate synthase/dCMPhydroxymethylase	d.117.1.1	Thymidylate synthase/dCMPhydroxymethylase	d.117.1.1
	1b49_AC				
	1f28_AB				
	1bsp_AB				
	1qzf_AB				

111	1kij_AB 1mx0_AB 1qzr_AB 1s16_AB 1ei1_AB	DNA gyrase/MutL, N-terminal domain	d.122.1.2	DNA gyrase/MutL, N-terminal domain	d.122.1.2
112	1m35_AD 1pv9_AB 1chm_AB	Creatinase/aminopeptidase	d.127.1.1	Creatinase/aminopeptidase	d.127.1.1
113	2qlv_AB 2v92_AB 2ooy_AB	Ssp2 C-terminal domain-like	d.129.6.2	AMPKBI-like	d.353.1.1
114	2gac_AB 1k2x_AB 1apz_AB	(Glycosyl)asparaginase	d.153.1.5	(Glycosyl)asparaginase	d.153.1.5
115	1sb2_AB 1ukm_AB 1x2w_AB 1fvu_AB 1uex_AB 1oz7_AB 1v4l_AB	C-type lectin domain	d.169.1.1	C-type lectin domain	d.169.1.1
116	1inn_AB 1j6w_AB 1j6x_AB	Autoinducer-2 production protein LuxS	d.185.1.2	Autoinducer-2 production protein LuxS	d.185.1.2
117	2pkh_AB 3bwg_AB 2ikk_AB 2ooi_AB 3ddv_AB	UTRA domain	d.190.1.2	UTRA domain	d.190.1.2

	2fa1_AB				
	2p19_AC				
	3cnv_AB				
118	2onf_AB	Ohr/OsmC resistance proteins	d.227.1.1	Ohr/OsmC resistance proteins	d.227.1.1
	1ukk_AB				
	1qwi_AB				
	1lql_AB				
	1n2f_AB				
	2d7v_AB				
	2opl_AB				
119	1twj_AB	PurS subunit of FGAM synthetase	d.284.1.1	PurS subunit of FGAM synthetase	d.284.1.1
	1gtd_AB				
	1vq3_AB				
120	1si8_AC	Heme-dependent catalases	e.5.1.1	Heme-dependent catalases	e.5.1.1
	1m7s_AC				
	1a4e_AB				
	1th2_AC				
	1ggj_AC				
121	1dk4_AB	Inositol monophosphatase/fructose-1,6-bisphosphatase-like	e.7.1.1	Inositol monophosphatase/fructose-1,6-bisphosphatase-like	e.7.1.1
	1vdw_AB				
	1lbv_AB				
	1imd_AB				
122	2e75_AD	ISP transmembrane anchor	f.23.12.1	Cytochrome b of cytochrome bc1 complex (Ubiquinol-cytochrome c reductase)	f.21.1.2
	1q90_BR				
	2ibz_CE				
	2bcc_CE				
123	2axt_AD	Bacterial photosystem II reaction centre, L and M subunits	f.26.1.1	Bacterial photosystem II reaction centre, L and M subunits	f.26.1.1
	2gmr_LM				
	1qov_LM				

	6prc_LM				
124	1pdg_AB	Platelet-derived growth factor-like	g.17.1.1	Platelet-derived growth factor-like	g.17.1.1
	1bj1_VW				
	1rv6_VW				
	1wq9_AB				
	1vpp_VW				
	1kat_VW				

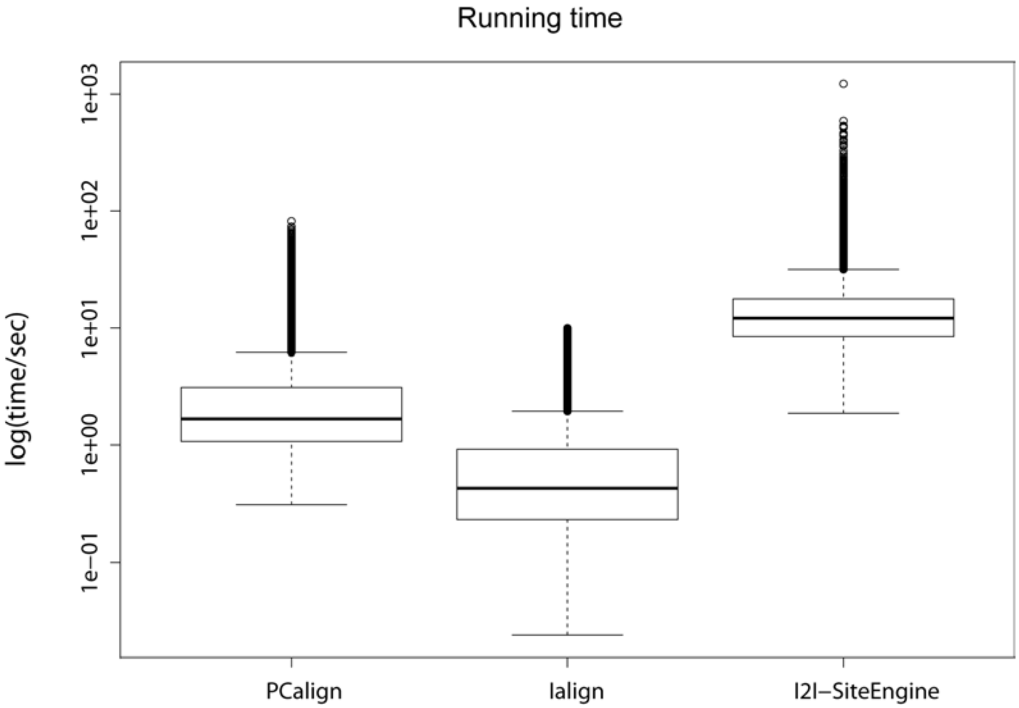

Figure B.6. **Running time reported by the three programs.** The time in seconds is plotted in log scale. PCalign finishes a comparison within a second on average, whereas Ialign completes within a fraction of a second. I2I-SiteEngine is roughly an order of magnitude slower. The longest computational time for these methods is spent on comparing (single-domained) dimers with over 200 interfacial residues, shown in circles on the very top of each box.

CHAPTER IV

Template-based protein inhibitor design for influenza hemagglutinin

4.1. Introduction:

The importance of protein-protein interactions in a wide range of cellular functions and diseases has instilled great interest in designing protein-protein interfaces for binding affinity, with the goal of modifying, creating or inhibiting protein-protein interactions. Part of the interest is intellectual; being able to predict proteins that bind with each other and with the correct interface configuration provides the ultimate test to our understanding of how proteins recognize one another to form macromolecular complexes, and the other part is the great therapeutic potential offered by designing proteins to target and interfere with disease-related interactions [202]. Recent advances in the field of affinity design have seen successes in redesigning antibodies for enhanced binding affinity [203], creation of a chimeric protein by fusing domains of different functionalities through a designed interface [204], modifying a protein that is physiologically monomeric to self-associate into stable oligomers [205, 206], and the de novo design of two protein inhibitors that bind with influenza hemagglutinin protein [113, 114]. The last work is especially of interest to us, as it demonstrates the feasibility of creating diagnostic and/or therapeutic agents to target an arbitrarily chosen protein surface.

Given the structural model of the target protein, David Baker and coworkers employed a computational design approach that entails two fundamental steps; first is the identification of hot spot residues which make highly optimized van der Waals, hydrogen bonding and electrostatic interactions with the target protein that essentially anchor the association, followed by stitching these hot spot residues onto a scaffold protein that has a patch which is geometrically complementary to the target protein surface [113, 114]. Out of the 88 computational designs created, two that survived experimental validation were further "evolved" experimentally to select favorable mutations to enhance binding affinity. This computational protocol is *de novo* in the sense that the scaffold proteins from which the designs were created are physiologically unrelated to the target protein of interest, and have their surfaces re-engineered to bring about binding activity that was absent with the wild type proteins.

In view of the success of the *de novo* approach in designing protein binders for a given target protein surface, we would like to test an alternative route to propose candidate proteins that may bind with a target protein, under the premise that a structural model exists for the target protein in complex with another protein at the desired binding site. This approach, which we term as template-based, is rationalized based on the observation that nature has a limited set of choices for interface designs [165], and dissimilar proteins can associate in similar fashions [207]. We illustrate our idea in Figure 4.1**,** where we screen the library of currently available structural models for protein-protein interfaces that resemble the one formed between our target protein and its native partner, and use those protein dimers with similar interfaces as a starting point for the design of potential proteins that can bind with the target. In this approach, we do not

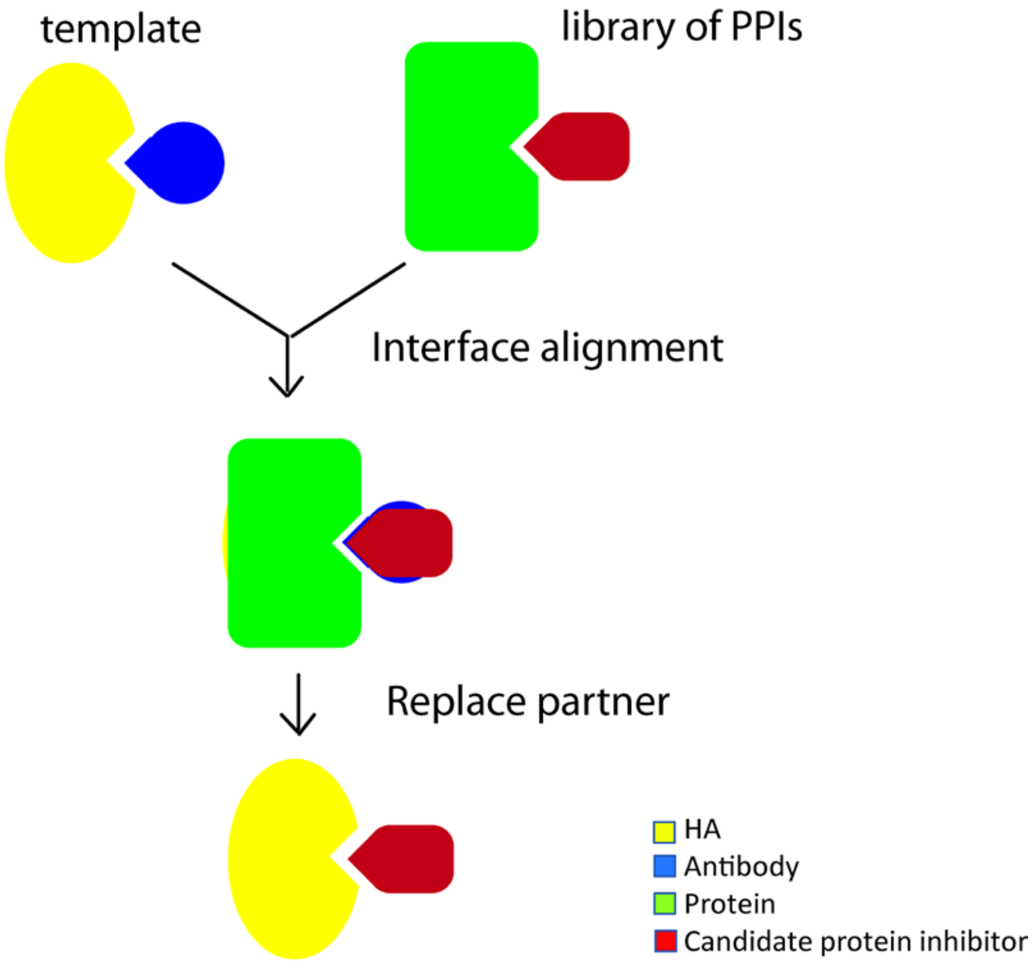

Figure 4.1. Template-based protein inhibitor design protocol. Using the structural model of the complex formed by hemagglutinin (yellow) and CR6261 (blue) as the template of a protein-protein interface, we search the Protein Data Bank for similar interfacial patterns between other proteins (green and red). Upon the identification of such dimers, we then replace the antibody (blue) by the protein that has an equivalent binding site as the paratope (red), based on the structural alignment of the interface region, and propose this protein as a potential binder for the target viral protein.

require the proposed protein binder to share structural similarity with the native partner of the target protein, as long as there is significant local similarity at the interface region.

The system we chose to test our template-based inhibitor design is the same as that in [113, 114], specifically the hemagglutinin (HA) protein from the 1918 H1N1 influenza pandemic [208]. Influenza viruses pose a serious threat to public health, leading to 250,000-500,000 deaths annually around the globe [209]. Despite intensive efforts on therapeutic and vaccine development, limited success has been achieved so far in combating this highly adaptable virus. The HA protein is a membrane glycoprotein found on the surface of this enveloped virus. Together with another glycoprotein, neuraminidase, they define the subtypes of different influenza viruses, such as the swine flu H1N1 and the avian flu H5N1. HA has been well characterized as a primary drug target [210], given its essential role in the earlier stage of viral infection- from binding of influenza viruses to their sialylated cell-surface receptors, to mediating the transfer of viral generic material into the cytoplasm via membrane fusion with the endosome. Being the major surface antigen recognized by neutralizing antibodies, the HA rapidly undergoes antigenic variation under immune pressure to escape detection, posing a great challenge in the design of vaccines and therapeutics that can effectively target different strains of influenza viruses [211]. This bottleneck underscores the need for novel inhibitor designs that adopt alternative mechanisms to provide a broader range of protection.

The HA complex consists of a trimer of two protein subunits, HA1 and HA2, which are disulfide-linked and are derived from the proteolysis of one polypeptide during maturation [212]. HA1 forms the head region (shown in Figure 4.2) which directly binds

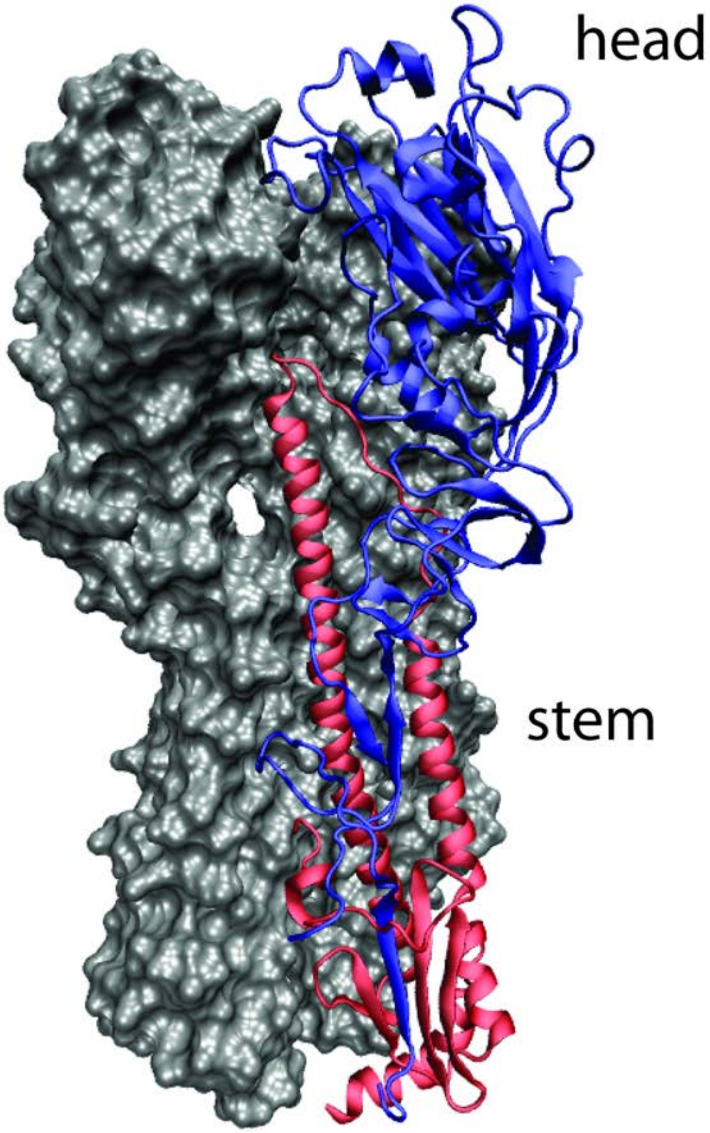

Figure 4.2. Structure of a HA trimer. Only one asymmetric unit of the three (PDB code: 3gbn) is shown in cartoon representation, and the other two are illustrated in grey surface representation. HA1 (blue) forms the globular head that recognizes and binds with cell surface receptors, while HA2 (red) forms the stem region and contains the fusion peptide that is required for membrane fusion with the endosome. Figure is generated by the software Visual Molecular Dynamics (VMD) [184].

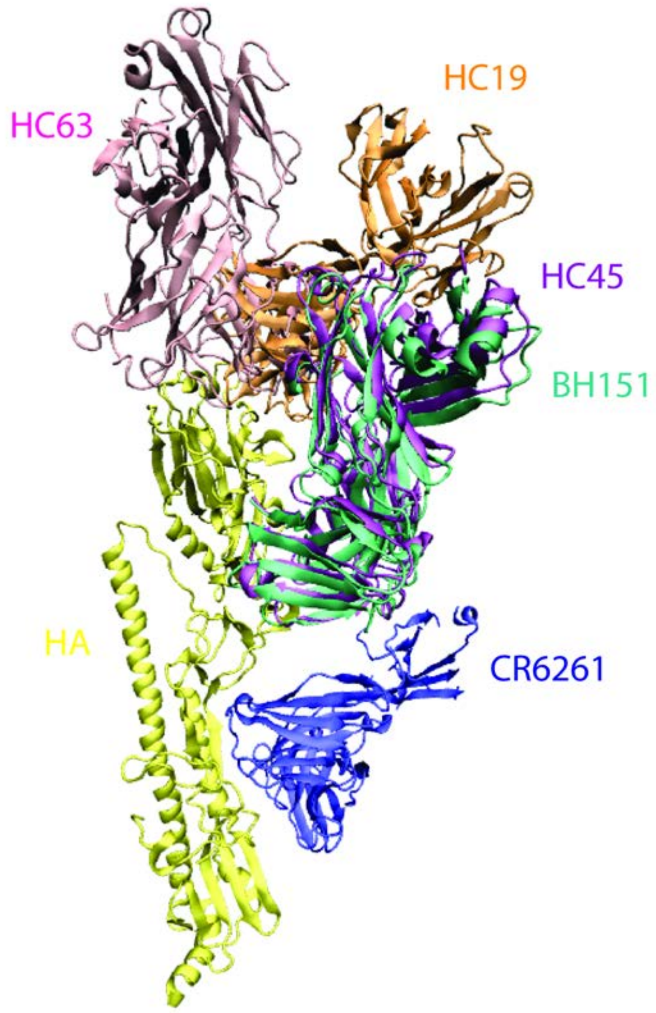

Figure 4.3. Antibodies targeting HA. Different antibodies target different regions of HA (yellow), with most of them neutralizing the hypervariable globular head region, shown here overlaid with one another by aligning the HA structure. The head-bound antibodies include BH151 (PDB code: 1eo8, shown in green), HC63 (PDB code: 1ken, shown in pink), HC45 (PDB code: 1qfu, shown in purple) and HC19 (PDB code: 2vir, shown in orange). The broadly neutralizing antibody, CR6261 (PDB code: 3gbn), shown in blue, binds instead to the stem region that is highly conserved across different strains of influenza.

to cellular receptors for endocytosis, while HA2 comprises the stem region, with the terminal arm forming the fusion peptide that is responsible for initiating membrane fusion upon entry. Inside the lowered pH environment of the endosome, HA undergoes a large conformational change that allows the insertion of the fusion peptide into the endosome membrane, and subsequently the viral genetic material gets released into the cytoplasm [213]. A majority of the neutralizing antibodies that bind with HA recognize the highly variable head region [214-217], thereby physically blocking the interaction between the head region and the cell surface receptors. These antibodies typically bind to different patches on the head domain (shown in Figure 4.3) and are strain-specific, offering suboptimal protection with limited breadth. In contrast, the HA stem region is highly conserved across different strains and subtypes of influenza, and can be recognized by several broadly neutralizing antibodies [218-220], one of which has been recently co-crystallized with both the HA from the 1918 H1N1 and that from the recent H5N1 avian influenza [221]. Given the promising heterosubtypic protection offered by inhibitors targeting this epitope that is constrained from mutation, we chose the structural model of the broadly neutralizing antibody, CR6261, in complex with the HA from H1N1 (PDB ID: 3gbn) as our template to identify naturally existing protein dimers that interact in a similar fashion, with the ultimate objective of proposing novel protein inhibitors that can offer expanded therapeutic options in the clinic.

4.2. Methods:

4.2.1. Interface library construction

To construct the library of protein dimer interfaces, we downloaded 75,694 structural models from Protein Data Bank (statistics as for Sep 2011) [112], which were screened against PISA [222] to extract 165257 protein dimer interfaces assigned to be biologically significant (as opposed to due to crystal packing). 123395 dimers remain after pruning those with too small an interface (<10 residues) or too large an interface (>200 residues). These dimers are then grouped based on sequence identity of 50% (we require both chains of the protein dimers, which can be homodimers or heterodimers, within the same group to have at least 50% sequence identity), in order to select representative dimer interfaces to reduce the size of the library for screening. Since two protein dimers with similar sequences can interact in different orientations, we perform an all-against-all pairwise structural comparison of the protein-protein interfaces within each group using PCalign [223], and further cluster using complete-linkage clustering of the dimers that have interface similarities of PC-scores greater than 0.5. We then select one representative dimer from each cluster to reduce the search space, and this collection of sequence-wise non-redundant protein dimers constitute our final interface library.

4.2.2. Initial candidate selection

The complex comprising of the 1918 H1N1 influenza virus HA and Fab CR6261 (PDB code 3gbn) is used as our template structure to screen the interface library for similarly arranged geometric and chemical patterns. Using the program PCalign, we first retain dimers that have interface similarities with the target structure of PC-scores greater than 0.36. This value was shown to discriminate similar interfaces against different ones based on our previous study [223]. We then apply a size cutoff between 50 and 100 amino acids to the protein chains that have equivalent binding sites as the antibody, as we believe peptides that are too small are prone to degradation while large bulky proteins may diffuse too slowly to associate effectively with the HA. Because our interface similarity measure takes into account both binding fragments, we extract from each dimer the chain that has an equivalent binding site as the antibody (we call it the candidate protein), and make a chimera complex of the HA and the candidate protein based on structural alignment of the interfaces of the two native complexes. This approach enables us to delineate the contribution to interface similarity by the candidate protein, and we can perform a re-ranking of all the candidate proteins using PC-scores between the chimera complexes and the HA-antibody complex.

4.2.3. Criteria for evaluating likelihood of binding activity

For any predicted candidate protein, we use three independent criteria to assess its binding potential to the target protein. First we calculate the approximate binding energy using classical molecular mechanics force field via the GBSA method [224], as in Equation 4.1:

$$\Delta G = \Delta E_{MM} + \Delta G_{solv,p} + \Delta G_{solv,np} - T\Delta S, \tag{4.1}$$

where ΔG is the binding free energy, ΔE_{MM} is the interaction energy between the target protein and the candidate protein, calculated as the enthalpy difference between the total system and the individual components, using the CHARMM force field (Equation 4.2) [225].

$$E_{MM} = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} k_\varphi(1 + \cos(n\varphi - \delta))$$
$$+ \sum_{improper} k_\omega(\omega - \omega_0)^2 + \sum_{Urey-Bradley} k_u(u - u_0)^2$$

$$+\sum_{non-bonded\ pairs}\varepsilon_{ij}\left[\left(\frac{r_{ij}^{min}}{r_{ij}}\right)^{12}-2\left(\frac{r_{ij}^{min}}{r_{ij}}\right)^{6}\right]+\sum_{non-bonded\ pairs}\frac{q_i q_j}{\epsilon r_{ij}}\,, \qquad (4.2)$$

where the first five terms account for the bond stretching, the angle bending, the torsion angles, the out of plane bending and the Urey-Bradley term which total up the bonded energy. The sixth term which is the van der Waals energy and the seventh term which is the electrostatic interactions contribute to the non-bonded energy. $\Delta G_{solv,p}$ is the electrostatic contribution to the solvation energy calculated using the Generalized Born equation, and $\Delta G_{solv,np}$ is an empirical term calculated as 0.015 kcal/Å^2 × SASA (solvent accessible surface area) to approximate the hydrophobic component of the solvation energy. The last term in Equation 4.1 accounts for the entropic cost of fixing the two proteins in the bound states with each other, and is neglected as an approximation in our calculation for reasons of the high computational cost involved. We believe this does not significantly affect our evaluation of the binding free energy comparisons across different candidate proteins with the same target protein. When only a single snapshot of a structure is available, such as the minimized configuration, the binding free energy is calculated based on that snapshot. When a simulation trajectory is available for a system, the last 10 ps of the trajectory is used to obtain an average binding free energy from all frames recorded.

The second criterion used is to gauge if blind docking analysis reveals our predicted binding configuration as favorable for any given protein candidate docked onto the target protein. For this purpose we adopt the software developed by Zhiping Weng’s group, Zdock [226], which is an initial-stage docking algorithm that explores the six degrees of freedom of individual proteins that are treated as rigid bodies. Their scoring function incorporates a pairwise statistical potential, shape complementarity, electrostatics, and desolvation in the fast fourier transform-based algorithm [227], based on which the top-scoring 2000 configurations are reported with the default parameters. Given this docking algorithm is relatively coarse-grained without the next refinement step, we also use a loose criterion to count “near-native” configurations, where we consider a configuration returned by Zdock as near-native if at least three contacts (residue-based) are captured out of all contacts from our proposed binding configuration. The fraction of near-native configurations contained in the top 2000 poses reported by Zdock then reflects how likely our candidate protein indeed associates with the target protein at the intended site.

Lastly, we also use the software FoldX to predict the binding energy of our proposed candidate proteins and the target [228]. The energy function differs from the classical molecular mechanics one described in our first criterion mainly in two aspects; one is that the different energy terms such as hydrogen bonding have been weighted empirically, and the other being the crude entropy estimation based on statistical analysis of the dihedral angles of the peptide backbone, whereas entropy is ignored in the first criterion.

4.2.4. Relaxation of initial structure

The candidate protein identified by PCalign is initially placed against the target protein based on the structural alignment of the interfaces of the template complex and of the query complex. This may unavoidably introduce steric clashes as the two proteins that are not natural interacting partners become forced together. A brief minimization of the initial configuration in implicit solvent is performed for all candidate proteins. For further relaxation to allow side chains to be repacked, a simulated annealing step described as follows is carried out.

Without any restraints applied to the system during the annealing procedure, it was observed that a candidate protein tends to quickly diffuse away from the target surface due to the sub-optimal interactions present initially to sufficiently stabilize the association. Thus the following restraints were added to keep a candidate protein within the proximity of HA. Based on the minimized configuration of the complex, a list of contacting residues is generated using the contact definition in [223]. These contacts are used for setting up distance restraints during the heating procedure, where a flat-bottom restraint defined in Equation 4.3 is applied to each pair of contacting residues to keep the two proteins from diffusing away from each other:

$$E(R) = \begin{cases} 0.5 \times k_{min} \times (R - R_{min})^2 & R < R_{min} \\ 0.0 & R_{min} < R < R_{max} \\ 0.5 \times k_{max} \times (R - R_{max})^2 & R_{max} < R < R_{lim} \\ f_{max} \times (R - (R_{lim} + R_{max})/2) & R > R_{lim} \end{cases}, \tag{4.3}$$

where *E(R)* is the biasing potential, and *R* is the distance in Å between the Cα atoms of two particular residues from the contact list at a given time point during simulation. The two spring constants, k_{min} and k_{max}, are both set to 1 kcal/mol/Å^2. Given the distance *d* between a particular pair of contacting residues in the starting configuration, R_{min} and R_{max} are set to be (*d*-2) Å and (*d*+2) Å respectively to ensure the two residues stay within close proximity within each other. f_{max} which is set to 2 kcal/mol/Å defines the final asymptote slope as the two residues drift further apart. Finally, R_{lim} is defined as R_{max} + f_{max}/k_{max} to ensure continuity of the derivative.

In addition to the distance restraints between the HA and the candidate protein, a harmonic restraint with respect to their starting positions is applied to the Cα atoms of the HA protein chains with a force constant of 5.0 kcal/mol/Å^2 to keep the target protein in place, and a harmonic restraint with respect to the starting configuration is applied to the Cα atoms of the candidate protein with a force constant of 5.0 kcal/mol/Å^2 to allow the candidate protein to explore around the binding site of the HA without permitting structural distortions. These restraints are applied to the system during the first 10 ps of the simulation when the system is heated up from 300 K to 400 K, after which the restraints are removed to allow the system to cool back down to 300 K over another 10 ps of simulation.

4.2.5. Selection of residues for redesign

Following the simulated annealing procedure, two rounds of stepwise mutation analysis are carried out to identify residues for redesign. Rather than attempting mutations for all interfacial residues, we focus on only those residues that have mismatched chemical types with their structural equivalent ones from the paratope based on the initial structural alignment generated by PCalign, and mutate them to any of the residues with matched chemical types, in order to maximize the binding free energy with minimal mutations introduced to the original candidate proteins. A few exceptions are left untouched; Cys, Pro and Gly residues are kept due to their special structural roles in proteins in general. Residues with their side chains facing away from the binding site are also left unchanged as they do not directly participate in stabilizing the complex association. Lastly, residues forming salt bridges with other residues are also not included for redesign, as introducing mutations at these positions may destabilize the candidate protein structure. The position of the Cα atom and the center of mass of the side chain are used for building the model of each mutated residue, which is further relaxed by the same annealing procedure. Following one mutation of a single residue, each mutation is evaluated by the estimated binding free energy described in Section 4.2.3, and the mutation that results in maximal improvement of binding free energy is accepted for the next round of mutation analysis. After two point mutations are introduced in each candidate protein, the final proposed binding configuration is assessed using all three criteria for evaluation.

4.3. **Results**:

4.3.1. Correlation between PC-score with the native complex and the predicted binding energy

To assess if interface similarity with the native complex may serve as a good predictor for binding activity, we take the 88 initial computational designs from [113], and compare the interface similarities between these structural models and the HA-CR6261 complex. As only two out of the 88 computational designs showed reproducible binding activity during experimental validation, we do not have the actual binding data for these 88 structural models except for the computed binding energy reported in the paper, which is what we use in this analysis. As shown in Figure 4.4, there is very limited correlation between the computed binding energy and the interface similarity with the native complex (reported by PC-score), with a correlation coefficient of -0.27. However, it is worth noting that the two designs that survived experimental validation are found to have a high PC-score, one being 0.565 and the other 0.664, implying the binding sites having high resemblance to the paratope on CR6261. This gives us confidence in choosing a good template as the starting point for computational design.

4.3.2. Top ranked predictions

Upon screening the interface library, we identified altogether 10 protein chains that have an interaction environment which highly resembles our template complex, based on a PC-score cutoff of 0.5. To judge how likely it is that these 10 protein will bind to the target site of the HA protein, we include three reference states to compare the candidates with; one being the crystal structure of HA-CR6261 (PDB code: 3gbn), and the other two being the crystal structures of the two *de novo* designs from [113, 114] (PDB codes: 4eef and 3r2x). Because the two designed inhibitors have undergone affinity maturation through experiments, we are also interested in how the initial computational designs perform using our evaluation criteria for comparison purposes, which are thus included in our analysis as well.

Table 4.1 summarizes a preliminary survey of characteristics of the predicted protein complexes, with the reference states highlighted in red. These statistics are collected based on the initial complex configurations generated using structural alignment of the interfaces only. With the exception of the antibody heavy chain that has over 200 residues, being a two-domain protein itself, the *de novo* designs and the predicted candidate proteins have about the same length, with comparable interface sizes as well.

These initial configurations of the candidate proteins placed next to the target proteins are then minimized and further relaxed by simulated annealing, which is shown to generally improve the proposed binding poses (Table 4.2). Especially when the estimated

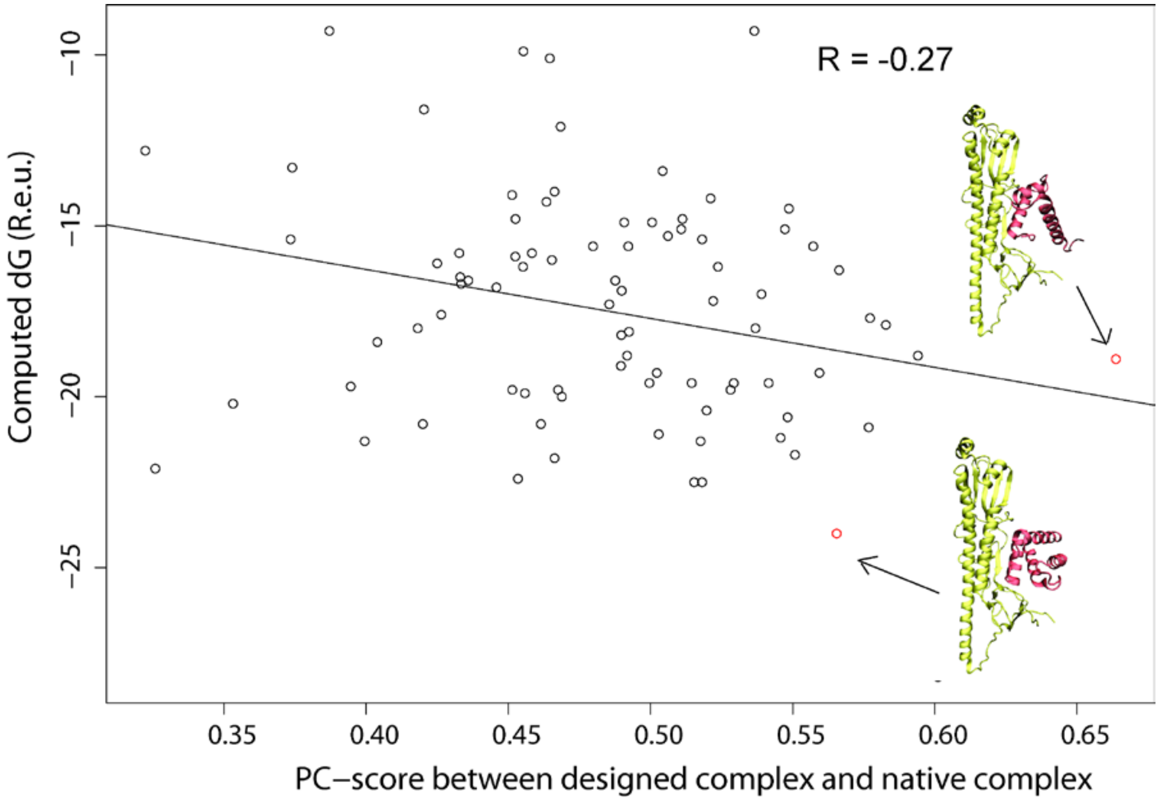

Figure 4.4. PC-scores of computational designs. Comparison between the 88 de novo computational designs with the native HA-CR6261 complex results in PC-scores that are poorly correlated with their computed energy in Rosetta energy units (R.e.u.). The two designs (HB36 and HB80) that showed actual binding activity are illustrated in red, bound to the target protein in yellow. Both scored high in terms of interface similarity with the native complex.

Table 4.1. Basic properties of candidate proteins. Compared with the three reference structures highlighted in red, our predicted protein inhibitors have about equal attributes. These include interface similarity measure with the native HA-CR6261 complex for the predicted or solved bound configuration of the candidate protein and HA, the size of the known binder or candidate protein, the interface size of the predicted or solved complex structure, and the buried surface area of the predicted or solved complex. Each protein is annotated by its PDB code and the chain identifier. The two entries in brackets, HB36 and HB80, are the original computational designs that have yet undergone experimental affinity maturation.

Binding protein	PC-score with native complex	Protein size (a.a.)	NO. of interfacial residues (a.a.)	ΔSASA (Å^2)
CR6261	1	226	36	1660
3r2xC (HB36)	0.558 (0.565)	93	37	1800
4eefG (HB80)	0.580 (0.664)	45	33	1660
1yrnA	0.535	61	28	1900
1ohzB	0.500	60	40	1800
2x9aC	0.519	65	40	1160
3ihqB	0.512	75	32	1500
1pk1B	0.561	89	22	1670
2b87A	0.525	58	38	1990
2y0sL	0.543	92	38	1860
3ossC	0.515	68	33	1840
1s1qB	0.584	76	32	2290
1qfnA	0.561	85	32	2590

binding free energy is considered, side-chain repacking led to much more favored complex configurations, rendering the results judged by the first criterion much closer to our three reference states. In comparison, in terms of blind docking analysis by Zdock and prediction by FoldX, there is still a significant gap between our predictions and the reference structures, suggesting room for further improvement.

The 10 candidate proteins are then subjected to stepwise mutation analysis. For the majority of these proteins, two point mutations appear sufficient to significantly improve the estimated binding free energies (Table 4.3), with the exception of one protein (PDB code 2y0s, chain L) for which no second mutation that brings about noticeable improvement can be identified. While not all three criteria reach a consensus regarding the effect of the mutations introduced for each candidate protein, the majority of the proteins showed improvement based on at least two of the three criteria. Remarkably, four proteins performed comparably with the reference states when evaluated by the three types of assessment, namely chain B of PDB code 1ohz, chain C of PDB code 2x9a, chain B of PDB code 1pk1, and chain A of PDB code 1qfn. It should be highlighted that although these results are only sub-optimal compared to the two *de novo* designs, which have gone through “evolution” with experiments, our predictions almost parallel those of the original computational designs from [113], with minimal mutations needed to bring about putative binding activity.

The four designs mentioned above in complex with the HA protein are illustrated in Figure 4.5. With the exception of chain C of 2x9a which is largely beta-sheet, the other proteins are mostly helical in terms of secondary structure content. Upon close examination of the candidate proteins, we see that substitutions recommended by our

Table 4.2. Structural relaxation with simulated annealing. For each of the three criteria used for evaluating the likelihood of binding, a comparison is given before simulated annealing (after minimization) and after. For majority of the candidate proteins, simulated annealing improves the proposed binding pose, demonstrated by better scoring with the evaluation criteria. Nonetheless, for blind docking analysis and energy estimated by FoldX, most of these candidate proteins scored more poorly compared to the reference states shown in red.

Binding protein	<$\Delta G_{binding}$> (kcal/mol)		Near-native poses identified by Zdock After minimization (%)		ΔG by FoldX (kcal/mol)	
	After minimization	After simulated annealing	After minimization	After simulated annealing	After minimization	After simulated annealing
CR6261	-85.1	-73.1	3.5	1.05	-14.15	-10.95
3r2xC (HB36)	-60.6 (-45.6)	-60.1 (-53.4)	6.95 (3.25)	2.15 (1.95)	-12.35 (-10.14)	-11.05 (-5.03)
4eefG (HB80)	-76.9 (-76.4)	-76.0 (-56.5)	6.9 (4.3)	8.35 (1.6)	-7.38 (-7.87)	-12.14 (-9.13)
1yrnA	-6.5	-43.4	0.35	0.55	9.62	1.38
1ohzB	-27.7	-47.3	2.05	3.85	-7.68	-6.36
2x9aC	-32.5	-34.7	0.4	0.95	-4.96	-0.43
3ihqB	-51.0	-47.2	0.8	1.4	-2.79	0.74
1pk1B	-13.8	-47.3	1.15	1.35	1.84	0.22
2b87A	4.9	-27.6	0.45	0.05	3.14	0.57
2y0sL	-42.1	-59.2	1.2	0.5	-5.68	-3.57
3ossC	-19.2	-37.8	0.25	0.95	1.97	1.54
1s1qB	-62.4	-63.5	1.2	1.05	3.10	-2.44
1qfnA	-23.2	-41.9	1	1.8	8.18	2.24

Table 4.3. Improvement with computational redesign. Each candidate protein is subjected to stepwise mutations, with the most favorable mutation retained for the next round. Each round improves over the previous one ($<\Delta G_{binding}>^1$ over $<\Delta G_{binding}>^0$, and $<\Delta G_{binding}>^2$ over $<\Delta G_{binding}>^1$). The final designs scored comparably with our reference states when evaluated by the three criteria. The last column lists the root mean squared distance (RMSD) of the final design based on Cα atoms with respect to its starting structure, and the low RMSD value suggests minimal structural distortion introduced to the candidate protein.

Binding protein	$<\Delta G_{binding}>^0$ (kcal/mol)	$<\Delta G_{binding}>^1$ (kcal/mol)	$<\Delta G_{binding}>^2$ (kcal/mol)	Zdock	ΔG by FoldX (kcal/mol)	RMSD (Å)
CR6261	-73.1	-	-	1.05	-10.95	2.3
3r2xC (HB36)	-60.1 (-53.4)	-	-	2.15 (1.95)	-11.05 (-5.03)	1.0 (1.0)
4eefG (HB80)	-76.0 (-56.5)	-	-	8.35 (1.6)	-12.14 (-9.13)	1.9 (1.2)
1yrnA	-43.4	-61.1 (E102R)	-68.1 (Q87E)	0.1	3.23	0.8
1ohzB	-47.3	-59.8 (S45W)	-65.7 (R53W)	2.1	-11.55	1.1
2x9aC	-34.7	-41.0 (D14N)	-48.4 (D46V)	1.8	-6.35	1.1
3ihqB	-47.2	-72.0 (E255W)	-81.7 (V260Y)	1.1	-3.63	1.1
1pk1B	-47.3	-55.9 (E56V)	-65.8 (S32M)	1.7	-6.21	0.9
2b87A	-27.6	-45.5 (D32N)	-59.3 (R35F)	0.5	-3.01	1.7
2y0sL	-59.2	-73.1 (H47Y)	-	1.45	-4.24	1.4
3ossC	-37.8	-49.1 (R137Q)	-57.5 (V150Q)	0.8	0.63	1.3
1s1qB	-63.5	-69.3 (Q62K)	-78.9 (I44W)	0.3	-1.7	0.9
1qfnA	-41.9	-56.4 (E56K)	-72.2 (D74R)	1.1	-5.84	1.7

computational analysis are generally intuitive (Figure 4.6). For the xylanase taken from the cellulosome complex in *Ruminiclostridium thermocellum* (chain B of PDB code 1ohz), the two substitutions (S45W and R53W) mainly serve to reduce the desolvation costs to favor binding, while modifying the side chain volume at the same time to better fill the void between the inhibitor and the target protein (Figure 4.6, A1 and A2). For the viral attachment protein G3P (2x9a chain C) from *Enterobacteria phage If1*, both of the mutations (D14N and D46V) remove the repulsion with like charges from the nearby residues (D46 and D19 of HA2 respectively) on the target site, thereby improving the binding energy (Figure 4.6, B1 and B2). The sex-comb-on-midleg protein found in *Drosophila melanogaster* (chain B of 1pk1) is optimized to undergo the mutation of E56V to avoid repulsion with like charges from the nearby (D19 of HA2) in HA (Figure 4.6, C1), and the mutation of S32M to reduce the desolvation cost and to favor hydrophobic interaction with the surrounding environment on the target site (Figure 4.6, C2). Finally, for glutaredoxin 1 from *Escherichia coli* (chain A of PDB code 1qfn), both substitutions (E56K and D74R) improve the electrostatic interaction by forming salt bridges with E57 of HA2 and D46 of HA2 from HA respectively (Figure 4.6, D1 and D2), which significantly improves the estimated binding free energy (Table 4.3).

4.4. **Discussion**:

This work presents a computational approach to design protein inhibitors for a target protein, when a structural model for the target protein in complex with a known binder exists. The key idea is to use the protein-protein interface for the known complex as a template to search for naturally present protein dimers that interact in similar fashions, from which a substitution of the known binder of the target protein can be made by its

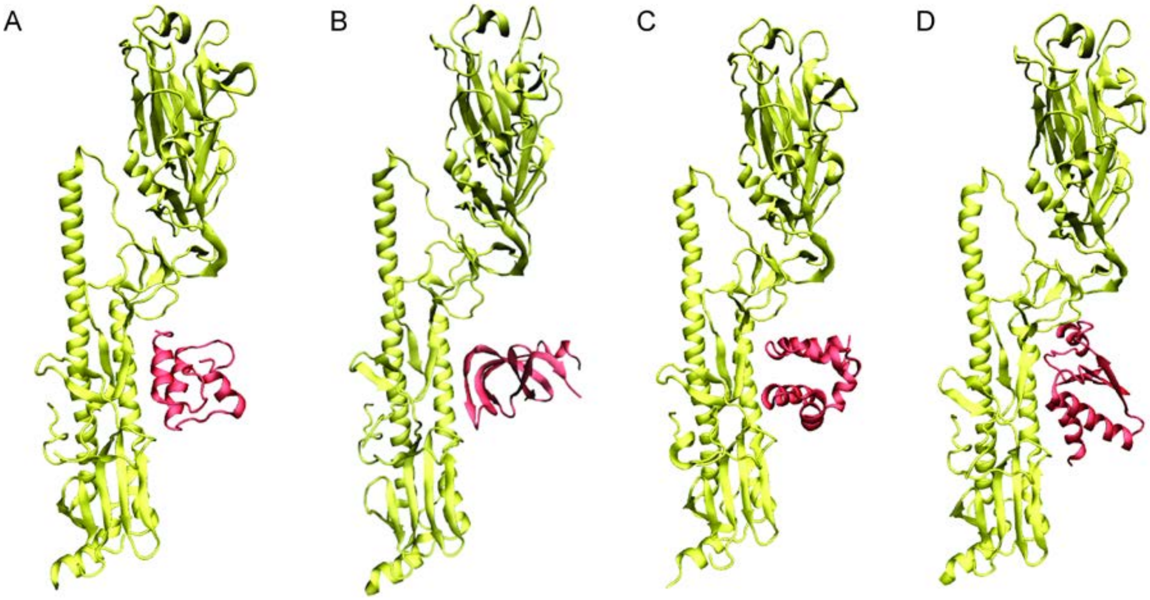

Figure 4.5. Top four designed inhibitors. In all panels, HA is colored in yellow while the predicted binding protein is colored in red; (A) chain B of PDB code 1ohz, (B) chain C of PDB code 2x9a, (C) chain B of PDB code 1pk1 and (D) chain A of PDB code 1pk1.

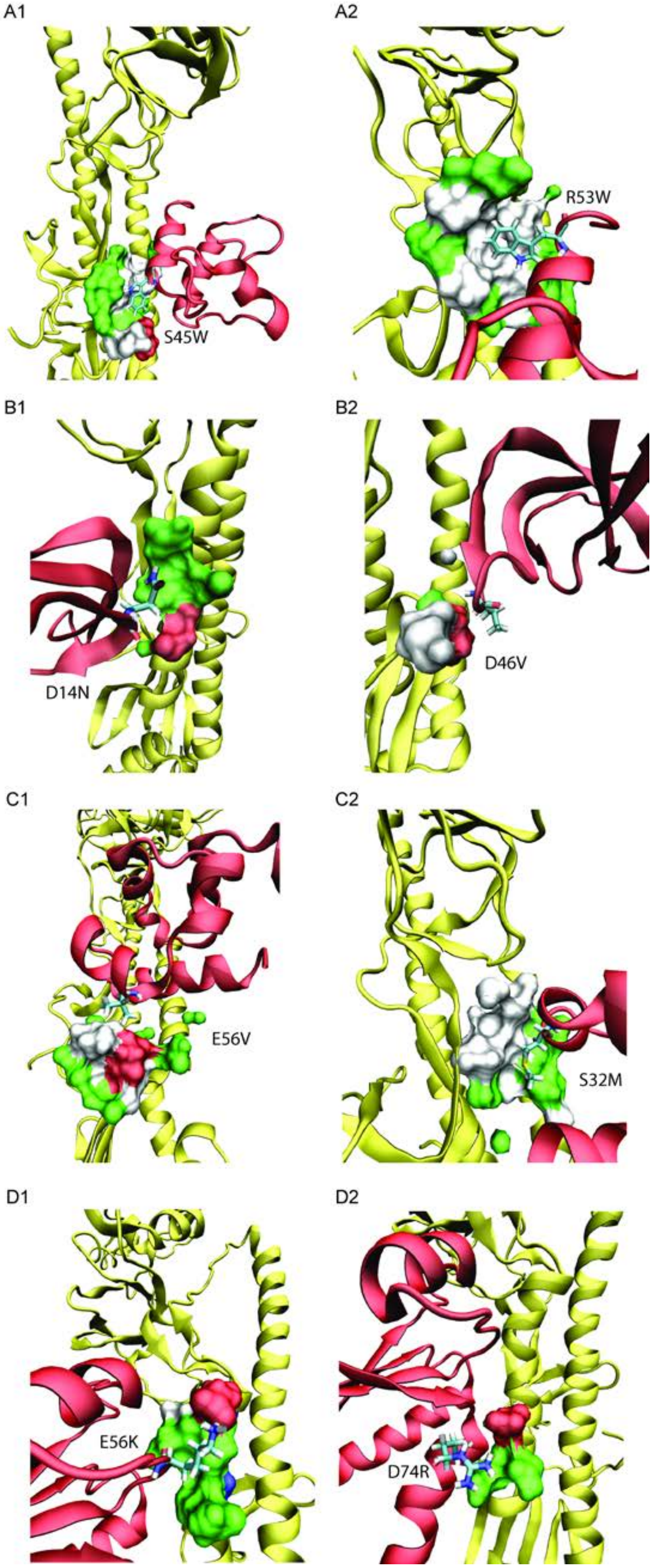
A1
A2
R53W
S45W
B1
B2
D14N
D46V
C1
C2
E56V
S32M
D1
D2
E56K
D74R

Figure 4.6. Mutations introduced in top four designs. In all panels, HA is colored in yellow while the designed binding protein is colored in red, both in cartoon representation. The mutated residues on the designed proteins are shown in bonds representation, colored by the atom types. The binding site on the target protein that is near the redesigned residues is drawn in surface representation, with non-polar residues colored in while, polar residues in green, acidic residues in red and basic residues in blue. (A1,2) illustrate the modified redidues for chain B of PDB code 1ohz. (B1,2) show the modified residues for chain C of PDB code 2x9a. (C1,2) are the mutations for chain B of PDB code 1pk1. (D1,2) show the mutations for chain A of PDB code 1qfn.

structural equivalent based on interface alignment. The concept of using templates in predicting protein-protein interactions is not new; based on sequence and/or structural similarity of the individual proteins, numerous studies have been carried out to model protein-protein interactions between some target proteins using the complex structure of their homologue pairs [229-241]. In addition, methods also exist that predict protein-protein interactions by focusing on local similarities at the interface region, either based on sequence-structure compatibility [242] or purely from a structural point of view [243-245], the latter of which is commonly applied in docking studies. The novel contribution of our work is that it is the first application, to the best of our knowledge, of the template-based method in therapeutic design.

The key to success of the template-based methods in modelling protein-protein interactions is the target-query similarity measure. Our metric used in this study, we believe, captures the minimally sufficient information to quantify interface similarity. First of all, the topology of interfacial residues is not considered, in recognition of the fact that interfaces consist of discontiguous fragments in space. Second, chemical types of the interfacial residues are accounted for when scoring the similarity. Lastly, only the coordinates of the Cα atoms and not the side chains are used for computing the scoring function, based on the rationale that the Cα atoms alone provide the skeleton of the protein scaffold that accommodates interfacial residues with certain chemical and spatial arrangement, while still allowing the flexibility of introducing mutations for re-design purposes. Therefore our method, PCalign, is well-suited for our intended inhibitor design for a given target protein.

While it is not clear at this stage whether our computational predictions are indeed valid without being confirmed by experiments, we have shown in our work a plausible alternative route to design protein binders for a target of interest when a structural template is available. Based on the evaluation by the different assessment criteria we used, this method appears to offer great promise in generating testable hypotheses. Beyond these computational predictions, we are in the process of setting up experimental collaborations to have these designs tested for binding activity. Once validated, we hope to improve our design through experiment-guided affinity maturation, using techniques such as site-directed mutagenesis or error-prone polymerase chain reaction to identify favorable mutations that enhance binding activity, which in turn should better inform us the physical principles that underlie molecular recognition. Within the scope of this work, we believe we have demonstrated a computational framework that puts us in a good starting point in exploring therapeutic options for a given drug target.

As an extension of the current work on HA inhibitor design, it is worth examining whether our proposed candidate proteins have cross-reactivity with other HA strains and subtypes, similar to the broadly neutralizing antibody CR6261, given that the intended binding site of our designed proteins should largely overlap with that of the antibody. Such tests can be carried out for different HAs using techniques such as gel filtration, yeast display or biolayer interferometry experiments. If inhibitory activity can indeed be detected, we will be more confident about the feasibility of our template-based drug design approach in providing additional diagnostic or therapeutic tools to battle against influenza viruses.

Once binding activity can be established for the predicted protein inhibitors, we are also interested in whether any toxic effects will be associated with these designs. In other words, identifying potential off-targets of the protein inhibitors is equally important in carefully evaluating the feasibility of any drug candidates, as interference with normal cellular activity involving these off-targets can be highly undesirable. For the 10 candidates we identified, five of them are bacterial proteins, one archaeal, one viral, and the remaining three eukaryotic proteins. With the exception of the ubiquitin from humans (chain B of PDB code 1s1q), the other two eukaryotic proteins are from yeasts and fruit flies respectively, which are considered to be on the lower end of the eukaryotic hierarchy. Therefore it is unlikely that these candidate proteins bind directly with human proteins to result in toxicity. However, precaution needs to be taken in examining whether the proposed target site binding region of the candidate proteins share significant sequence similarity with their human homologues, if any. Conservation of the interface region of the candidate proteins across different kingdoms could lead to unwanted off-target effects if they cross-react with the human homologues of their binding partners, disrupting the associated cellular function in the host system. Other than interference that results from shared binding sites for the HA equivalent between the candidate proteins and their human homologues, another source of toxicity can also come from remote binding sites with other proteins located far away from the designed ones on the candidate proteins. This is especially the case if the candidate protein is found in an oligomer as opposed to a dimer in its physiological state, so that the candidate protein is simultaneously interacting with other macromolecules. Three out of the 10 predictions we have belong to this category, for which redesign for interfacial residues located on the secondary binding sites may be necessary to prevent undesirable association with the human counterparts of those additional partners, should a homologous functional complex exists in our system.

CHAPTER V

Protein-protein interfaces in viral capsids are structurally unique

5.1. Introduction:

Viral capsids display an elaborate and symmetrical architecture that is not seen in other macromolecular complexes formed by cellular proteins. Most available structural models of viral capsids observe icosahedral symmetry, where 60 copies of the icosahedral asymmetric unit tile the 20 triangular faces. Each icosahedral asymmetric unit contains one or more protein subunits, with the number denoted by the Triangulation number or T-number. In T=1 viruses, all 60 capsid proteins are placed in an identical environment. The majority of the T>1 viruses with more than 60 capsid proteins in the capsid shell obey the quasi-equivalence principle proposed by Caspar and Klug [11], permitting slightly varied modes of interaction between capsid proteins in different structural environments, with a few exceptions that either still adhere to the overall icosahedral symmetry [246, 247] or slightly distort such symmetry [248].

The unique functional role of viral capsid proteins prompts questions regarding their structural characteristics, given that function follows form. Previously, our study on comparing the folded topology of viral capsid proteins and generic cellular proteins revealed the lack of connectivity in the structural space between the two [200], highlighting the geometric constraints to which the building blocks must conform to form a closed shell. In the current chapter, we concern ourselves with whether the protein-protein interfaces in viral capsids are similarly subjected to such selective pressure, that is, whether evolution plays a role in shaping how capsid proteins associate, or if physical laws of complementarity alone are sufficient in guiding the self-assembly of proteins in general.

Other than their distinct higher order organization, capsid proteins display an interesting feature of interface plasticity, allowing different intermolecular contacts to be formed by sequence-wise identical proteins in one static quaternary structure [104], as opposed to dynamically controlled variation in intermolecular interactions in some other biological systems [249]. Understanding how such fine-tuning of protein-protein interactions is achieved in viral capsids will enhance our knowledge of the fundamental principles of protein association, which we can apply to areas such as making better predictions in protein-protein docking studies by improving the scoring functions accordingly. This is only true, however, if the knowledge we gain from the wealth of interaction data in viral capsids is generalizable to all proteins, which necessitates the assessment of whether the inter-subunit interfaces in viral capsids span the structural interface space of all proteins including cellular ones.

A third aspect that motivates the comparison of interfaces in viral capsids versus those in cellular protein complexes is whether the design of antiviral drugs targeting viral capsid assembly can be rationalized to minimize toxicity. While focus of current treatment for viral diseases has largely been on viral enzymes [250], viral capsid proteins are emerging as a highly promising yet underexplored therapeutic target, with some recent advances in small molecule and peptide inhibitors that can bind HIV capsid proteins [99, 100] at the inter-subunit interfaces to disrupt the initial viral capsid assembly. As computational efforts towards structure-based drug design continue to propose new candidates for experimental validation [251], precaution needs to be taken to evaluate the selectivity of these drug candidates. If we find that the inter-subunit interfaces in capsids are indeed uniquely found in viruses, we can focus on these pathogen-specific sites for therapeutic development without worrying about off-target effects that may disrupt normal cellular activities.

To address the question of whether the modes of protein-protein recognition seen in viral capsids are representative of those found in small oligomeric protein-protein complexes that are typically formed by cellular proteins, we perform a structural comparative analysis across all pairwise-interacting dimers in viral capsids versus those in generic protein complexes for all structural models available to date. The same question was raised by Joel Janin and coworkers [129], but with a different approach that comprehensively surveyed various structural features of capsid protein-protein interfaces, crystal contacts and protein-protein complexes, such as buried surface area, chemical composition and atomic packing. In our work, we make the first attempt to directly test the hypothesis that the inter-subunit protein-protein interfaces found in viral capsids are structurally unique. Figure 5.1 outlines the design of our analysis. Specifically, we would like to examine if the structural overlap in the protein-protein interface space between the set of all capsid-forming proteins and the set of all non-capsid forming proteins is significantly small. If this is true, we then have statistical evidence supporting the uniqueness of inter-subunit interfaces in viral capsids.

5.2. Materials and Methods:

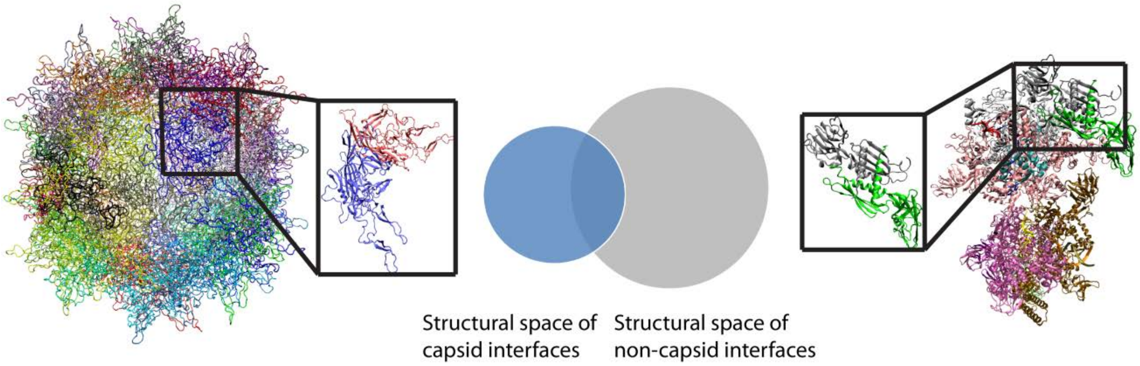

Figure 5.1. Comparison between inter-subunit interfaces in viral capsids and protein-protein interfaces in generic protein complexes. The protein-protein interfaces formed by capsid proteins, illustrated on the left (PDB code: 3kic), constitutes the structural space drawn in blue, while the generic protein-protein interfaces, exemplified by those found in an RNA polymerase elongation complex shown on the right (PDB code: 2o5i), constitutes the structural space drawn in grey. Overlap in the Venn diagram refers to the subset of generic protein-protein interfaces that sufficiently resemble some capsid inter-subunit interfaces, based on certain quantitative criterion for structural similarity, and hence signifies the extent to which interfacial patterns found in cellular protein complexes can be represented by those in viral capsids.

5.2.1. Data collection

Each viral capsid shell contains 60 copies of the icosahedral asymmetric unit (IAU), which consists of T-number of protein subunits. Therefore all pairwise dimer interfaces involving at least one protein subunit from the first IAU sufficiently represent all unique protein-protein interfacial patterns found in that particular virus. For the 421 entries in the database VIrus Particle ExploreR (VIPERdb) [10], which include procapsids (capsids in the premature form) as well as subviral particles that form as a result of modified conditions, yet all obeying icosahedral symmetry, we collect 1930 dimers involving subunits from the first asymmetric unit, with at least 10 interfacial residues and not exceeding 200 interfacial residues in those dimers. These cutoffs were chosen based on the rationale that pairwise interacting protein dimers with too small an interface are insufficiently stabilized in the complexed state, and too large a protein-protein interface typically involves multiple domains. Moreover, 50 additional entries of capsid protein dimers that we identified in the generic interface set (described in the following section) were removed from that set and added to the capsid interface set (the complete list of these 50 entries is found in Table C.1). These 1980 capsid dimers were then clustered using complete-linkage hierarchical agglomerative clustering cut at a PC-score (explained in Section 5.2.2) of 0.5 into 551 groups, from each of which a medoid structure is chosen to be included in the final representative set of capsid protein-protein interfaces. It is worth noting that the final clustering procedure has precluded many quasi-equivalent interfaces to be collectively included in the capsid set and only one or few representatives of those are retained. However, we believe that the representative set captures the coarse overall geometric and chemical moieties found in capsid shells and thus suffices for our comparative study.

For the generic interface set, 75694 structural models were collected from the Protein Data Bank [112], which were screened against PISA [222] to extract 165257 protein dimer interfaces assigned to be biologically significant (as opposed to due to crystal packing). 123395 remained after pruning those with fewer than 10 interfacial residues or more than 200 interfacial residues. This set was then grouped into dimers that have pairwise sequence identity of at least 50%. Within each group, representative dimers are chosen so that their interfaces have PC-scores lower than 0.5 with one another, given similar monomers can interact in dissimilar ways. The entries corresponding to inter-subunit interfaces in viral capsids are removed from this reduced set and appended to the viral capsid set as described earlier. Finally, the same structural clustering procedure is performed to retain representative dimers only to constitute our generic dimer set consisting of 20014 pairwise generic interfaces.

5.2.2. Comparison metric

For quantifying the similarity between two interfaces of a given pair of protein dimers, we use the program PCalign [223], which returns a normalized score, PC-score, based on structural alignment of two interfaces. PC-score takes the following form:

$$\text{PC-score} = \text{PC-score}_{\text{raw}}{}^{\ln 0.3/\ln(0.14+0.29\times 0.97^{L_{ave}})}, \qquad (5.1)$$

where L_{ave} is the average of the number of interfacial residues for the pair of protein dimers compared, and PC-score$_{\text{raw}}$ is computed by

$$\text{PC-score}_{\text{raw}} = \frac{f_c}{L_{ave}} \sum_{i=1}^{L_{ali}} \frac{1}{1+0.25\times\left(1-I_{ii}(\text{same chem type})\right)+\frac{d_{ii}^2}{16}} , \tag{5.2}$$

based on the structural alignment of the two interfaces identified by PCalign. Here f_c is the ratio of common contacts between the two interfaces aligned given by

$$f_c = \frac{\overleftrightarrow{N^1}\cdot\overleftrightarrow{N^2}}{(\overleftrightarrow{N^1}\cdot\overleftrightarrow{N^1}+\overleftrightarrow{N^2}\cdot\overleftrightarrow{N^2})/2} , \tag{5.3}$$

where $\overleftrightarrow{N^1}$ and $\overleftrightarrow{N^2}$ are matrices representing the contact maps of the aligned interfacial residues in the pair of interfaces being compared. The dot operation represents the inner product of two matrices. I_{ii}(same chem type) is the indicator function that takes the value of 1 if the i^{th} pair of aligned residues share the same chemical functional group and 0 if they do not (see [223]). d_{ii} is the spatial distance in Å between the Cα atoms of the i^{th} pair of aligned residues. This scoring function ranges from 0 to 1, with 1 resulting from comparison of identical interfaces (self-comparison).

5.2.3. Quantify overlap in structural space

To have a quantitative measure of the extent to which generic protein-protein interfaces represent patterns found at capsid inter-subunit interfaces, we perform the following analysis. We carry out $M\times N$ comparisons of protein-protein interfaces using PCalign between all M members in the capsid set and all N members in the generic set. For each of the N generic protein-protein interfaces, we find its nearest neighbor in the capsid set, where the structural distance is defined by (1 – PC-score). For a given generic protein-protein interface, the distance between itself and its nearest neighbor in the capsid set reflects how connected it is to the capsid set in the interface space. We then select the generic protein-protein interfaces whose distances with their respective nearest neighbor are less than 0.5, and this set of generic interfaces is considered to share significant structural similarity with some inter-subunit interface in viral capsids, representing the overlapping region in the Venn diagram shown in Figure 5.1. The count of these capsid-like generic interfaces is used as the test statistic for our hypothesis testing.

5.2.4. Statistical significance of the test statistic

To assess whether the shared structural space between generic protein-protein interfaces and inter-subunit interfaces found in viral capsids is significantly small, we estimate the p-value of our test statistic by a permutation test, as summarized in Figure 5.2. Similar to what was done previously [200], the total set of protein-protein interfaces is first randomly partitioned into two sets, A and B, that have the same number of interfaces as the capsid set and the generic set respectively. We then count the number of interfaces in the larger set B that highly resemble at least one interface in the smaller set A, which is our variable of interest. By repeating this experiment 10,000 times, we obtain the distribution of structural overlap between any two mutually exclusive sets of the given sizes, and can therefore estimate the probability of obtaining a value smaller than or equal to our test statistic by random chance.

5.3. Results:

5.3.1. Similar sizes but different oligomerization states

For the representative 551 capsid protein-protein interfaces and 20014 generic protein-protein interfaces, we first examine the size distribution of each set, specifically the number of interfacial residues present in a protein dimer. From Figure 5.3, we see that

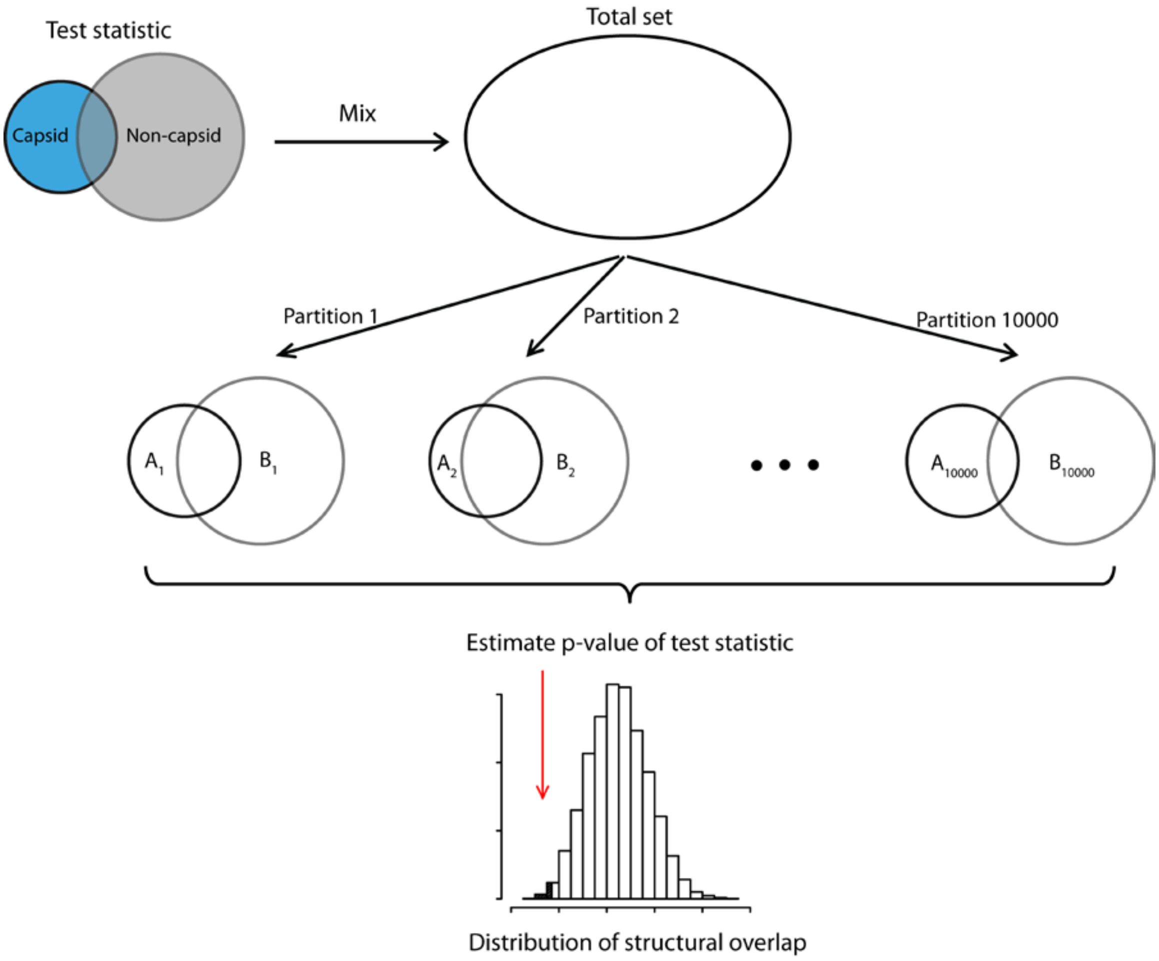

Figure 5.2. Permutation test for estimating the statistical significance. Under the assumption of our null hypothesis that inter-subunit interfaces in viral capsids are no different from generic protein-protein interfaces, we can exchange labels between the capsid set and the non-capsid set (i.e. partition the total set) in a random fashion to obtain a set A that mirrors the capsid set, and their complement set B that is the equivalent of the generic set. All possible values of the structural overlap between set A and set B under the rearrangement of the labels give us the distribution of structural overlap, from which we can obtain the statistical significance of our test statistic.

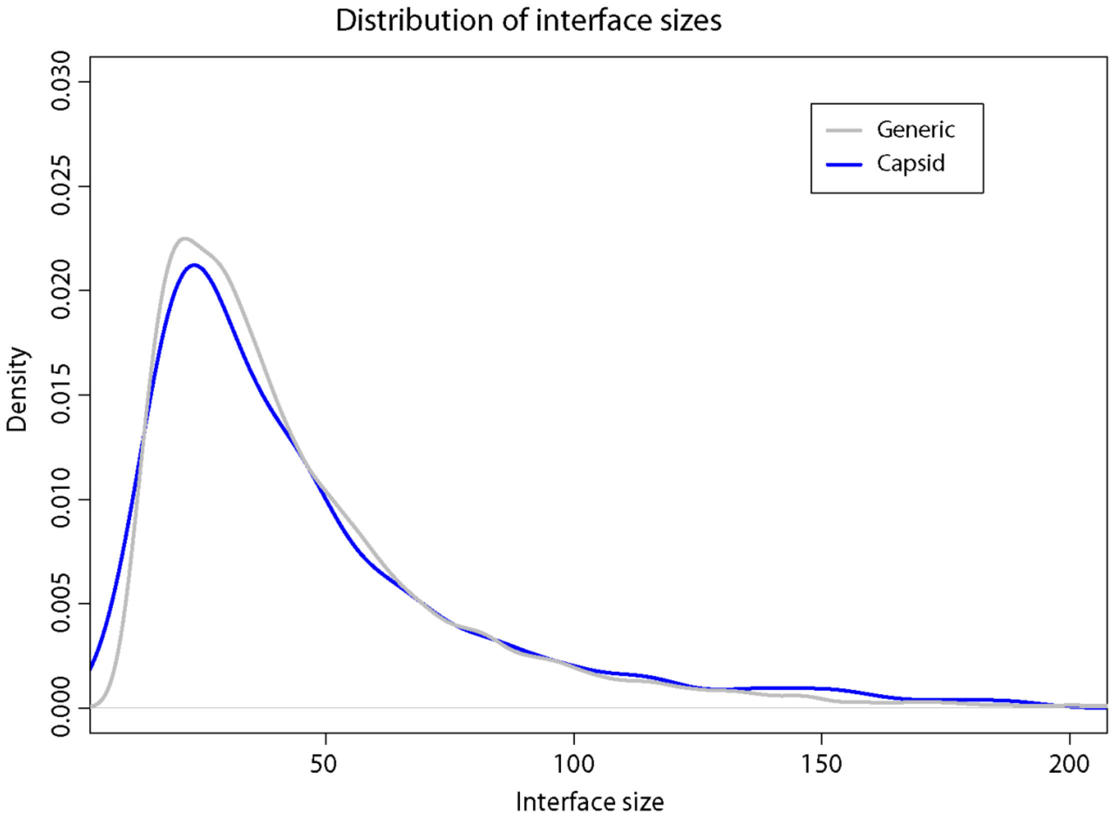

Figure 5.3. Density distribution of interface sizes for the generic set and the capsid set. The two largely overlap, with the generic set having a marginally larger proportion of interfaces in the range of 15 to 40 interfacial residues.

there is no pronounced difference in the distribution of interface sizes in the two sets, at least not in the range of interface sizes (between 10 and 200) we have considered in our analysis. The capsid set compared to the generic set has marginally a larger number of residues making contacts at an interface. The most frequently seen interface size for a generic protein dimer is 22 interfacial residues, while an inter-subunit interface in viral capsids typically has 24 interfacial residues. Nonetheless, this difference is well within the standard deviation of 30 residues in the distribution and thus negligible. Having comparable interface sizes in the two sets implies that differences identified, if any, should be mainly attributed to the geometry and the chemical properties of the interfacial residues.

The oligomerization states, however, differ significantly between the two sets, as expected. Similar to what was done in [129], we examined, for each protein of all dimers included in our analysis, the number of pairwise protein-protein interfaces that the protein is simultaneously involved in. This number of interacting partners gives a crude description of the higher order organization of proteins in their quaternary states. As shown in Figure 5.4, a typical generic protein forms a dimer, while capsid proteins are frequently interacting with 5 other proteins within the same shell, highlighting the overall complexity of capsid shell formation. The different oligomerization states may underline potential differences we find in the two sets of proteins.

5.3.2. Protein-protein interfaces formed by capsid proteins are distinct from those formed by cellular proteins

The all-against-all comparison between pairwise protein-protein interfaces in viral

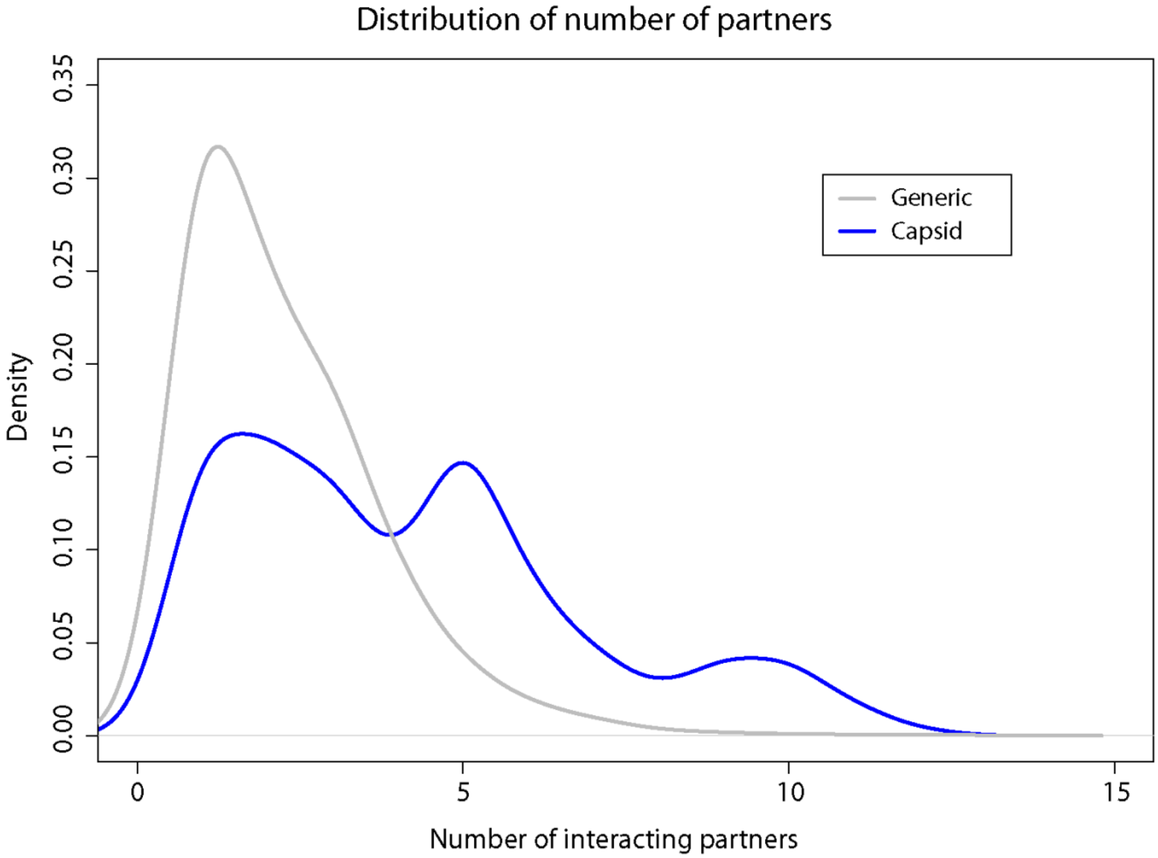

Figure 5.4. Different oligomerization states in the two sets. This figure plots the density distribution of the number of simultaneous interacting partners in the same protein complex for a given protein that forms an interface in the generic set and the capsid set**.** A capsid protein, on average, has 4.2 neighboring capsid proteins within the same shell, while the average number of interacting partners for a generic protein is only 2.2.

capsids and cellular protein complexes identified altogether 418 generic interfaces that resemble capsid ones, where the similarity is defined based on a distance cutoff of 0.5. Compared with the distribution of any two sets of the same sizes, estimated from the 10,000 permutation tests, this overlap in structural space falls on the extreme left as shown in Figure 5.5, which is significantly small with a one-tailed p-value < 0.0001. We further show in Figure 5.6 that such a disconnectivity between the capsid set and the generic set in the protein-protein interface space is not the result of differences in the sampling density of the two sets; in fact, the intra-set connectivity is about equal for the two sets. We thus have statistical evidence that in terms of interfacial patterns formed between protein dimers, inter-subunit interfaces in viral capsids are not representative of generic protein-protein interfaces found in cellular protein complexes.

The same conclusion can be arrived at even when a different distance cutoff is chosen. Figure 5.7 plots the cumulative fraction of 20014 interfaces that are within a certain structural distance of their nearest neighbor in the complementary set of 551 interfaces. The blue curve corresponding to the comparison between the capsid set and the generic set is slightly shifted to the right of the grey curves representing the 10,000 permutations, which suggests that inter-subunit interfaces in viral capsids are more different from generic protein-protein interfaces compared to what happens as a result of random chance. This holds true across the entire spectrum of structural distances, including the cutoff of 0.5 chosen previously.

5.3.3. Overlap in structural space of protein-protein interfaces

As we examine the 418 generic interfaces that resemble at least one inter-subunit

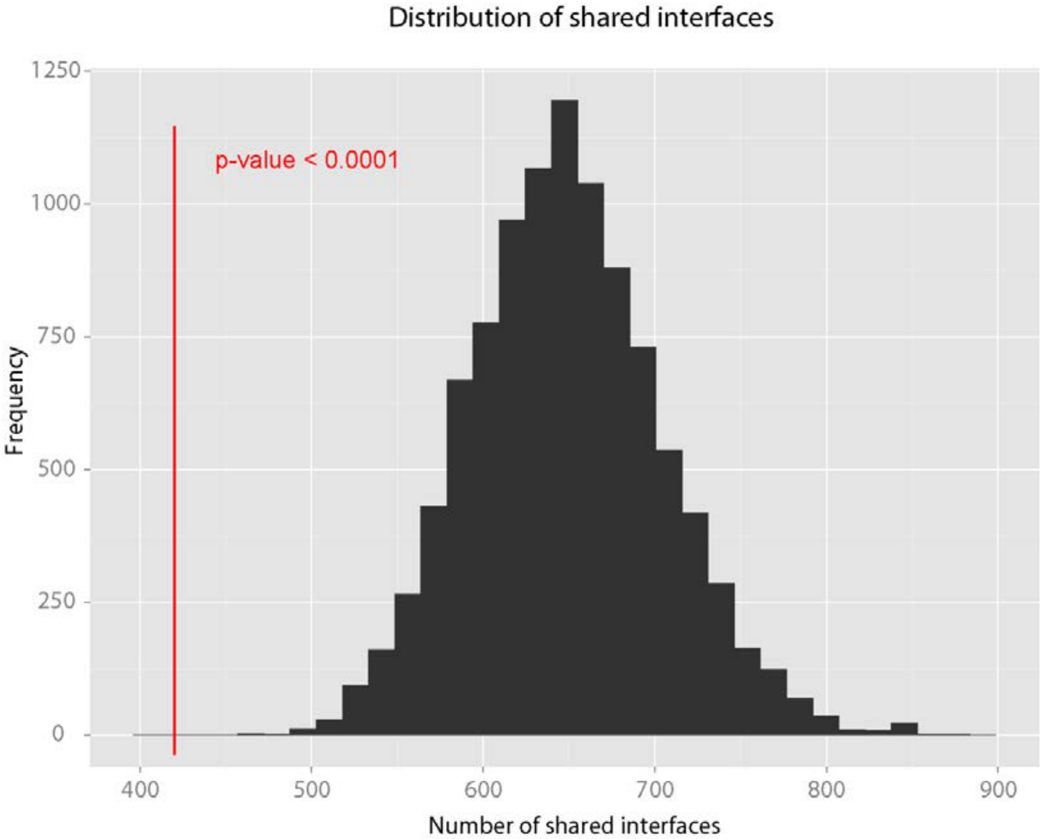

Figure 5.5. Statistical significance of the test statistic. Out of 10,000 permutations, no single case results in 418 or fewer structurally similar interfaces identified between a randomly selected set of 551 interfaces and their complement set of 20014 interfaces, which makes the one-tailed p-value of our test statistic less than 0.0001. Hence there is strong statistical evidence supporting the hypothesis that capsid interfaces are unlike generic ones.

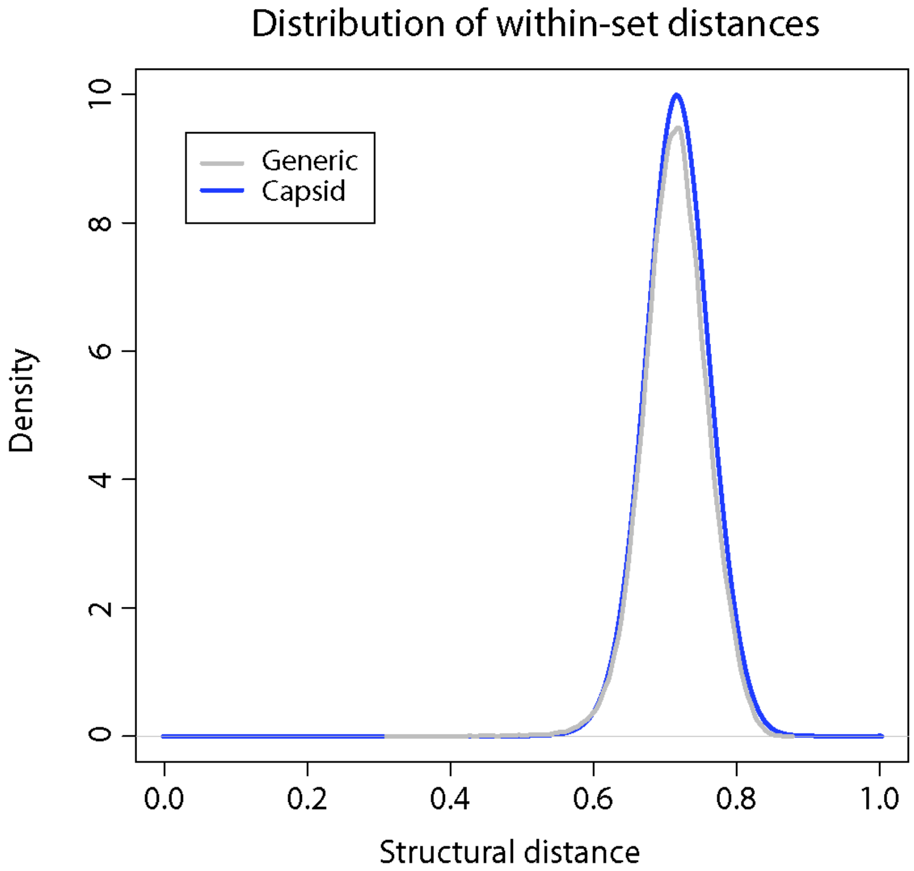

Figure 5.6. Equal connectivity in the two sets. This plot shows that the within-set structural distance distribution for the capsid interface set is well overlaid on top of that for the generic interface set, which suggests that the structural space of both sets have been sampled equally densely to give rise to the data sets, demonstrating the uniformness of the total interface set that justifies the soundness of the permutation test.

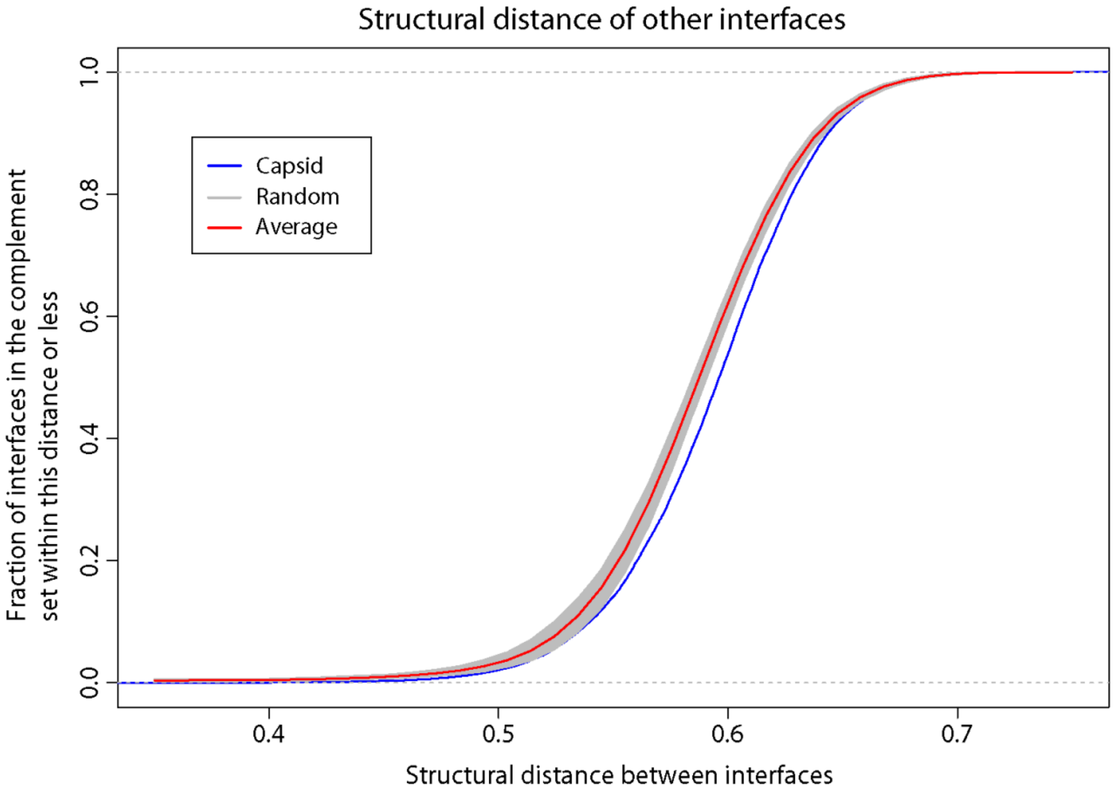

Figure 5.7. Capsid protein-protein interfaces are different from generic ones. Shown here in the empirical cumulative fraction distribution of distances between one set of 551 protein-protein interfaces and their nearest neighbor in the complementary set, we see that capsid protein-protein interfaces are structurally more distant from non-capsid ones, represented by the blue curve, than the random background, represented by 10,000 grey curves. The average empirical cumulative fraction distribution is colored in red. The range of structural distances plotted is from 0.35 to 0.75 to show a better resolved picture, while in theory this can range from 0 to 1.

interface in some capsid, we see their structural analogues in the capsid set are generally small in size, with more than 70% of them having fewer than 20 interfacial residues (Figure 5.8). For these protein dimers that form a small area of contact, there are most likely only limited ways of arranging a few points (i.e. interfacial residues) spatially. Thus for the smaller interfaces in viral capsids, it is intuitively easier to find similar patterns among generic interfaces. Furthermore, interactions rendered by a few interfacial residues are unlikely to contribute to a great amount of binding energies compared with larger interfaces (which should not be confused with the idea of a few hot spot residues anchoring protein-protein association in general). In fact, most of these protein-protein interfaces in viral capsids that overlap with generic ones are found in between capsomeres, which are pentamers or hexamers and generally considered to be stable assembly intermediates for viral capsids, rather than within a capsomere, indicating that they are less active players in the shell formation process. This overlap in structural space of protein-protein interfaces in the two sets of data thus may not have significant biological implications.

Unlike the case with comparing the folded topology between capsid proteins and generic proteins, where we found a few representative classes of cellular proteins that resemble capsid ones which also forms symmetric oligomers, no similar trend can be easily identified with the capsid-like generic interfaces in terms of the functional annotation of the constituent protein monomers. A complete list of the diverse range of GO terms associated with these protein interfaces is summarized in Table C.2. Among these, the most abundant hits are associated with proteins in the immune system. This is not surprising, given the frequent encounters and close interaction between the host

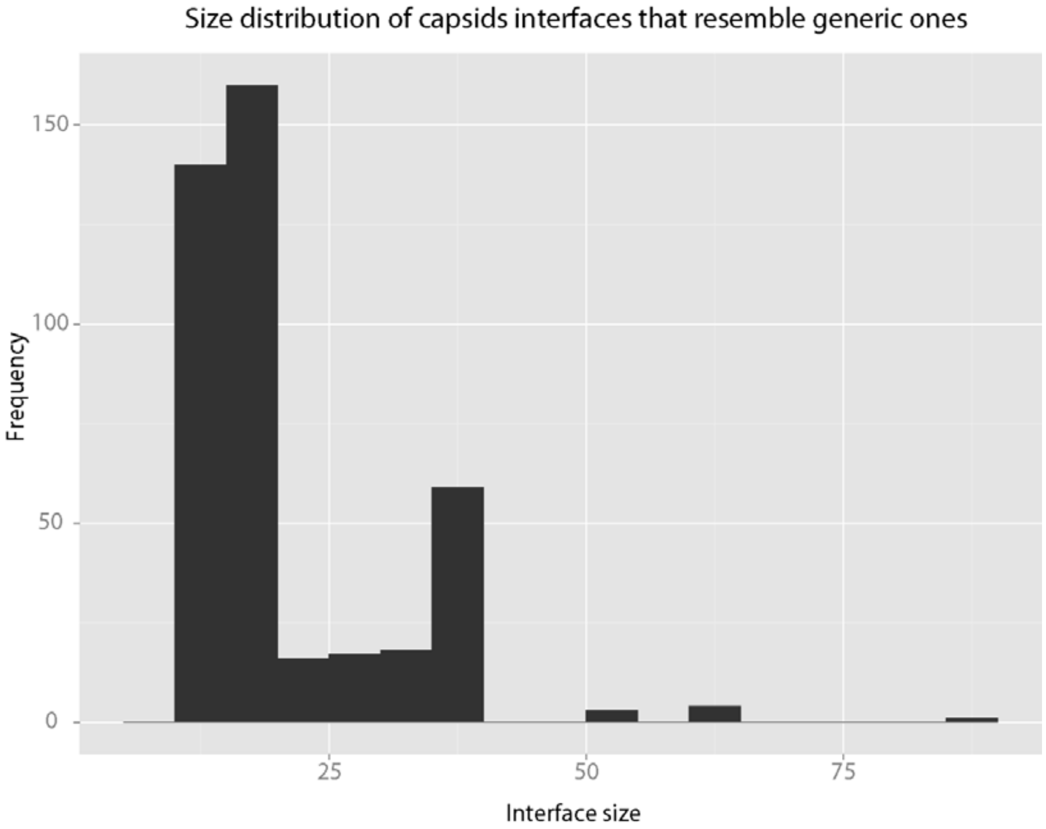

Figure 5.8. Sizes of the 418 capsid protein-protein interfaces in the overlap region. Majority of these capsid interfaces that structurally resemble some generic interfaces are small in size, with the exception of one interface that has a size of 85 interfacial residues, being a long coiled coil structure.

immune system and the viral coat proteins before the virus gains access to the cell interior. What appears most frequent in these small, capsid-like generic interfaces are clusters of discontiguous fragments from loop regions, rather than well-defined secondary structural elements, and the constituent proteins forming these small contacts are typically part of larger oligomeric complexes as opposed to homodimers. Figure 5.9 illustrates a few examples of these cases, with the constituent proteins covering functional classes such as transferases (Figure 5.9 (A)), cell adhesion (Figure 5.9 (B)) and ligases (Figure 5.9(C)).

5.4. Discussion:

5.4.1. Result of comparative study is not sensitive to quality of structural data

Structural comparison between biological molecules relies on having reasonable structural models solved by experimental techniques. The large, elaborate nature of macromolecular complexes such as viral capsids often creates limits on the sample availability and regularity, resulting in lower resolution of the structures determined. Consequently, the atomic coordinates of these structural models are less accurate and sometimes with details missing, compared to those of small protein complexes involving cellular proteins. Furthermore, in the case of viral capsids, the symmetry of the assembled architecture is utilized for averaging over the experimental density of all asymmetric units for coordinate derivation, leading to additional sources of imprecision. Nonetheless, the structural comparison tool that we used in this study scores the similarity between a pair of protein-protein interfaces with reduced representation of their structural information, both geometrically and chemically, and we have demonstrated previously [200] that our comparison metric can recognize significant similarity between related

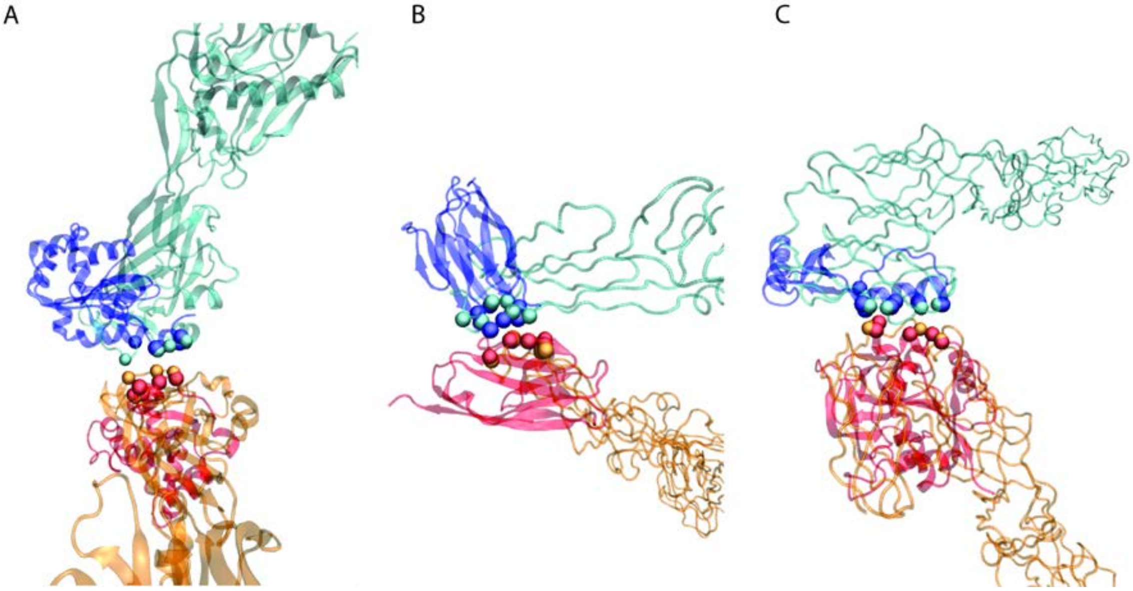

Figure 5.9. Examples of similar interfaces from the capsid set and the generic set. In all panels, the viral capsid proteins are colored in cyan and orange, while the cellular protein dimers are colored in blue and red. The monomer proteins are drawn in either cartoon representation or tube representation, the latter for structural models lacking side-chain information. The Cα atoms of the aligned interfacial residues are shown in van der Waals representation. (A) The interface formed by the generic protein dimer (PDB code 1xiq, chain IDs B and F) is aligned to the capsid interface (PDB code 1x9p, chain A in the first IAU and chain A in the 16th IAU). (B) The generic interface between chain A and chain B of PDB code 3m45 is aligned to the capsid interface formed by chain B and chain C within the same IAU in PDB code 1n6g. (C) The generic interface between chain G and chain H of PDB code 3d54 is aligned to the capsid interface between chain F of the first IAU and chain F of the second IAU in PDB entry 3muw.

protein- protein interfaces in artificially created data that are highly corrupted. Given the robustness of our method against noisy data, it is unlikely that the conclusion drawn based on our comparative study will be affected by the inaccuracy of the atomic positions in these structural models.

5.4.2. Domain-swapped interfaces are not treated differently

Domain swapping is commonly found in homomeric complexes, and it refers to two identical proteins exchanging the same structural elements to form dimer interfaces that replace the original intramolecular contacts in the monomeric states. Since its first recovery in diphtheria toxin [252], many domain-swapped structures of diverse origin have been identified [253]. These interfaces are best characterized by their highly intertwined nature, blurring the boundary between intramolecular contacts and intermolecular ones. In our work, we have not delineated this class of protein-protein interfaces from the rest of generic interfaces for special consideration, for the main reason that despite the possible advantage of the domain swapping mechanism in formation of large, stable protein complexes, domain swapping is still rare (on the order of 10) among currently solved structures [254] and remains less well understood for unambiguous annotation. This being the case, we do not expect the negligible fraction contributed by this class of protein-protein interfaces to have significant bearing on the overall conclusion of our study.

5.4.3. Implications of protein-protein interfaces in capsids being unique

Viral capsid shells represent large macromolecular assemblies that exist on a scale not seen in most of cellular macromolecular complexes. Given the unique function of capsid proteins in making shells, we are interested in whether the sticky patches on these building blocks that piece them together are also structurally unique, in order to pinpoint features critical for the design principle of large biological containers. Our hypothesis testing has provided strong statistical evidence of the distinctiveness of the interfacial patterns found between capsid proteins from those between cellular proteins, which has the important implication that how capsid proteins recognize and associate with one another is governed not simply by physical principles of shape complementarity alone, but is retained specifically by evolution to favor the higher order organization. This conclusion that capsid interfaces are subjected to evolutionary constraints is in agreement with the previous findings that changes in inter-subunit geometry have direct impact on the oligomerization states of protein subunits [255], which in turn should confer selective pressure on protein-protein interactions. Combined with our earlier observation that the folded topology of viral capsid proteins is also segregated from those of cellular proteins, we arrive at the conclusion that the basic shape of these Lego pieces and the molecular recognition sites on their edges act concertedly to create the sophisticated shell architecture as designed.

In terms of rationalizing therapeutic design of antiviral drugs, the finding of this work favors the view that pathogen specificity can be achieved, given that protein-protein interfaces in viral capsids are significantly different from those involved in cellular activities. While this statement holds true, we would like to caution against the danger of introducing inhibitors targeting the smaller interfaces in viral capsids, especially those not found within the same capsomere, which can possibly fall into the structural overlap between capsid protein-protein interfaces and generic ones. As shown in Table C.2, these potential off-targets in the cellular domain cover a wide range of biological functions, including various enzymatic activities and gene regulation that are crucial for life. A more conservative approach would be to focus on those larger interfaces that provide greater stabilization energy for the capsid shell, with the hope of not only having greater inhibitive power for the assembly process but also with fewer side effects.

Lastly, our study has confirmed that interfacial patterns in viral capsids are not representative of those in cellular protein complexes. While most cellular proteins function by forming binary interactions, either transiently or permanently, viral capsid proteins are characterized by interacting with multiple partners simultaneously in the multiple-component assemblies. What we are most interested in with regards to properties of inter-subunit interfaces in viral capsids is the control of quasi-equivalence, which provides a great example of exquisite fine-tuning of modes of interaction via conformational switching of identical gene products. However, as we established in this study, the larger interfaces within capsomeres do not share much structural similarity with generic protein-protein interfaces, and it is precisely the alternation between the concave pentamers and the flat hexamers that best manifests the quasi-equivalent property. Therefore, what we learn about physical principles governing protein-protein recognition from capsid proteins may not easily extend to protein-protein interactions at large.

The above aspects cover the implications of our findings in this work. What our conclusion does not imply, however, is the general uniqueness of structural characteristics of all viral proteins. The exclusive sets of folded topology and distinct patterns of association with one another found in viral capsid proteins are correlated with their specific function that is missing for cellular life forms. Unlike capsid proteins, other viral proteins may assume roles that partially overlap with cellular proteins. Examples include viral proteases that process polypeptides for maturation, which contain motifs for hydrolysis of peptide bonds similar to cellular proteases that lyse misfolded peptides for recycling, and reverse transcriptases in retroviruses necessarily share similar nucleotide binding sites with DNA polymerases in cellular machinery. Related studies on the structural relationship between general classes of viral proteins and cellular proteins have been carried out, with focuses on virus-host interaction in humans [182] and thermodynamic stabilities [256]. The first study established extensive overlap between virus-host interactions and endogenous interactions within the host for competitive binding, suggesting that the uniqueness of capsid protein-protein interfaces is a result of functional requirement rather than their viral origin. The second study concluded on the high adaptability of viral proteins for effective interaction with proteins in the host, as opposed to thermostable proteins, which are less tolerant of the deleterious effects of mutations. This second study presents an additional aspect of viral proteins, specifically their chemical composition and disorder propensity. One needs to be precautious in extrapolating this unique biophysical property of viral proteins to all protein-protein interfaces that are viral derived without further investigation.

5.4.4. Concluding remarks

To conclude, we have shown in this study via rigorous hypothesis testing that inter-subunit interfaces in the large, elaborate viral capsid assemblies are significantly different from protein-protein interfaces found in the smaller and simpler cellular protein complexes. This difference in capsid interfaces is most likely the consequence of the functional requirement for making the specific architecture of shells for encapsulation of the viral genome, which on average involves more neighbors per protein subunit as well as assumes a more expanded organization compared with cellular complexes that are often times binary and more compact and collapsed in nature. Our results should provide deeper understanding of the nature of self-association in large biological containers for creating predictable designs in a variety of therapeutic and materials science applications.

Appendix C

Additional information for CHAPTER V

Table C.1. List of additional capsid protein dimers. These entries are not deposited in the VIPERdb, and were subsequently moved from the generic interface set to the capsid set before a structural redundancy filter was applied to both sets.

PDB code	Chain identifiers	Virus name
1ahs	AB	African horse sickness virus
1ak4	CD	Human immunodeficiency virus 1
1bvp	56	Bluetongue virus
1c5e	AC	Enterobacteria phage lambda
1ebo	DF	Ebola virus
1ei7	AB	Tobacco mosaic virus
1em9	AB	Rous sarcoma virus
1fn9	AB	Reovirus
1hiw	AR	Human immunodeficiency virus 1
1hiw	QR	Human immunodeficiency virus 1
1jmu	BH	Reovirus
1jmu	DE	Reovirus
1jmu	DF	Reovirus
1jmu	EF	Reovirus
1jmu	FI	Reovirus
1no7	AB	Human herpesvirus 1
1p65	AB	Porcine reproductive and respiratory syndrome virus
1r6r	AB	Dengue virus 2
1u7k	EF	AKR (endogenous) murine leukemia virus
1vps	AE	Murine polyomavirus
1wnc	BF	SARS coronavirus
1zv8	GK	SARS coronavirus
2beq	DF	SARS coronavirus
2cse	1V	Mammalian orthoreovirus 1 Lang
2cse	1W	Mammalian orthoreovirus 1 Lang
2cse	AK	Mammalian orthoreovirus 1 Lang
2cse	AR	Mammalian orthoreovirus 1 Lang
2cse	AY	Mammalian orthoreovirus 1 Lang

2cse	JU	Mammalian orthoreovirus 1 Lang
2cse	LU	Mammalian orthoreovirus 1 Lang
2cse	QX	Mammalian orthoreovirus 1 Lang
2cse	VW	Mammalian orthoreovirus 1 Lang
2cse	VZ	Mammalian orthoreovirus 1 Lang
2cse	WZ	Mammalian orthoreovirus 1 Lang
2eia	AB	Equine infectious anemia virus
2ezo	BC	Simian immunodeficiency virus
2gic	AC	Vesicular stomatitis Indiana virus
2gic	BC	Vesicular stomatitis Indiana virus
2gsy	BD	Infectious bursal disease virus
2gsy	GN	Infectious bursal disease virus
2i2s	AB	Porcine rotavirus
2vvf	AB	Pseudoalteromonas phage PM2
2zl5	AB	Norwalk virus
3bwq	AE	Simian virus 40
3h00	BD	Human immunodeficiency virus type 1
3izo	BC	Human adenovirus 5
3rkc	AB	Hepatitis E virus
3sln	AC	Norovirus Hu
3sln	BJ	Norovirus Hu
3sln	DH	Norovirus Hu

Table C.2. Functional annotation of capsid-like generic interfaces. The 418 generic interfaces that structurally resemble some inter-subunit interfaces in viral capsids are highly diverse in terms of the biological functions they are involved in. This table summarizes the counts of entries in each functional category.

Functional annotation	NO. of capsid-like generic interfaces
Alpha-helical bundle	1
Antimicrobial protein	1
Apoptosis	2
Bacterial cell division	1
Beta-clamp	1
Biosynthetic protein	1
Blood clotting	3
CAMP-binding protein	1
Catalytic antibody	1
Cell adhesion	8
Cell cycle	3
Cell invasion	1
Cell motility protein	1
Chaperone	5
Chemokine	1
Chemotaxis	1
Chromatin regulator	1
Chromosome segregation	1
Complex (ligand/receptor)	4
Contractile protein	4
Cytokine	2
Cytotoxin	1
Decarboxylase	1
De novo protein	6
DNA binding protein	10
DNA integration	1
Electron transport	8
Endocytosis/Exocytosis	3
Enterotoxin	1
Gene regulation	8
Growth factor	1
Heme-binding	1
Hormone	6
Hydrolase	56
Immune system	28
Immunoglobulin	3
Isomerase	1
Lectin	2

Leucine zipper	1
Ligase	5
Light-harvesting protein	1
Lipid binding protein	1
Lipoprotein	1
Lyase	6
Membrane protein	11
Metal binding protein	3
Murine class I MHC	1
Neuropeptide	2
Neurotoxin	1
Nucleic acid binding protein	2
Oxidoreductase	22
Oxygen storage/transport	5
Peptide binding protein	1
Peroxidase	1
Phosphoribosyl transferase	1
Photosynthesis	10
Plant protein	1
Proteinase inhibitor	1
Protein binding	9
Protein transport	4
Replication	1
Retinoic-acid transport	1
Ribosomal protein	24
RNA binding protein	3
Signaling protein	13
Structural genomics, unknown	20
Structural protein	6
Sugar binding protein	3
Synthetic protein model	1
T cell receptor	1
Toxin	6
Transcription	23
Transferase	34
Tryptophan biosynthesis	1
Unknown bacterial hydrolase	1
Unknown function	6
Viral protein	12

CHAPTER VI

Conclusion

Viral capsids exhibit elaborate and symmetrical architectures not seen in cellular multimeric protein complexes. The limited genome coding capacity of viruses necessitates the economic use of a single or a few types of capsid proteins, which can efficiently and faithfully self-assemble into a homogeneous particle of a defined size. This thesis work is aimed at pinpointing structural characteristics of viral capsid proteins that govern the functional assembly of capsid shells, with the ultimate goal of gaining a deeper understanding of the design principles that control such events for various biomedical and materials science applications.

In Chapter II [200], I explored the folded topology of all viral capsid proteins for which structural data are available, and compared them with a representative set of generic proteins, in order to test the hypothesis that these shell-forming building blocks assume unique geometry. This large scale comparative analysis, using Template Modeling (TM)-score, identified only 2078 non-capsid proteins that have marginal structural resemblance to capsid proteins, covering altogether 210 folds following the definition in SCOP, which is far less than the number of shared folds between randomly partitioned sets of the same sizes. The statistical significance of this test statistic, estimated from our permutation test, is very high with a p-value less than 0.0001. This leads to our conclusion that viral capsid proteins are segregated in structural fold space. During this work, I developed a novel statistical framework that is suitable for assessing whether a structural feature is uniquely found in any functional class of proteins. Incidentally, the same analysis performed on several other functional classes of proteins, such as muscle proteins and kinases, did not establish exclusive sets of folds for these classes, which makes the major conclusion of the work even more significant. The results of this study provide novel insights into how structural folds of capsid proteins, which directly determine the shape of these building blocks, might be constrained during evolution by requirement of the assembled cage-like architecture. Additionally, my work highlights the guideline for synthetic maneuvers of virus-based nanoplatforms when imparting new functionalities, which is to adhere to the library of folds adopted by capsid proteins that nature has created specifically to make shells.

Chapter III [223] describes a computational tool I developed to compare the physicochemical properties of interfaces between two given protein dimers, focusing on local similarities at the interface region that participate in molecular recognition while disregarding the evolutionary relatedness between the constituent protein monomers. Unlike existing methods which require atomic details of the input structural models, my method can deal with relatively poor quality data determined at lower resolutions, allowing structural comparison across models with different levels of detail. This major contribution greatly expands the structural space that can be studied by including protein-protein interaction patterns found in large macromolecular complexes, exemplified by viral capsids, which are typically solved by lower-resolution experimental techniques. Such improvement was made possible by defining intermolecular contacts using minimal structural information available, based on statistical knowledge from mining the Protein Data Bank, and our amino acid type-specific definition of contacts is shown to outperform the generic distance cutoff criterion used traditionally. Additionally, the geometric hashing algorithm used in my method effectively addresses the fact that interfacial residues are discontiguous fragments clustered spatially regardless of their connectivity, which optimally captures the central properties of the protein-protein interfaces that interest us. I demonstrated the power of the PCalign method in accurately recognizing highly similar interfaces across related protein dimers both on a real dataset and on a simulated "corrupted" dataset, showing its robustness against noise. Furthermore, through comparison with existing methods, I showed that my method has better sampling on average to yield optimal structural alignment, mapping a larger fraction of structurally equivalent interfacial residues with lower root mean squared error. Finally, I illustrate the usefulness of the PCalign method in recognizing interesting biological relationships masked by apparent lack of structural similarity by a few examples of viral mimicry, where viruses convergently evolved similar binding sites on viral proteins to target cellular proteins in the host system, without these viral proteins bearing any evolutionary relationship with the native partners of their target proteins.

Chapter IV illustrates an application of the method developed in Chapter III to computer-aided drug design. This work is rationalized on the hypothesis that nature has a limited set of choices in arranging interfacial residues spatially and chemically. Using the recently solved crystal structure of the 1918 influenza virus hemagglutinin protein in complex with a neutralizing antibody CR6261 as a template, I screened a non-redundant subset of all protein dimers in the Protein Data bank to look for similar protein-protein interaction patterns to those between the epitope and the paratope. Encouraged by our finding that two recently designed protein inhibitors [113, 114] for the same target via a de novo approach scored discriminatively higher by our program PCalign than the computational models that failed experimental validation, I selected 10 top-ranking proteins which act as the antibody equivalent based on the structural alignment of the interfaces as a starting point for designing binders for our target protein. These candidates were further optimized by rationally selecting mutations at the binding site to favor the interaction. Out of the 10 initial candidates, four of them gave comparable measures of likelihood of binding to the viral protein as that of the original neutralizing antibody, based on several independent evaluation criteria such as estimated binding free energies and percent of near-native poses correctly docked. Experimental collaboration is now under way to have our predictions tested for binding activity. This work presents a generalizable pipeline of computationally designing novel protein-based inhibitors for any given therapeutic target, using structural templates identified by the PCalign tool.

Having established that viral capsid proteins are represented by unique folded topologies in Chapter II, I next investigated whether protein-protein interfaces in viral capsids are representative of generic ones found in cellular protein complexes in Chapter V. The former attribute essentially concerns the shape of the Lego pieces for shell formation, while the latter addresses the sticky patches that hold the pieces together. From the comparative analysis of a non-redundant set of 551 inter-subunit interfaces in viral capsids and a non-redundant set of 20014 protein-protein interfaces in non-capsid protein complexes, we found that 418 generic protein-protein interfaces (about 2%) share similar physicochemical patterns with some protein-protein interface in the capsid set.

This overlap in the structural space of protein-protein interfaces is again significantly small with a p-value < 0.0001, based on a permutation test on the total set of protein-protein interfaces. We thus conclude that the exclusive protein-protein interaction patterns found in viral capsid proteins are necessary, but not the only, determinants of their higher order organization. This observation is in agreement with Chothia and coworkers' finding of interfaces having direct influence on the oligomeric states of homomeric tetramers and hexamers from eleven protein families [255]. Together with my earlier study of the folded topology of viral capsid proteins, we can establish the importance of both having the geometry that conforms with the mathematical constraints and having special physicochemical complementarity among individual building blocks in directing the final complex formation of viral capsid shells.

Looking beyond, a related question that remains to be fully understood is the control of quasi-equivalence in viruses. Viruses display a highly diverse range of mechanisms in fine-tuning interactions between capsid proteins to acquire a defined shell size. For small viruses, conformational switching of capsid protein subunits allows the formation of different intermolecular contacts, which typically involves a molecular switch in the terminal arm that does not appear highly conserved in sequence. For the majority of the canonical T=3 and T=4 viruses, this switch is ordered in subunits in a certain environment while disordered in others. Leviviruses provide an exception to this rule, without any apparent conformational switching mechanism among their protein subunits, yet still assuming a T=3 icosahedral symmetry. Larger viruses use more complex ways to dictate the size of their shells, often with the help of additional scaffolding proteins, exemplified by the tape-measure protein in PRD1 viruses. Given that capsid proteins of most icosahedral viruses of varying sizes have a common structural core, represented by the ubiquitous jellyroll fold, it is particularly intriguing what the unifying molecular signature at the protein-protein interfaces is, if there is any, that plays a role in determining the T-number of the assembled architecture, in addition to the interplay with RNA/DNA molecules.

A preliminary survey was carried out for T=1, T=3 and T=4 viruses by analyzing the set of unique capsid inter-subunit interfaces in their pentamers and hexamers. While a pentamer is convex, with pairwise interacting dimers forming a bent interface, a hexamer is planar, and interfaces in a hexamer can be either bent along the 5-fold symmetry axis, or relatively flat along the 3-fold symmetry axis. The capsid proteins in these small plant viruses all have the signature jellyroll topology, with small variations in the connecting loops of the beta-strands. Comparing the protein-protein interfaces across the board therefore can reveal potential molecular fingerprints related to T-numbers.

Figure 6.1 shows the comparison result for inter-subunit interfaces in a few representative T=1, T=3 and T=4 viruses, taking one capsid from each family from the VIPERdb database. T=1 viruses include a satellite virus (PDB code: 1stm), a sobemovirus (PDB code: 1vak), a bromovirus (PDB code: 1yc6), and a circovirus (PDB code: 3r0r), of which the sobemovirus and the bromovirus are both T=1 subviral particles resulting from mutations of what are otherwise T=3 viruses. T=3 viruses include a tymovirus (PDB code: 1ddl), a bromovirus (PDB code: 1js9), a calicivirus (PDB code: 2gh8), a sobemovirus (PDB code: 2izw), a tombusvirus (PDB code: 2tbv) and a nodavirus (PDB code: 2z2q). Finally, one tetravirus (PDB code: 1ohf) represents a T=4 virus. All dimer interfaces in unique environments on the icosahedral shell are compared

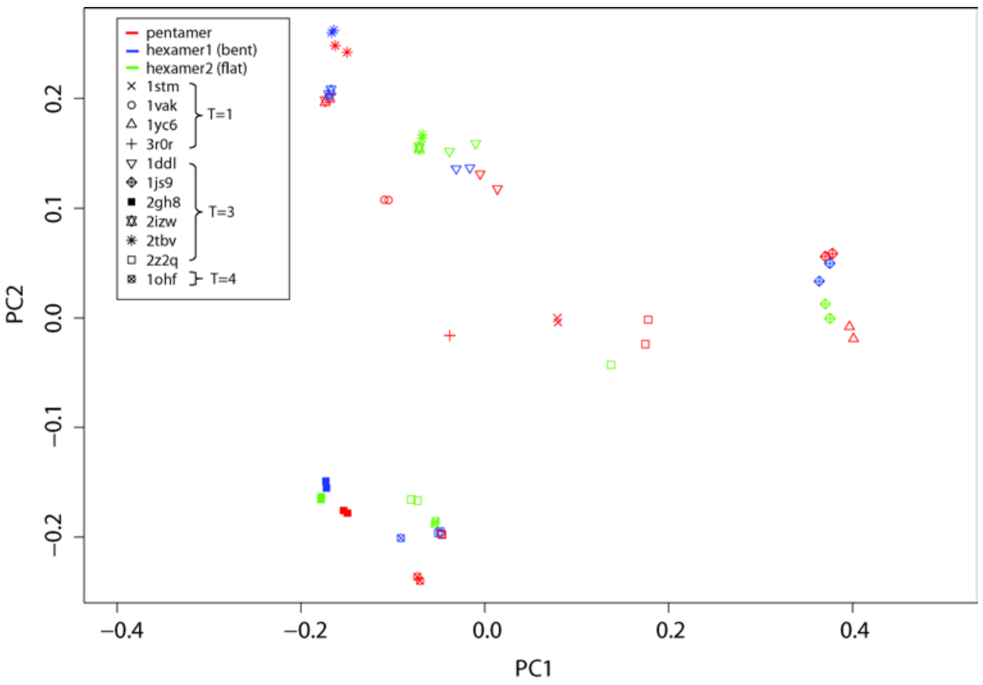

Figure 6.1. Comparison of inter-subunit interfaces in small plant viruses. The separation between different interfaces approximately represents the structural distance between them. While interfaces found within the same capsid tend to cluster together, there is no apparent similarity among interfaces found within capsids of the same T-number, or interfaces found within equivalent structural environment (pentamers or hexamers).

against one another using PCalign, and their mutual distance, defined as (1 - PC-score), is visualized via a multi-dimensional scaling plot shown in Figure 6.1. This plot captures the structural distance between different interfaces based on projection onto the first two principle components (PC1 and PC2). The color grouping refers to interfaces found at a pentamer, the bent interfaces in a hexamer (hexamer1), and the flat interfaces in a hexamer (hexamer2).

The distribution of different interfaces in Figure 6.1 shows that generally inter-subunit interfaces within the same family are more similar to one another than across different families, and this is also true for the two cases of the T=1 mutant and the T=3 wild type in the same family, which are clustered together respectively. The plot does not show a clear pattern that groups interfaces by the T-number of the capsid they are found in. Also worth noting is that the pentamer interfaces resemble the bent hexamer interfaces more than the flat hexamer interfaces, as expected. Overall, there is lack of correlation between interface similarity and the higher-order organization of capsid proteins. An interesting observation is that for the T=3 nodavirus and the T=4 tetravirus, there have been reported evolutionary relatedness in terms of capsid maturation strategies [105, 257]. Despite the different conformational switching mechanisms and particle architectures for these two viruses, the interfacial patterns show remarkable similarity, shown here as clustered together, suggesting alternative molecular motifs that converge to the same regulatory elements for tuning the hinge angles between interacting dimers.

The diversity of conformational switching mechanisms prompts the question of whether a general underlying rule exists to dictate the size of the assembled capsid. Previously, capsomeres (i.e. pentamers and hexamers) mixed in pre-defined ratios are modeled in simulations to study the dynamic assembly of viral capsids of different T-numbers [258], which has largely reproduced assembled icosahedral capsids of the expected size in silico in addition to other aberrant morphologies. This leads to our hypothesis that the T-number of icosahedral viral shells is determined by the pre-existing ratio of the different capsomeres, which in turn is influenced by the pre-disposition of the capsid protein subunits to form pairwise bent or flat interfaces. Indeed, when examining the capsid proteins in different environments on an icosahedral shell, we found tractable trends about the proportion of different conformations, specifically the order/disorder distribution, corresponding to the T-number. In T=1 viruses where all protein subunits are in identical environment, there is only one conformation. In T=4 viruses, there are four different positions for the protein subunits, which are by convention labeled as A, B, C and D. For providence virus and Nudaurelia capensis ω virus found in the tetraviridae family, two out of the four subunits assume an ordered configuration in the terminal arms while the other two have the disordered configuration, although one virus uses an N-terminal switch and the other a C-terminal switch. For T=3 viruses, there are different classes depending on the overall morphology of the shell, known as the truncated icosahedron and the rhombic triacontahedron [104]. Out of the three subunits A, B and C, only one assumes an ordered configuration in the terminal arm in the rhombic triacontahedron subtype, whereas there are two subunits that are ordered in the terminal arms in the truncated icosahedron subtype. In other words, the order/disorder ratios seem correlated with the T-numbers and the subtypes within the same T-number. If indeed this is the key regulating factor of the higher order organization, one can in theory make reliable predictions by estimating the relative frequencies of the ordered/disordered states in the conformational ensemble of a given capsid protein.

In practice, however, the above hypothesis is difficult to be tested computationally. This is mainly due to the fact that a ratio of 2:2 for the two states in T=4 viruses corresponds to the two states being in equilibrium with each other, and the ratio of 2:1 or 1:2 for the two states in T=3 viruses corresponds to a small free energy difference of about 0.4 kcal/mol, which is below the sensitivity of any molecular dynamics simulation techniques. An alternative approach may be to estimate the disorder propensity of a molecular switch using only information from its peptide sequence via Bioinformatics tools, such as DynaMine [259], and validate the predictions through experimental means, the latter of which can be laborious and expensive. Deciphering the Da Vinci code hidden in individual capsid proteins that ultimately defines the shell size remains a challenging task to explore in future.

The structural determination of the large and complex assembly of viral capsids in sub-nano resolution marks a milestone in structural biology, bringing exciting opportunities to learn about the natural design principles behind these finely-controlled biological containers. This dissertation has contributed a novel statistical framework as well as a computational tool that enabled the systematic survey of different structural features of viral capsid proteins crucial for their function. Through rigorous hypothesis testing, I have delineated the folded topology as well as interfacial patterns of viral capsid proteins as important attributes retained specifically by evolution for shell assembly. My findings should provide guidelines for designing and modulating the assembly of biological building blocks into defined higher order structures in various nanotechnology applications, as well as rationalize the design of antiviral drugs towards targeting pathogen-specific sites for improved therapeutic efficacy and reduced toxicity.

BIBLIOGRAPHY

1. Suttle CA: **Viruses in the sea**. *Nature* 2005, **437**(7057):356-361.
2. Douglas T, Young M: **Viruses: making friends with old foes**. *Science* 2006, **312**(5775):873-875.
3. Abroi A, Gough J: **Are viruses a source of new protein folds for organisms? - Virosphere structure space and evolution**. *BioEssays : news and reviews in molecular, cellular and developmental biology* 2011, **33**(8):626-635.
4. Nasir A, Forterre P, Kim KM, Caetano-Anolles G: **The distribution and impact of viral lineages in domains of life**. *Frontiers in microbiology* 2014, **5**:194.
5. Khayat R, Tang L, Larson ET, Lawrence CM, Young M, Johnson JE: **Structure of an archaeal virus capsid protein reveals a common ancestry to eukaryotic and bacterial viruses**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(52):18944-18949.
6. Fokine A, Leiman PG, Shneider MM, Ahvazi B, Boeshans KM, Steven AC, Black LW, Mesyanzhinov VV, Rossmann MG: **Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(20):7163-7168.
7. Krupovic M, Bamford DH: **Virus evolution: how far does the double beta-barrel viral lineage extend?** *Nature reviews Microbiology* 2008, **6**(12):941-948.
8. Bamford DH, Grimes JM, Stuart DI: **What does structure tell us about virus evolution?** *Current opinion in structural biology* 2005, **15**(6):655-663.
9. Roos WH, Ivanovska IL, Evilevitch A, Wuite GJ: **Viral capsids: mechanical characteristics, genome packaging and delivery mechanisms**. *Cellular and molecular life sciences : CMLS* 2007, **64**(12):1484-1497.
10. Carrillo-Tripp M, Shepherd CM, Borelli IA, Venkataraman S, Lander G, Natarajan P, Johnson JE, Brooks CL, 3rd, Reddy VS: **VIPERdb2: an enhanced and web API enabled relational database for structural virology**. *Nucleic acids research* 2009, **37**(Database issue):D436-442.
11. Caspar DL, Klug A: **Physical principles in the construction of regular viruses**. *Cold Spring Harbor symposia on quantitative biology* 1962, **27**:1-24.
12. Mannige RV, Brooks CL, III: **Periodic table of virus capsids: implications for natural selection and design**. *PloS one* 2010, **5**(3):e9423.
13. Keef T, Twarock R, Elsawy KM: **Blueprints for viral capsids in the family of polyomaviridae**. *Journal of theoretical biology* 2008, **253**(4):808-816.
14. Naitow H, Tang J, Canady M, Wickner RB, Johnson JE: **L-A virus at 3.4 A resolution reveals particle architecture and mRNA decapping mechanism**. *Nature structural biology* 2002, **9**(10):725-728.
15. Cheng RH, Reddy VS, Olson NH, Fisher AJ, Baker TS, Johnson JE: **Functional implications of quasi-equivalence in a T = 3 icosahedral animal virus established by cryo-electron microscopy and X-ray crystallography**. *Structure* 1994, **2**(4):271-282.

16. Mannige RV, Brooks CL, III: **Tilable nature of virus capsids and the role of topological constraints in natural capsid design**. *Physical review E, Statistical, nonlinear, and soft matter physics* 2008, **77**(5 Pt 1):051902.

17. Lin T, Porta C, Lomonossoff G, Johnson JE: **Structure-based design of peptide presentation on a viral surface: the crystal structure of a plant/animal virus chimera at 2.8 A resolution**. *Folding & design* 1996, **1**(3):179-187.

18. McLain L, Durrani Z, Wisniewski LA, Porta C, Lomonossoff GP, Dimmock NJ: **Stimulation of neutralizing antibodies to human immunodeficiency virus type 1 in three strains of mice immunized with a 22 amino acid peptide of gp41 expressed on the surface of a plant virus**. *Vaccine* 1996, **14**(8):799-810.

19. Langeveld JP, Kamstrup S, Uttenthal A, Strandbygaard B, Vela C, Dalsgaard K, Beekman NJ, Meloen RH, Casal JI: **Full protection in mink against mink enteritis virus with new generation canine parvovirus vaccines based on synthetic peptide or recombinant protein**. *Vaccine* 1995, **13**(11):1033-1037.

20. Obeid OE, Partidos CD, Howard CR, Steward MW: **Protection against morbillivirus-induced encephalitis by immunization with a rationally designed synthetic peptide vaccine containing B- and T-cell epitopes from the fusion protein of measles virus**. *Journal of virology* 1995, **69**(3):1420-1428.

21. Porta C, Spall VE, Loveland J, Johnson JE, Barker PJ, Lomonossoff GP: **Development of cowpea mosaic virus as a high-yielding system for the presentation of foreign peptides**. *Virology* 1994, **202**(2):949-955.

22. Usha R, Rohll JB, Spall VE, Shanks M, Maule AJ, Johnson JE, Lomonossoff GP: **Expression of an animal virus antigenic site on the surface of a plant virus particle**. *Virology* 1993, **197**(1):366-374.

23. Pumpens P, Grens E: **HBV core particles as a carrier for B cell/T cell epitopes**. *Intervirology* 2001, **44**(2-3):98-114.

24. Stahl SJ, Murray K: **Immunogenicity of peptide fusions to hepatitis B virus core antigen**. *Proceedings of the National Academy of Sciences of the United States of America* 1989, **86**(16):6283-6287.

25. Kratz PA, Bottcher B, Nassal M: **Native display of complete foreign protein domains on the surface of hepatitis B virus capsids**. *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**(5):1915-1920.

26. Birkett A, Lyons K, Schmidt A, Boyd D, Oliveira GA, Siddique A, Nussenzweig R, Calvo-Calle JM, Nardin E: **A modified hepatitis B virus core particle containing multiple epitopes of the Plasmodium falciparum circumsporozoite protein provides a highly immunogenic malaria vaccine in preclinical analyses in rodent and primate hosts**. *Infection and immunity* 2002, **70**(12):6860-6870.

27. Oliveira GA, Wetzel K, Calvo-Calle JM, Nussenzweig R, Schmidt A, Birkett A, Dubovsky F, Tierney E, Gleiter CH, Boehmer G *et al*: **Safety and enhanced immunogenicity of a hepatitis B core particle Plasmodium falciparum malaria vaccine formulated in adjuvant Montanide ISA 720 in a phase I trial**. *Infection and immunity* 2005, **73**(6):3587-3597.

28. Jegerlehner A, Tissot A, Lechner F, Sebbel P, Erdmann I, Kundig T, Bachi T, Storni T, Jennings G, Pumpens P *et al*: **A molecular assembly system that renders antigens of choice highly repetitive for induction of protective B cell responses**. *Vaccine* 2002, **20**(25-26):3104-3112.

29. Billaud JN, Peterson D, Barr M, Chen A, Sallberg M, Garduno F, Goldstein P, McDowell W, Hughes J, Jones J *et al*: **Combinatorial approach to hepadnavirus-like particle vaccine design**. *Journal of virology* 2005, **79**(21):13656-13666.

30. Netter HJ, Macnaughton TB, Woo WP, Tindle R, Gowans EJ: **Antigenicity and immunogenicity of novel chimeric hepatitis B surface antigen particles with exposed hepatitis C virus epitopes**. *Journal of virology* 2001, **75**(5):2130-2141.

31. Sojikul P, Buehner N, Mason HS: **A plant signal peptide-hepatitis B surface antigen fusion protein with enhanced stability and immunogenicity expressed in plant cells**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(5):2209-2214.

32. Bisht H, Chugh DA, Swaminathan S, Khanna N: **Expression and purification of Dengue virus type 2 envelope protein as a fusion with hepatitis B surface antigen in Pichia pastoris**. *Protein expression and purification* 2001, **23**(1):84-96.

33. Bisht H, Chugh DA, Raje M, Swaminathan SS, Khanna N: **Recombinant dengue virus type 2 envelope/hepatitis B surface antigen hybrid protein expressed in Pichia pastoris can function as a bivalent immunogen**. *Journal of biotechnology* 2002, **99**(2):97-110.

34. Schlienger K, Mancini M, Riviere Y, Dormont D, Tiollais P, Michel ML: **Human immunodeficiency virus type 1 major neutralizing determinant exposed on hepatitis B surface antigen particles is highly immunogenic in primates**. *Journal of virology* 1992, **66**(4):2570-2576.

35. Eckhart L, Raffelsberger W, Ferko B, Klima A, Purtscher M, Katinger H, Ruker F: **Immunogenic presentation of a conserved gp41 epitope of human immunodeficiency virus type 1 on recombinant surface antigen of hepatitis B virus**. *The Journal of general virology* 1996, **77 (Pt 9)**:2001-2008.

36. Frazer IH, Quinn M, Nicklin JL, Tan J, Perrin LC, Ng P, O'Connor VM, White O, Wendt N, Martin J *et al*: **Phase 1 study of HPV16-specific immunotherapy with E6E7 fusion protein and ISCOMATRIX adjuvant in women with cervical intraepithelial neoplasia**. *Vaccine* 2004, **23**(2):172-181.

37. Dale CJ, Liu XS, De Rose R, Purcell DF, Anderson J, Xu Y, Leggatt GR, Frazer IH, Kent SJ: **Chimeric human papilloma virus-simian/human immunodeficiency virus virus-like-particle vaccines: immunogenicity and protective efficacy in macaques**. *Virology* 2002, **301**(1):176-187.

38. Liu WJ, Liu XS, Zhao KN, Leggatt GR, Frazer IH: **Papillomavirus virus-like particles for the delivery of multiple cytotoxic T cell epitopes**. *Virology* 2000, **273**(2):374-382.

39. Griffiths JC, Berrie EL, Holdsworth LN, Moore JP, Harris SJ, Senior JM, Kingsman SM, Kingsman AJ, Adams SE: **Induction of high-titer neutralizing antibodies, using hybrid human immunodeficiency virus V3-Ty viruslike particles in a clinically relevant adjuvant**. *Journal of virology* 1991, **65**(1):450-456.

40. Weber J, Cheinsong-Popov R, Callow D, Adams S, Patou G, Hodgkin K, Martin S, Gotch F, Kingsman A: **Immunogenicity of the yeast recombinant p17/p24:Ty virus-like particles (p24-VLP) in healthy volunteers**. *Vaccine* 1995, **13**(9):831-834.

41. Gilbert SC, Plebanski M, Harris SJ, Allsopp CE, Thomas R, Layton GT, Hill AV: **A protein particle vaccine containing multiple malaria epitopes**. *Nature biotechnology* 1997, **15**(12):1280-1284.

42. Deml L, Speth C, Dierich MP, Wolf H, Wagner R: **Recombinant HIV-1 Pr55gag virus-like particles: potent stimulators of innate and acquired immune responses**. *Molecular immunology* 2005, **42**(2):259-277.

43. Niikura M, Takamura S, Kim G, Kawai S, Saijo M, Morikawa S, Kurane I, Li TC, Takeda N, Yasutomi Y: **Chimeric recombinant hepatitis E virus-like particles as an oral vaccine vehicle presenting foreign epitopes**. *Virology* 2002, **293**(2):273-280.

44. Maurer P, Jennings GT, Willers J, Rohner F, Lindman Y, Roubicek K, Renner WA, Muller P, Bachmann MF: **A therapeutic vaccine for nicotine dependence: preclinical efficacy, and Phase I safety and immunogenicity**. *European journal of immunology* 2005, **35**(7):2031-2040.

45. Young LS, Searle PF, Onion D, Mautner V: **Viral gene therapy strategies: from basic science to clinical application**. *The Journal of pathology* 2006, **208**(2):299-318.

46. Jolly D: **Viral vector systems for gene therapy**. *Cancer gene therapy* 1994, **1**(1):51-64.

47. Cavazzana-Calvo M, Hacein-Bey S, de Saint Basile G, Gross F, Yvon E, Nusbaum P, Selz F, Hue C, Certain S, Casanova JL *et al*: **Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease**. *Science* 2000, **288**(5466):669-672.

48. Hacein-Bey-Abina S, Le Deist F, Carlier F, Bouneaud C, Hue C, De Villartay JP, Thrasher AJ, Wulffraat N, Sorensen R, Dupuis-Girod S *et al*: **Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy**. *The New England journal of medicine* 2002, **346**(16):1185-1193.

49. Imren S, Fabry ME, Westerman KA, Pawliuk R, Tang P, Rosten PM, Nagel RL, Leboulch P, Eaves CJ, Humphries RK: **High-level beta-globin expression and preferred intragenic integration after lentiviral transduction of human cord blood stem cells**. *The Journal of clinical investigation* 2004, **114**(7):953-962.

50. Ostedgaard LS, Rokhlina T, Karp PH, Lashmit P, Afione S, Schmidt M, Zabner J, Stinski MF, Chiorini JA, Welsh MJ: **A shortened adeno-associated virus expression cassette for CFTR gene transfer to cystic fibrosis airway epithelia**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(8):2952-2957.

51. Chatterji A, Ochoa WF, Ueno T, Lin T, Johnson JE: **A virus-based nanoblock with tunable electrostatic properties**. *Nano letters* 2005, **5**(4):597-602.

52. Yildiz I, Shukla S, Steinmetz NF: **Applications of viral nanoparticles in medicine**. *Current opinion in biotechnology* 2011, **22**(6):901-908.

53. Lee LA, Wang Q: **Adaptations of nanoscale viruses and other protein cages for medical applications**. *Nanomedicine : nanotechnology, biology, and medicine* 2006, **2**(3):137-149.

54. Leong HS, Steinmetz NF, Ablack A, Destito G, Zijlstra A, Stuhlmann H, Manchester M, Lewis JD: **Intravital imaging of embryonic and tumor neovasculature using viral nanoparticles**. *Nature protocols* 2010, **5**(8):1406-1417.

55. Steinmetz NF, Cho CF, Ablack A, Lewis JD, Manchester M: **Cowpea mosaic virus nanoparticles target surface vimentin on cancer cells**. *Nanomedicine (Lond)* 2011, **6**(2):351-364.

56. Lewis JD, Destito G, Zijlstra A, Gonzalez MJ, Quigley JP, Manchester M, Stuhlmann H: **Viral nanoparticles as tools for intravital vascular imaging**. *Nature medicine* 2006, **12**(3):354-360.

57. Jung B, Rao AL, Anvari B: **Optical nano-constructs composed of genome-depleted brome mosaic virus doped with a near infrared chromophore for potential biomedical applications**. *ACS nano* 2011, **5**(2):1243-1252.

58. Li F, Zhang ZP, Peng J, Cui ZQ, Pang DW, Li K, Wei HP, Zhou YF, Wen JK, Zhang XE: **Imaging viral behavior in Mammalian cells with self-assembled capsid-quantum-dot hybrid particles**. *Small* 2009, **5**(6):718-726.

59. Datta A, Hooker JM, Botta M, Francis MB, Aime S, Raymond KN: **High relaxivity gadolinium hydroxypyridonate-viral capsid conjugates: nanosized MRI contrast agents**. *Journal of the American Chemical Society* 2008, **130**(8):2546-2552.

60. Allen M, Bulte JW, Liepold L, Basu G, Zywicke HA, Frank JA, Young M, Douglas T: **Paramagnetic viral nanoparticles as potential high-relaxivity magnetic resonance contrast agents**. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine* 2005, **54**(4):807-812.

61. Anderson EA, Isaacman S, Peabody DS, Wang EY, Canary JW, Kirshenbaum K: **Viral nanoparticles donning a paramagnetic coat: conjugation of MRI contrast agents to the MS2 capsid**. *Nano letters* 2006, **6**(6):1160-1164.

62. Prasuhn DE, Jr., Yeh RM, Obenaus A, Manchester M, Finn MG: **Viral MRI contrast agents: coordination of Gd by native virions and attachment of Gd complexes by azide-alkyne cycloaddition**. *Chem Commun (Camb)* 2007(12):1269-1271.

63. Steinmetz NF, Hong V, Spoerke ED, Lu P, Breitenkamp K, Finn MG, Manchester M: **Buckyballs meet viral nanoparticles: candidates for biomedicine**. *Journal of the American Chemical Society* 2009, **131**(47):17093-17095.

64. Stephanopoulos N, Tong GJ, Hsiao SC, Francis MB: **Dual-surface modified virus capsids for targeted delivery of photodynamic agents to cancer cells**. *ACS nano* 2010, **4**(10):6014-6020.

65. Ren Y, Wong SM, Lim LY: **Folic acid-conjugated protein cages of a plant virus: a novel delivery platform for doxorubicin**. *Bioconjugate chemistry* 2007, **18**(3):836-843.

66. Lu Y, Low PS: **Folate-mediated delivery of macromolecular anticancer therapeutic agents**. *Advanced drug delivery reviews* 2002, **54**(5):675-693.

67. Brown WL, Mastico RA, Wu M, Heal KG, Adams CJ, Murray JB, Simpson JC, Lord JM, Taylor-Robinson AW, Stockley PG: **RNA bacteriophage capsid-mediated drug delivery and epitope presentation**. *Intervirology* 2002, **45**(4-6):371-380.

68. Abbing A, Blaschke UK, Grein S, Kretschmar M, Stark CM, Thies MJ, Walter J, Weigand M, Woith DC, Hess J *et al*: **Efficient intracellular delivery of a protein and a low molecular weight substance via recombinant polyomavirus-like particles**. *The Journal of biological chemistry* 2004, **279**(26):27410-27421.

69. Suci PA, Varpness Z, Gillitzer E, Douglas T, Young M: **Targeting and photodynamic killing of a microbial pathogen using protein cage architectures functionalized with a photosensitizer**. *Langmuir : the ACS journal of surfaces and colloids* 2007, **23**(24):12280-12286.

70. Kramer RM, Li C, Carter DC, Stone MO, Naik RR: **Engineered protein cages for nanomaterial synthesis**. *Journal of the American Chemical Society* 2004, **126**(41):13282-13286.

71. Flynn CE, Mao CB, Hayhurst A, Williams JL, Georgiou G, Iverson B, Belcher AM: **Synthesis and organization of nanoscale II-VI semiconductor materials using evolved peptide specificity and viral capsid assembly**. *Journal of Materials Chemistry* 2003, **13**(10):2414-2421.

72. Mao C, Solis DJ, Reiss BD, Kottmann ST, Sweeney RY, Hayhurst A, Georgiou G, Iverson B, Belcher AM: **Virus-based toolkit for the directed synthesis of magnetic and semiconducting nanowires**. *Science* 2004, **303**(5655):213-217.

73. Sarikaya M, Tamerler C, Jen AK, Schulten K, Baneyx F: **Molecular biomimetics: nanotechnology through biology**. *Nature materials* 2003, **2**(9):577-585.

74. Whaley SR, English DS, Hu EL, Barbara PF, Belcher AM: **Selection of peptides with semiconductor binding specificity for directed nanocrystal assembly**. *Nature* 2000, **405**(6787):665-668.

75. Douglas T, Young M: **Host-guest encapsulation of materials by assembled virus protein cages**. *Nature* 1998, **393**(6681):152-155.

76. Knez M, Bittner AM, Boes F, Wege C, Jeske H, Maiss E, Kern K: **Biotemplate synthesis of 3-nm nickel and cobalt nanowires**. *Nano letters* 2003, **3**(8):1079-1082.

77. Dujardin E, Peet C, Stubbs G, Culver JN, Mann S: **Organization of metallic nanoparticles using tobacco mosaic virus templates**. *Nano letters* 2003, **3**(3):413-417.

78. Endo M, Fujitsuka M, Majima T: **Porphyrin light-harvesting arrays constructed in the recombinant tobacco mosaic virus scaffold**. *Chemistry* 2007, **13**(31):8660-8666.

79. Miller RA, Presley AD, Francis MB: **Self-assembling light-harvesting systems from synthetically modified tobacco mosaic virus coat proteins**. *Journal of the American Chemical Society* 2007, **129**(11):3104-3109.

80. Sen Gupta S, Kuzelka J, Singh P, Lewis WG, Manchester M, Finn MG: **Accelerated bioorthogonal conjugation: a practical method for the ligation of diverse functional molecules to a polyvalent virus scaffold**. *Bioconjugate chemistry* 2005, **16**(6):1572-1579.

81. Blum AS, Soto CM, Wilson CD, Cole JD, Kim M, Gnade B, Chatterji A, Ochoa WF, Lin TW, Johnson JE *et al*: **Cowpea mosaic virus as a scaffold for 3-D patterning of gold nanoparticles**. *Nano letters* 2004, **4**(5):867-870.

82. Wang Q, Lin T, Tang L, Johnson JE, Finn MG: **Icosahedral virus particles as addressable nanoscale building blocks**. *Angew Chem Int Ed Engl* 2002, **41**(3):459-462.

83. Tang J, Johnson JM, Dryden KA, Young MJ, Zlotnick A, Johnson JE: **The role of subunit hinges and molecular "switches" in the control of viral capsid polymorphism**. *Journal of structural biology* 2006, **154**(1):59-67.

84. Klem MT, Resnick DA, Gilmore K, Young M, Idzerda YU, Douglas T: **Synthetic control over magnetic moment and exchange bias in all-oxide materials encapsulated within a spherical protein cage**. *Journal of the American Chemical Society* 2007, **129**(1):197-201.

85. De Clercq E: **Strategies in the design of antiviral drugs**. *Nature reviews Drug discovery* 2002, **1**(1):13-25.

86. McGuigan C, Yarnold CJ, Jones G, Velazquez S, Barucki H, Brancale A, Andrei G, Snoeck R, De Clercq E, Balzarini J: **Potent and selective inhibition of varicella-zoster virus (VZV) by nucleoside analogues with an unusual bicyclic base**. *Journal of medicinal chemistry* 1999, **42**(22):4479-4484.

87. McGuigan C, Barucki H, Carangio A, Blewett S, Andrei G, Snoeck R, De Clercq E, Balzarini J, Erichsen JT: **Highly potent and selective inhibition of varicella-zoster virus by bicyclic furopyrimidine nucleosides bearing an aryl side chain**. *Journal of medicinal chemistry* 2000, **43**(26):4993-4997.

88. Patick AK, Potts KE: **Protease inhibitors as antiviral agents**. *Clinical microbiology reviews* 1998, **11**(4):614-627.

89. von Itzstein M, Wu WY, Kok GB, Pegg MS, Dyason JC, Jin B, Van Phan T, Smythe ML, White HF, Oliver SW *et al*: **Rational design of potent sialidase-based inhibitors of influenza virus replication**. *Nature* 1993, **363**(6428):418-423.

90. Barnett JM, Cadman A, Gor D, Dempsey M, Walters M, Candlin A, Tisdale M, Morley PJ, Owens IJ, Fenton RJ *et al*: **Zanamivir susceptibility monitoring and characterization of influenza virus clinical isolates obtained during phase II clinical efficacy studies**. *Antimicrobial agents and chemotherapy* 2000, **44**(1):78-87.

91. Hayden FG, Osterhaus AD, Treanor JJ, Fleming DM, Aoki FY, Nicholson KG, Bohnen AM, Hirst HM, Keene O, Wightman K: **Efficacy and safety of the neuraminidase inhibitor zanamivir in the treatment of influenzavirus infections. GG167 Influenza Study Group**. *The New England journal of medicine* 1997, **337**(13):874-880.

92. Hayden FG, Treanor JJ, Fritz RS, Lobo M, Betts RF, Miller M, Kinnersley N, Mills RG, Ward P, Straus SE: **Use of the oral neuraminidase inhibitor oseltamivir in experimental human influenza: randomized controlled trials for prevention and treatment**. *JAMA : the journal of the American Medical Association* 1999, **282**(13):1240-1246.

93. Nicholson KG, Aoki FY, Osterhaus AD, Trottier S, Carewicz O, Mercier CH, Rode A, Kinnersley N, Ward P: **Efficacy and safety of oseltamivir in treatment of acute influenza: a randomised controlled trial. Neuraminidase Inhibitor Flu Treatment Investigator Group**. *Lancet* 2000, **355**(9218):1845-1850.

94. Whalley K: **Probing viral assembly**. *Nature Reviews Drug Discovery* 2006, **5**(4):282-283.

95. Stray SJ, Johnson JM, Kopek BG, Zlotnick A: **An in vitro fluorescence screen to identify antivirals that disrupt hepatitis B virus capsid assembly**. *Nature biotechnology* 2006, **24**(3):358-362.

96. Hadfield AT, Oliveira MA, Kim KH, Minor I, Kremer MJ, Heinz BA, Shepard D, Pevear DC, Rueckert RR, Rossmann MG: **Structural Studies on Human Rhinovirus-14 Drug-Resistant Compensation Mutants**. *J Mol Biol* 1995, **253**(1):61-73.

97. Desmond RA, Accortt NA, Talley L, Villano SA, Soong SJ, Whitley RJ: **Enteroviral meningitis: Natural history and outcome of pleconaril therapy**. *Antimicrobial agents and chemotherapy* 2006, **50**(7):2409-2414.

98. Logan D, Abughazaleh R, Blakemore W, Curry S, Jackson T, King A, Lea S, Lewis R, Newman J, Parry N *et al*: **Structure of a Major Immunogenic Site on Foot-and-Mouth-Disease Virus**. *Nature* 1993, **362**(6420):566-568.

99. Ternois F, Sticht J, Duquerroy S, Krausslich HG, Rey FA: **The HIV-1 capsid protein C-terminal domain in complex with a virus assembly inhibitor**. *Nature structural & molecular biology* 2005, **12**(8):678-682.

100. Tang C, Loeliger E, Kinde I, Kyere S, Mayo K, Barklis E, Sun Y, Huang M, Summers MF: **Antiviral inhibition of the HIV-1 capsid protein**. *J Mol Biol* 2003, **327**(5):1013-1020.

101. Ganser-Pornillos BK, Cheng A, Yeager M: **Structure of full-length HIV-1 CA: a model for the mature capsid lattice**. *Cell* 2007, **131**(1):70-79.

102. Sticht J, Humbert M, Findlow S, Bodem J, Muller B, Dietrich U, Werner J, Krausslich HG: **A peptide inhibitor of HIV-1 assembly in vitro**. *Nature structural & molecular biology* 2005, **12**(8):671-677.

103. Kelly BN, Kyere S, Kinde I, Tang C, Howard BR, Robinson H, Sundquist WI, Summers MF, Hill CP: **Structure of the antiviral assembly inhibitor CAP-1 complex with the HIV-1 CA protein**. *J Mol Biol* 2007, **373**(2):355-366.

104. Johnson JE: **Functional implications of protein-protein interactions in icosahedral viruses**. *Proceedings of the National Academy of Sciences of the United States of America* 1996, **93**(1):27-33.

105. Speir JA, Taylor DJ, Natarajan P, Pringle FM, Ball LA, Johnson JE: **Evolution in action: N and C termini of subunits in related T = 4 viruses exchange roles as molecular switches**. *Structure* 2010, **18**(6):700-709.

106. Barcena J, Verdaguer N, Roca R, Morales M, Angulo I, Risco C, Carrascosa JL, Torres JM, Caston JR: **The coat protein of Rabbit hemorrhagic disease virus contains a molecular switch at the N-terminal region facing the inner surface of the capsid**. *Virology* 2004, **322**(1):118-134.

107. Abrescia NG, Cockburn JJ, Grimes JM, Sutton GC, Diprose JM, Butcher SJ, Fuller SD, San Martin C, Burnett RM, Stuart DI *et al*: **Insights into assembly from structural analysis of bacteriophage PRD1**. *Nature* 2004, **432**(7013):68-74.

108. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**(4):536-540.

109. Zhang Y, Skolnick J: **Scoring function for automated assessment of protein structure template quality**. *Proteins* 2004, **57**(4):702-710.

110. Rossmann MG, Morais MC, Leiman PG, Zhang W: **Combining X-ray crystallography and electron microscopy**. *Structure* 2005, **13**(3):355-362.

111. Frank J: **Single-particle imaging of macromolecules by cryo-electron microscopy**. *Annual review of biophysics and biomolecular structure* 2002, **31**:303-319.

112. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic acids research* 2000, **28**(1):235-242.

113. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Wilson IA, Baker D: **Computational design of proteins targeting the conserved stem region of influenza hemagglutinin**. *Science* 2011, **332**(6031):816-821.

114. Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Mattos C, Myers CA, Kamisetty H, Blair P, Wilson IA *et al*: **Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing**. *Nature biotechnology* 2012, **30**(6):543-548.

115. Rux JJ, Burnett RM: **Spherical viruses**. *Current opinion in structural biology* 1998, **8**(2):142-149.

116. Harrison SC: **The familiar and the unexpected in structures of icosahedral viruses**. *Current opinion in structural biology* 2001, **11**(2):195-199.

117. Johnson JE, Chiu W: **Structures of virus and virus-like particles**. *Current opinion in structural biology* 2000, **10**(2):229-235.

118. Filman DJ, Hogle JM: **Virology. Architecture with a difference**. *Nature* 1991, **351**(6322):100-101.

119. Harrison SC: **What do viruses look like?** *Harvey lectures* 1989, **85**:127-152.

120. Abad-Zapatero C, Abdel-Meguid SS, Johnson JE, Leslie AG, Rayment I, Rossmann MG, Suck D, Tsukihara T: **Structure of southern bean mosaic virus at 2.8 A resolution**. *Nature* 1980, **286**(5768):33-39.

121. Jones EY, Stuart DI, Walker NP: **Structure of tumour necrosis factor**. *Nature* 1989, **338**(6212):225-228.

122. Hosur MV, Schmidt T, Tucker RC, Johnson JE, Gallagher TM, Selling BH, Rueckert RR: **Structure of an insect virus at 3.0 A resolution**. *Proteins* 1987, **2**(3):167-176.

123. Chapman MS, Liljas L: **Structural folds of viral proteins**. *Advances in protein chemistry* 2003, **64**:125-196.

124. Chelvanayagam G, Heringa J, Argos P: **Anatomy and evolution of proteins displaying the viral capsid jellyroll topology**. *J Mol Biol* 1992, **228**(1):220-242.

125. Krupovic M, Bamford DH: **OPINION Virus evolution: how far does the double beta-barrel viral lineage extend?** *Nature Reviews Microbiology* 2008, **6**(12):941-948.

126. Choi HK, Tong L, Minor W, Dumas P, Boege U, Rossmann MG, Wengler G: **Structure of Sindbis virus core protein reveals a chymotrypsin-like serine proteinase and the organization of the virion**. *Nature* 1991, **354**(6348):37-43.

127. Conway JF, Cheng N, Zlotnick A, Wingfield PT, Stahl SJ, Steven AC: **Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy**. *Nature* 1997, **386**(6620):91-94.

128. Rey FA, Heinz FX, Mandl C, Kunz C, Harrison SC: **The Envelope Glycoprotein from Tick-Borne Encephalitis-Virus at 2 Angstrom Resolution**. *Nature* 1995, **375**(6529):291-298.

129. Bahadur RP, Rodier F, Janin J: **A dissection of the protein-protein interfaces in icosahedral virus capsids**. *J Mol Biol* 2007, **367**(2):574-590.

130. Pandit SB, Skolnick J: **Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score**. *BMC bioinformatics* 2008, **9**:531.

131. Xu JR, Zhang Y: **How significant is a protein structure similarity with TM-score=0.5?** *Bioinformatics* 2010, **26**(7):889-895.

132. Gao M, Skolnick J: **Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(52):22517-22522.

133. Murzin AG, Brenner SE, Hubbard T, Chothia C: **Scop - a Structural Classification of Proteins Database for the Investigation of Sequences and Structures**. *J Mol Biol* 1995, **247**(4):536-540.

134. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE: **The ASTRAL Compendium in 2004**. *Nucleic acids research* 2004, **32**(Database issue):D189-192.

135. Donadini R, Liew CW, Kwan AH, Mackay JP, Fields BA: **Crystal and solution structures of a superantigen from Yersinia pseudotuberculosis reveal a jelly-roll fold**. *Structure* 2004, **12**(1):145-156.

136. Varela PF, Romero A, Sanz L, Romao MJ, Topfer-Petersen E, Calvete JJ: **The 2.4 angstrom resolution crystal structure of boar seminal plasma PSP-I/PSP-II: a zona pellucida-binding glycoprotein heterodimer of the spermadhesin family built by a CUB domain architecture**. *J Mol Biol* 1997, **274**(4):635-649.

137. Bahadur RP, Janin J: **Residue conservation in viral capsid assembly**. *Proteins* 2008, **71**(1):407-414.

138. Mintseris J, Weng ZP: **Structure, function, and evolution of transient and obligate protein-protein interactions**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(31):10930-10935.

139. Valdar WSJ, Thornton JM: **Protein-protein interfaces: Analysis of amino acid conservation in homodimers**. *Proteins-Structure Function and Bioinformatics* 2001, **42**(1):108-124.

140. Guharoy M, Chakrabarti P: **Conservation and relative importance of residues across protein-protein interfaces**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15447-15452.

141. Resch-Genger U, Grabolle M, Cavaliere-Jaricot S, Nitschke R, Nann T: **Quantum dots versus organic dyes as fluorescent labels**. *Nat Methods* 2008, **5**(9):763-775.

142. McNerny DQ, Leroueil PR, Baker JR: **Understanding specific and nonspecific toxicities: a requirement for the development of dendrimer-based pharmaceuticals**. *Wires Nanomed Nanobi* 2010, **2**(3):249-259.

143. Jesorka A, Orwar O: **Liposomes: Technologies and Analytical Applications**. *Annu Rev Anal Chem* 2008, **1**:801-832.

144. Giljohann DA, Seferos DS, Daniel WL, Massich MD, Patel PC, Mirkin CA: **Gold Nanoparticles for Biology and Medicine**. *Angew Chem Int Edit* 2010, **49**(19):3280-3294.

145. Fortin JP, Wilhelm C, Servais J, Menager C, Bacri JC, Gazeau F: **Size-sorted anionic iron oxide nanomagnets as colloidal mediators for magnetic hyperthermia**. *Journal of the American Chemical Society* 2007, **129**(9):2628-2635.

146. Qiao RR, Yang CH, Gao MY: **Superparamagnetic iron oxide nanoparticles: from preparations to in vivo MRI applications**. *Journal of Materials Chemistry* 2009, **19**(35):6274-6293.

147. Khemtong C, Kessinger CW, Gao J: **Polymeric nanomedicine for cancer MR imaging and drug delivery**. *Chem Commun (Camb)* 2009(24):3497-3510.

148. Uchida M, Klem MT, Allen M, Suci P, Flenniken M, Gillitzer E, Varpness Z, Liepold LO, Young M, Douglas T: **Biological containers: Protein cages as multifunctional nanoplatforms**. *Adv Mater* 2007, **19**(8):1025-1042.

149. Rice G, Tang L, Stedman K, Roberto F, Spuhler J, Gillitzer E, Johnson JE, Douglas T, Young M: **The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(20):7716-7720.

150. Douglas T, Strable E, Willits D, Aitouchen A, Libera M, Young M: **Protein engineering of a viral cage for constrained nanomaterials synthesis**. *Adv Mater* 2002, **14**(6):415-+.

151. Lewis JD, Destito G, Zijlstra A, Gonzalez MJ, Quigley JP, Manchester M, Stuhlmann H: **Viral nanoparticles as tools for intravital vascular imaging**. *Nat Med* 2006, **12**(3):354-360.

152. Arora PS, Kirshenbaum K: **Nano-tailoring; stitching alterations on viral coats**. *Chemistry & biology* 2004, **11**(4):418-420.

153. Plummer EM, Manchester M: **Viral nanoparticles and virus-like particles: platforms for contemporary vaccine design**. *Wires Nanomed Nanobi* 2011, **3**(2):174-196.

154. Knez M, Bittner AM, Boes F, Wege C, Jeske H, Maiss E, Kern K: **Biotemplate synthesis of 3-nm nickel and cobalt nanowires**. *Nano Lett* 2003, **3**(8):1079-1082.

155. Speir JA, Munshi S, Wang GJ, Baker TS, Johnson JE: **Structures of the Native and Swollen Forms of Cowpea Chlorotic Mottle Virus Determined by X-Ray Crystallography and Cryoelectron Microscopy**. *Structure* 1995, **3**(1):63-78.

156. Pokorski JK, Steinmetz NF: **The Art of Engineering Viral Nanoparticles**. *Mol Pharmaceut* 2011, **8**(1):29-43.

157. Nooren IM, Thornton JM: **Diversity of protein-protein interactions**. *The EMBO journal* 2003, **22**(14):3486-3492.

158. Bravo J, Aloy P: **Target selection for complex structural genomics**. *Current opinion in structural biology* 2006, **16**(3):385-392.

159. Bahadur RP, Chakrabarti P, Rodier F, Janin J: **A dissection of specific and non-specific protein-protein interfaces**. *J Mol Biol* 2004, **336**(4):943-955.

160. Jones S, Thornton JM: **Principles of protein-protein interactions**. *Proceedings of the National Academy of Sciences of the United States of America* 1996, **93**(1):13-20.

161. Sheinerman FB, Norel R, Honig B: **Electrostatic aspects of protein-protein interactions**. *Current opinion in structural biology* 2000, **10**(2):153-159.

162. Bogan AA, Thorn KS: **Anatomy of hot spots in protein interfaces**. *J Mol Biol* 1998, **280**(1):1-9.

163. Morrow JK, Zhang S: **Computational prediction of protein hot spot residues**. *Current pharmaceutical design* 2012, **18**(9):1255-1265.

164. Gao M, Skolnick J: **iAlign: a method for the structural comparison of protein-protein interfaces**. *Bioinformatics* 2010, **26**(18):2259-2265.

165. Gao M, Skolnick J: **Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(52):22517-22522.

166. Shulman-Peleg A, Mintz S, Nussinov R, Wolfson HJ: **Protein-protein interfaces: recognition of similar spatial and chemical organizations**. In: *Algorithms in Bioinformatics.* Berline: Springer; 2004: 194-205.

167. Kolinski A, Skolnick J: **Discretized Model of Proteins .1. Monte-Carlo Study of Cooperativity in Homopolypeptides**. *J Chem Phys* 1992, **97**(12):9412-9426.

168. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R: **A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique**. *J Mol Biol* 1996, **260**(4):604-620.

169. Nussinov R, Wolfson HJ: **Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques**. *Proceedings of the National Academy of Sciences of the United States of America* 1991, **88**(23):10495-10499.

170. Bachar O, Fischer D, Nussinov R, Wolfson H: **A computer vision based technique for 3-D sequence-independent structural comparison of proteins**. *Protein engineering* 1993, **6**(3):279-288.

171. Fischer D, Wolfson H, Lin SL, Nussinov R: **Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding**. *Protein science : a publication of the Protein Society* 1994, **3**(5):769-778.

172. Schmitt S, Kuhn D, Klebe G: **A new method to detect related function among proteins independent of sequence and fold homology**. *J Mol Biol* 2002, **323**(2):387-406.

173. Sander O, Sing T, Sommer I, Low AJ, Cheung PK, Harrigan PR, Lengauer T, Domingues FS: **Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage**. *PLoS computational biology* 2007, **3**(3):e58.

174. Dewerra D: **Algorithmic Graph-Theory - Gibbons,A**. *Eur J Oper Res* 1986, **26**(1):179-179.

175. Kuhn HW: **The Hungarian Method for the assignment problem**. *Nav Res Log* 2005, **52**(1):7-21.

176. Kabsch W: **Solution for Best Rotation to Relate 2 Sets of Vectors**. *Acta Crystallogr A* 1976, **32**(Sep1):922-923.

177. Damodaran KV, Reddy VS, Johnson JE, Brooks CL, 3rd: **A general method to quantify quasi-equivalence in icosahedral viruses**. *Journal of molecular biology* 2002, **324**(4):723-737.

178. Winter C, Henschel A, Kim WK, Schroeder M: **SCOPPI: a structural classification of protein-protein interfaces**. *Nucleic acids research* 2006, **34**(Database issue):D310-314.

179. Alcami A: **Viral mimicry of cytokines, chemokines and their receptors**. *Nature reviews Immunology* 2003, **3**(1):36-50.

180. Hajishengallis G, Lambris JD: **Microbial manipulation of receptor crosstalk in innate immunity**. *Nature reviews Immunology* 2011, **11**(3):187-200.

181. Lopez S, Arias CF: **Rotavirus-host cell interactions: an arms race**. *Current opinion in virology* 2012, **2**(4):389-398.

182. Franzosa EA, Xia Y: **Structural principles within the human-virus protein-protein interaction network**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(26):10538-10543.

183. Alexander-Brett JM, Fremont DH: **Dual GPCR and GAG mimicry by the M3 chemokine decoy receptor**. *The Journal of experimental medicine* 2007, **204**(13):3157-3172.

184. Humphrey W, Dalke A, Schulten K: **VMD: visual molecular dynamics**. *Journal of molecular graphics* 1996, **14**(1):33-38, 27-38.

185. Devreotes PN, Zigmond SH: **Chemotaxis in eukaryotic cells: a focus on leukocytes and Dictyostelium**. *Annual review of cell biology* 1988, **4**:649-686.

186. Proudfoot AE, Handel TM, Johnson Z, Lau EK, LiWang P, Clark-Lewis I, Borlat F, Wells TN, Kosco-Vilbois MH: **Glycosaminoglycan binding and oligomerization are essential for the in vivo activity of certain chemokines**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(4):1885-1890.

187. Lubkowski J, Bujacz G, Boque L, Domaille PJ, Handel TM, Wlodawer A: **The structure of MCP-1 in two crystal forms provides a rare example of variable quaternary interactions**. *Nature structural biology* 1997, **4**(1):64-69.

188. Precious B, Childs K, Fitzpatrick-Swallow V, Goodbourn S, Randall RE: **Simian virus 5 V protein acts as an adaptor, linking DDB1 to STAT2, to facilitate the ubiquitination of STAT1**. *Journal of virology* 2005, **79**(21):13434-13441.

189. Didcock L, Young DF, Goodbourn S, Randall RE: **The V protein of simian virus 5 inhibits interferon signalling by targeting STAT1 for proteasome-mediated degradation**. *Journal of virology* 1999, **73**(12):9928-9933.

190. Scrima A, Konickova R, Czyzewski BK, Kawasaki Y, Jeffrey PD, Groisman R, Nakatani Y, Iwai S, Pavletich NP, Thoma NH: **Structural basis of UV DNA-damage recognition by the DDB1-DDB2 complex**. *Cell* 2008, **135**(7):1213-1223.

191. Shiyanov P, Hayes SA, Donepudi M, Nichols AF, Linn S, Slagle BL, Raychaudhuri P: **The naturally occurring mutants of DDB are impaired in stimulating nuclear import of the p125 subunit and E2F1-activated transcription**. *Molecular and cellular biology* 1999, **19**(7):4935-4943.

192. Leupin O, Bontron S, Strubin M: **Hepatitis B virus X protein and simian virus 5 V protein exhibit similar UV-DDB1 binding properties to mediate distinct activities**. *Journal of virology* 2003, **77**(11):6274-6283.

193. Lin GY, Lamb RA: **The paramyxovirus simian virus 5 V protein slows progression of the cell cycle**. *Journal of virology* 2000, **74**(19):9152-9166.

194. Li T, Chen X, Garbutt KC, Zhou P, Zheng N: **Structure of DDB1 in complex with a paramyxovirus V protein: viral hijack of a propeller cluster in ubiquitin ligase**. *Cell* 2006, **124**(1):105-117.

195. Negrete OA, Levroney EL, Aguilar HC, Bertolotti-Ciarlet A, Nazarian R, Tajyar S, Lee B: **EphrinB2 is the entry receptor for Nipah virus, an emergent deadly paramyxovirus**. *Nature* 2005, **436**(7049):401-405.

196. Bonaparte MI, Dimitrov AS, Bossart KN, Crameri G, Mungall BA, Bishop KA, Choudhry V, Dimitrov DS, Wang LF, Eaton BT *et al*: **Ephrin-B2 ligand is a functional receptor for Hendra virus and Nipah virus**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(30):10652-10657.

197. Bowden TA, Aricescu AR, Gilbert RJ, Grimes JM, Jones EY, Stuart DI: **Structural basis of Nipah and Hendra virus attachment to their cell-surface receptor ephrin-B2**. *Nature structural & molecular biology* 2008, **15**(6):567-572.

198. Chrencik JE, Brooun A, Kraus ML, Recht MI, Kolatkar AR, Han GW, Seifert JM, Widmer H, Auer M, Kuhn P: **Structural and biophysical characterization of the EphB4*ephrinB2 protein-protein interaction and receptor specificity**. *The Journal of biological chemistry* 2006, **281**(38):28185-28192.

199. Lee B, Ataman ZA, Jin L: **Evil versus 'eph-ective' use of ephrin-B2**. *Nature structural & molecular biology* 2008, **15**(6):540-542.

200. Cheng S, Brooks CL, III: **Viral capsid proteins are segregated in structural fold space**. *PLoS computational biology* 2013, **9**(2):e1002905.

201. Li KB: **ClustalW-MPI: ClustalW analysis using distributed and parallel computing**. *Bioinformatics* 2003, **19**(12):1585-1586.

202. Kortemme T, Baker D: **Computational design of protein-protein interactions**. *Current opinion in chemical biology* 2004, **8**(1):91-97.

203. Lippow SM, Wittrup KD, Tidor B: **Computational design of antibody-affinity improvement beyond in vivo maturation**. *Nature biotechnology* 2007, **25**(10):1171-1176.

204. Chevalier BS, Kortemme T, Chadsey MS, Baker D, Monnat RJ, Stoddard BL: **Design, activity, and structure of a highly specific artificial endonuclease**. *Molecular cell* 2002, **10**(4):895-905.

205. Salgado EN, Ambroggio XI, Brodin JD, Lewis RA, Kuhlman B, Tezcan FA: **Metal templated design of protein interfaces**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(5):1827-1832.

206. Stranges PB, Machius M, Miley MJ, Tripathy A, Kuhlman B: **Computational design of a symmetric homodimer using beta-strand assembly**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(51):20562-20567.

207. Keskin O, Tsai CJ, Wolfson H, Nussinov R: **A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications**. *Protein science : a publication of the Protein Society* 2004, **13**(4):1043-1055.

208. Gamblin SJ, Haire LF, Russell RJ, Stevens DJ, Xiao B, Ha Y, Vasisht N, Steinhauer DA, Daniels RS, Elliot A *et al*: **The structure and receptor binding properties of the 1918 influenza hemagglutinin**. *Science* 2004, **303**(5665):1838-1842.

209. **World Health Organization** [http://www.who.int/mediacentre/factsheets/fs211/en/]

210. Wiley DC, Skehel JJ: **The structure and function of the hemagglutinin membrane glycoprotein of influenza virus**. *Annual review of biochemistry* 1987, **56**:365-394.

211. Pica N, Palese P: **Toward a universal influenza virus vaccine: prospects and challenges**. *Annual review of medicine* 2013, **64**:189-202.

212. Wilson IA, Skehel JJ, Wiley DC: **Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 A resolution**. *Nature* 1981, **289**(5796):366-373.

213. Bullough PA, Hughson FM, Skehel JJ, Wiley DC: **Structure of influenza haemagglutinin at the pH of membrane fusion**. *Nature* 1994, **371**(6492):37-43.

214. Fleury D, Daniels RS, Skehel JJ, Knossow M, Bizebard T: **Structural evidence for recognition of a single epitope by two distinct antibodies**. *Proteins* 2000, **40**(4):572-578.

215. Barbey-Martin C, Gigant B, Bizebard T, Calder LJ, Wharton SA, Skehel JJ, Knossow M: **An antibody that prevents the hemagglutinin low pH fusogenic transition**. *Virology* 2002, **294**(1):70-74.

216. Fleury D, Barrere B, Bizebard T, Daniels RS, Skehel JJ, Knossow M: **A complex of influenza hemagglutinin with a neutralizing antibody that binds outside the virus receptor binding site**. *Nature structural biology* 1999, **6**(6):530-534.

217. Fleury D, Wharton SA, Skehel JJ, Knossow M, Bizebard T: **Antigen distortion allows influenza virus to escape neutralization**. *Nature structural biology* 1998, **5**(2):119-123.

218. Julien JP, Lee PS, Wilson IA: **Structural insights into key sites of vulnerability on HIV-1 Env and influenza HA**. *Immunological reviews* 2012, **250**(1):180-198.

219. Throsby M, van den Brink E, Jongeneelen M, Poon LL, Alard P, Cornelissen L, Bakker A, Cox F, van Deventer E, Guan Y *et al*: **Heterosubtypic neutralizing monoclonal antibodies**

cross-protective against H5N1 and H1N1 recovered from human IgM+ memory B cells. *PloS one* 2008, **3**(12):e3942.

220. Kashyap AK, Steel J, Oner AF, Dillon MA, Swale RE, Wall KM, Perry KJ, Faynboym A, Ilhan M, Horowitz M *et al*: **Combinatorial antibody libraries from survivors of the Turkish H5N1 avian influenza outbreak reveal virus neutralization strategies**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(16):5986-5991.

221. Ekiert DC, Bhabha G, Elsliger MA, Friesen RH, Jongeneelen M, Throsby M, Goudsmit J, Wilson IA: **Antibody recognition of a highly conserved influenza virus epitope**. *Science* 2009, **324**(5924):246-251.

222. Krissinel E, Henrick K: **Inference of macromolecular assemblies from crystalline state**. *J Mol Biol* 2007, **372**(3):774-797.

223. Cheng S, Zhang Y, Brooks CL, III: **PCalign: a method to quantify physicochemical similarity of protein-protein interfaces**. *(submitted)* 2014.

224. Lee MS, Feig M, Salsbury FR, Jr., Brooks CL, III: **New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations**. *Journal of computational chemistry* 2003, **24**(11):1348-1356.

225. Brooks BR, III CLB, Jr ADM, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S *et al*: **CHARMM: The biomolecular simulation program**. *Journal of computational chemistry* 2009, **30**(10):1545-1614.

226. Pierce BG, Hourai Y, Weng Z: **Accelerating protein docking in ZDOCK using an advanced 3D convolution library**. *PloS one* 2011, **6**(9):e24657.

227. Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R, Weng Z: **Integrating statistical pair potentials into protein complex prediction**. *Proteins* 2007, **69**(3):511-520.

228. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L: **The FoldX web server: an online force field**. *Nucleic acids research* 2005, **33**(Web Server issue):W382-388.

229. Mosca R, Ceol A, Aloy P: **Interactome3D: adding structural details to protein networks**. *Nat Methods* 2013, **10**(1):47-53.

230. Aloy P, Russell RB: **Interrogating protein interaction networks through structural biology**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(9):5896-5901.

231. Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin AC, Bork P, Superti-Furga G, Serrano L *et al*: **Structure-based assembly of protein complexes in yeast**. *Science* 2004, **303**(5666):2026-2029.

232. Guerler A, Govindarajoo B, Zhang Y: **Mapping monomeric threading to protein-protein structure prediction**. *Journal of chemical information and modeling* 2013, **53**(3):717-725.

233. Mukherjee S, Zhang Y: **Protein-protein complex structure predictions by multimeric threading and template recombination**. *Structure* 2011, **19**(7):955-966.

234. Lu L, Lu H, Skolnick J: **MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading**. *Proteins* 2002, **49**(3):350-364.

235. Chen H, Skolnick J: **M-TASSER: an algorithm for protein quaternary structure prediction**. *Biophysical journal* 2008, **94**(3):918-928.

236. Hosur R, Peng J, Vinayagam A, Stelzl U, Xu J, Perrimon N, Bienkowska J, Berger B: **A computational framework for boosting confidence in high-throughput protein-protein interaction datasets**. *Genome biology* 2012, **13**(8):R76.

237. Singh R, Park D, Xu J, Hosur R, Berger B: **Struct2Net: a web service to predict protein-protein interactions using a structure-based approach**. *Nucleic acids research* 2010, **38**(Web Server issue):W508-515.

238. Kundrotas PJ, Lensink MF, Alexov E: **Homology-based modeling of 3D structures of protein-protein complexes using alignments of modified sequence profiles**. *International journal of biological macromolecules* 2008, **43**(2):198-208.

239. Kundrotas PJ, Alexov E: **Predicting 3D structures of transient protein-protein complexes by homology**. *Biochimica et biophysica acta* 2006, **1764**(9):1498-1511.

240. Fukuhara N, Kawabata T: **HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures**. *Nucleic acids research* 2008, **36**(Web Server issue):W185-189.

241. Tyagi M, Hashimoto K, Shoemaker BA, Wuchty S, Panchenko AR: **Large-scale mapping of human protein interactome using structural complexes**. *EMBO reports* 2012, **13**(3):266-271.

242. Hosur R, Xu J, Bienkowska J, Berger B: **iWRAP: An interface threading approach with application to prediction of cancer-related protein-protein interactions**. *J Mol Biol* 2011, **405**(5):1295-1310.

243. Zhang QC, Petrey D, Garzon JI, Deng L, Honig B: **PrePPI: a structure-informed database of protein-protein interactions**. *Nucleic acids research* 2013, **41**(Database issue):D828-833.

244. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T *et al*: **Structure-based prediction of protein-protein interactions on a genome-wide scale**. *Nature* 2012, **490**(7421):556-560.

245. Sinha R, Kundrotas PJ, Vakser IA: **Protein docking by the interface structure similarity: how much structure is needed?** *PloS one* 2012, **7**(2):e31349.

246. Liddington RC, Yan Y, Moulai J, Sahli R, Benjamin TL, Harrison SC: **Structure of simian virus 40 at 3.8-A resolution**. *Nature* 1991, **354**(6351):278-284.

247. Stehle T, Yan Y, Benjamin TL, Harrison SC: **Structure of murine polyomavirus complexed with an oligosaccharide receptor fragment**. *Nature* 1994, **369**(6476):160-163.

248. Gipson P, Baker ML, Raytcheva D, Haase-Pettingell C, Piret J, King JA, Chiu W: **Protruding knob-like proteins violate local symmetries in an icosahedral marine virus**. *Nature communications* 2014, **5**:4278.

249. Milligan RA: **Protein-protein interactions in the rigor actomyosin complex**. *Proceedings of the National Academy of Sciences of the United States of America* 1996, **93**(1):21-26.

250. Bocanegra R, Rodriguez-Huete A, Fuertes MA, Del Alamo M, Mateu MG: **Molecular recognition in the human immunodeficiency virus capsid and antiviral design**. *Virus research* 2012, **169**(2):388-410.

251. ElSawy KM, Twarock R, Verma CS, Caves LS: **Peptide inhibitors of viral assembly: a novel route to broad-spectrum antivirals**. *Journal of chemical information and modeling* 2012, **52**(3):770-776.

252. Bennett MJ, Choe S, Eisenberg D: **Domain swapping: entangling alliances between proteins**. *Proceedings of the National Academy of Sciences of the United States of America* 1994, **91**(8):3127-3131.

253. Liu Y, Eisenberg D: **3D domain swapping: as domains continue to swap**. *Protein science : a publication of the Protein Society* 2002, **11**(6):1285-1299.

254. Rousseau F, Schymkowitz J, Itzhaki LS: **Implications of 3D domain swapping for protein folding, misfolding and function**. *Advances in experimental medicine and biology* 2012, **747**:137-152.

255. Perica T, Chothia C, Teichmann SA: **Evolution of oligomeric state through geometric coupling of protein interfaces**. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**(21):8127-8132.

256. Tokuriki N, Oldfield CJ, Uversky VN, Berezovsky IN, Tawfik DS: **Do viral proteins possess unique biophysical features?** *Trends in biochemical sciences* 2009, **34**(2):53-59.

257. Johnson JE, Munshi S, Liljas L, Agrawal D, Olson NH, Reddy V, Fisher A, McKinney B, Schmidt T, Baker TS: **Comparative studies of T = 3 and T = 4 icosahedral RNA insect viruses**. *Archives of virology Supplementum* 1994, **9**:497-512.

258. Nguyen HD, Brooks CL, III: **Generalized structural polymorphism in self-assembled viral particles**. *Nano letters* 2008, **8**(12):4574-4581.

259. Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF: **From protein sequence to dynamics and disorder with DynaMine**. *Nature communications* 2013, **4**:2741.