COMPUTATIONAL METHODS FOR PROTEIN STRUCTURE COMPARISON AND ANALYSIS

by

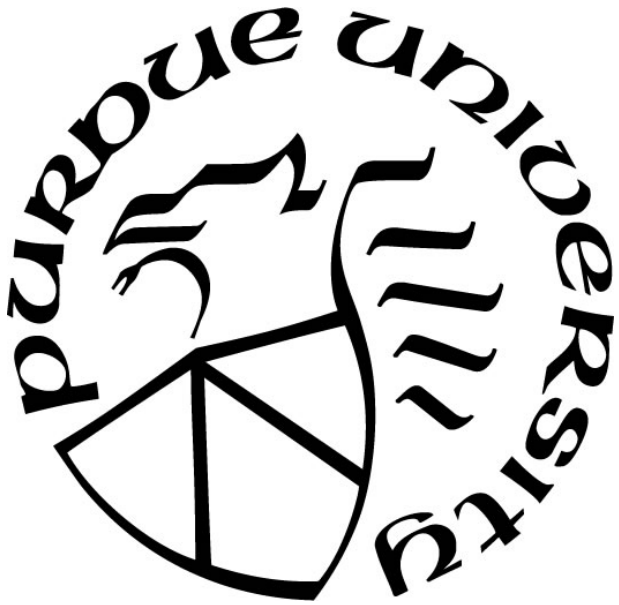
Xusi Han

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy



Department of Biological Sciences

West Lafayette, Indiana

May 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Daisuke Kihara, Chair

Department of Biological Sciences and Department of Computer Science

Dr. Michael Gribskov

Department of Biological Sciences

Dr. Wen Jiang

Department of Biological Sciences

Dr. Cynthia Stauffacher

Department of Biological Sciences

Approved by:

Dr. Jason Evans

To my mother, for always loving and supporting me

ACKNOWLEDGMENTS

First and foremost, I am sincerely thankful to my advisor, Professor Daisuke Kihara, for guiding me and teaching me through my Ph.D. journey. Under his guidance, I have learnt and finished much more than what I have expected at the beginning of my graduate study.

I also thank my committee members, Professors Michael Gribskov, Wen Jiang, and Cynthia Stauffacher for their guidance, encouragement and suggestions. They lent me their expertise to provide the context, guidance, and the occasional reality check as my pursed this research.

I am also grateful to all the members of Kihara lab: Tunde Aderinwale, Eman Alnabati, Siyang Chen, Charles Christoffer, Ziyun Ding, Juan Esquivel-Rodriguez, Aashish Jain, Xuejiao Kang, Ishita Khan, Hyung-Rae Kim, Lyman Monroe, Lenna Peterson, Sai Maddhuri, Daipayan Sarkar, Woong-Hee Shin, Atilla Sit, Genki Terashi, Yoichiro Togawa, Jacob Verburgt, Sean Flannery, Xiao Wang, Qing Wei, Yi Xiong, Jian Zhang, and Xiaolei Zhu. Thanks for all the questions, discussions, coding help, encouragement, and so many hours spent together.

I also deeply appreciate the supports and accompany from all my friends: Ninghai Gan, Yingyuan Sun, Zheng Xing, Boning Zhang, Xiangying Mao, Dan Xie, Heng Wu, Longfei Wang, Ruoxing Wang, Chen Li, and Jiemin Zhao. They help me go through the hard times and give me the confidence to be myself.

Finally, I need to thank my family: my parents, grandparents, uncles, and cousins. I am incredibly thankful to my mother, Xiumei Ma, for always loving and supporting me in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3D	three-dimensional
3DZD	3D Zernike descriptor
Å	Angstrom
ABC	ATP-binding cassette
ADP	adenosine 5'-diphosphate
AFP	aligned fragment pairs
BU	biological unit
CATH	Class/Architecture/Topology/Homology
CC	cross-correlation
CE	combinatorial extension
DED	direct electron detector
DDSD	Diffusion Distance Shape Descriptor
EM	electron microscopy
EMDB	Electron Microscopy Data Bank
ET	electron tomography
FAD	flavin-adenine dinucleotide
FFT	fast Fourier transform
FNR	False Negative Rate
FSSP	Families of Structurally Similar Proteins
FT	first tier
GDF	Gaussian distribution function
GMM	Gaussian mixture model
HAPT	Histograms of Area Projection Transform
iDR	incremental distance rank
IDSS	Inner Distance Shape Signature
LAD	Local Average Distance
LDP	local dense point
MAP	mean average precision
MVEE	minimum volume enclosing ellipsoid

NMR	nuclear magnetic resonance
NN	nearest neighbor
PC	principal component
PCA	principal component analysis
PCNA	proliferating cell nuclear antigen
PDB	Protein Data Bank
RMSD	root-mean-square deviation
SCOP	Structural Classification of Proteins
SCOPe	Structural Classification of Proteins – extended
SES	solvent excluded surface
SHREC	SHape REtrieval Contest
SNR	signal-to-noise ratio
ST	second tier
TP	True Positive
VESPER	VEctor-based local SPace ElectRon density map alignment

ABSTRACT

Proteins are involved in almost all functions in a living cell, and functions of proteins are realized by their tertiary structures. Protein three-dimensional structures can be solved by multiple experimental methods, but computational approaches serve as an important complement to experimental methods for comparing and analyzing protein structures. Protein structure comparison allows the transfer of knowledge about known proteins to a novel protein and plays an important role in function prediction. Obtaining a global perspective of the variety and distribution of protein structures also lays a foundation for our understanding of the building principle of protein structures. This dissertation introduces our computational method to compare protein 3D structures and presents a novel mapping of protein shapes that represents the variety and the similarities of 3D shapes of proteins and their assemblies. The methods developed in this work can be applied to obtain new biological insights into protein atomic structures and electron density maps.

CHAPTER 1. INTRODUCTION

Proteins are the most versatile macromolecules in living systems and serve crucial functions in essentially all biological processes [1, 2]. They can function as enzymes, transport and store other molecules, provide mechanical support, and regulate growth and differentiation. This incredible array of functions derives from the diverse set of three-dimensional (3D) structures, which are determined by amino acid sequences. Comparison and analysis of the 3D structure of proteins using an appropriate representation is thus crucial for understanding the universe of protein structure, function, and evolution [3].

Protein atomic structures can be determined by different experimental techniques. Computational approaches are also an indispensable part of protein structure analysis. They help tackle problems that are experimentally intractable and make testable hypotheses to guide experimental work. Computational approaches have applications in many different areas, including database search [4-7], sequence and structure alignment [8-12], and protein function prediction [13-18]. This work expands the ways in which we can compare and analyze protein structures.

1.1 Multiple levels of protein structures

Protein 3D structure is organized at four levels: primary, secondary, tertiary, and quaternary (Figure 1.1). The primary structure of a protein refers to the specific sequence of amino acids. Amino acids are held together by peptide bond, where the carboxyl group of one amino acid reacts with the amino group of the other amino acid and causes the release of a molecule of water.

Secondary structure refers to the characteristic folding of the polypeptide backbone. There are two main types of secondary structure, α-helix and β-sheet. In α-helix, the polypeptide backbone forms a repeating helical structure that is stabilized by hydrogen bonds occurring at regular intervals. In β-sheet, the structure is stabilized by hydrogen bonds formed between the carbonyl oxygen and the amine hydrogen of amino acid in adjacent strands.

Tertiary structure refers to the 3D structure of a polypeptide, which results from the interactions between the side chains of amino acids. It is the complete structure for protein with only one polypeptide chain, or single-chain protein. Quaternary structure is for protein complex,

which is composed of two or more polypeptide chains. Each polypeptide chain adopts its own tertiary structure and then assemble with each other via intermolecular interactions.

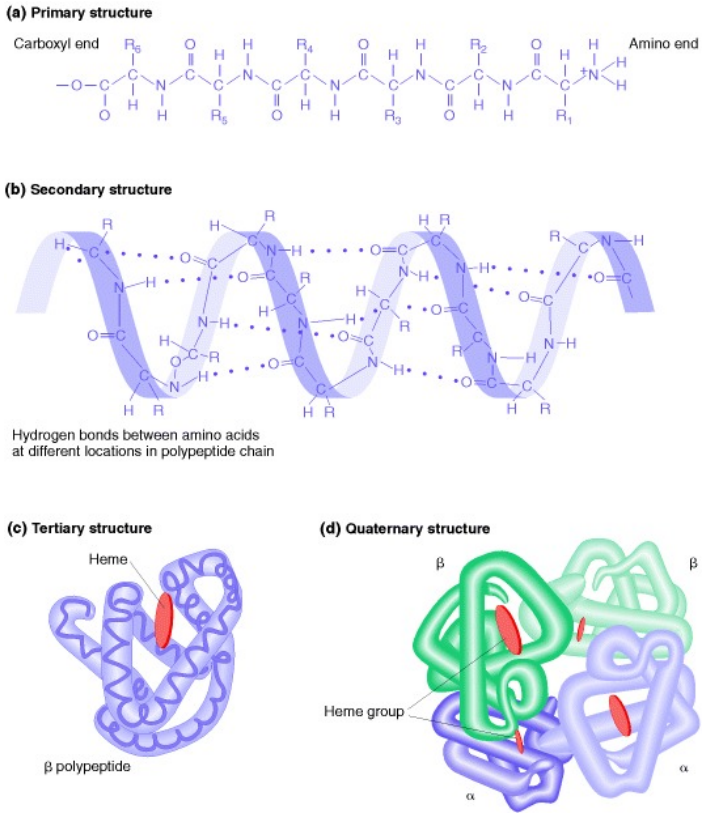


Figure 1.1: Four levels of protein structure. Figure is taken from [19].

A single-chain protein or a subunit of a protein complex can be further divided into domains, which are conserved part of a given protein sequence and tertiary structure that can fold and function in isolation. The concept of a domain was first used to describe the spatially distinct structural subunits in lysozyme [20] and ribonuclease [21]. Researchers have then gradually discovered that domains can recur either in different structural contexts or in multiple copies in the same polypeptide chain [22].

There are three main databases that classify individual domains of protein structures: the Structural Classification of Proteins (SCOP) database [23-25], the Class/Architecture/Topology/Homology (CATH) database [26], and the Families of Structurally Similar Proteins (FSSP) database [27]. SCOP is based on the visual inspection of folds and the manual curation of corresponding groups. This is a hierarchical classification with seven levels:

Class, Fold, Superfamily, Family, Protein, Species, and Domain. CATH database is constructed in a similar principle and classifies domains into four levels, as suggested in the name of the database. FSSP database is based on fully automated fold classification that groups proteins into clusters sharing similar folds. There is a large percentage of agreement between the three databases, although their philosophy and design are different [28].

1.2 Experimental structural biology

Structural biology aims to provide a comprehensive understanding of how molecular architecture performs the biological functions that are central to life. Atomic structures of biological macromolecules are often determined by one of the three major methods: X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM) [29]. The Protein Data Bank (PDB) is the central repository of three-dimensional macromolecular structural data [30]. As of December 2019, there were 158,549 entries in PDB, of which 89.0% (141,165 entries) were determined by X-ray crystallography, 8.1% (12,835 entries) by NMR, and 2.6% (4,092 entries) by cryo-EM.

1.2.1 X-ray crystallography

X-ray crystallography is the experimental technique determining the three-dimensional protein structure by X-ray diffraction of its crystals. After crystallization of the protein of interest, the crystal is placed in an intense beam of X-rays, producing regular pattern of reflections. A three-dimensional electron density map can be created by measuring the intensities of the diffraction pattern. From this electron density map, the mean positions of the atoms in the crystal can be determined. One example of protein atomic structure solved by X-ray crystallography is shown in Figure 1.2. The structure file shows x, y, and z coordinates of each atom in the protein.

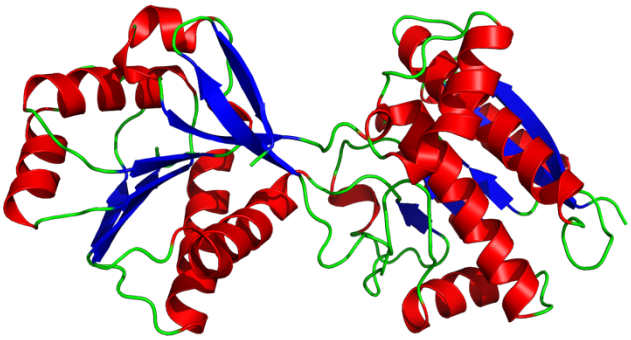


Figure 1.2: Example of protein structure solved by X-ray crystallography. This structure is 1usg-A [31], leucine-binding protein from *E. coli.* α-helix and β-sheet are colored in red and blue, respectively.

This technique originates from the discovery of X-ray by Wilhelm Roentgen in 1895. Subsequently, Max von Laue showed X-ray diffraction pattern of crystals in 1912, and William Lawrence Bragg derived the Bragg's Law and showed that X-ray diffraction could be used in the atomic structure determination [32]. The first crystal structure of a macromolecule was solved in 1957 by John Kendrew [33]. Since then, several other crystallographic structures, including the structure of vitamin B12 [34], insulin [35], and photosynthetic reaction center [36, 37], have been solved and awarded the Nobel Prize.

High-quality crystal is required by X-ray crystallography for structural analysis. If the target protein is difficult to arrange in crystals, it is impossible to understand the atomic structure of the biomolecule.

1.2.2 Nuclear magnetic resonance spectroscopy

NMR spectroscopy is a powerful tool to study the structure, dynamics, and interactions of biological macromolecules. The discovery of nuclear magnetic resonance was made by two groups (Felix Bloch *et al.* [38] and Edward Purcell *et al.* [39]) independently in 1946. A few years later, Proctor and Yu discovered that two different signals were generated by two nitrogens in NH_4NO_3 [40]. This chemical shift observation was confirmed later when three lines were detected in the spectrum of ethanol. The first published protein NMR spectrum was recorded at 40 MHz for bovine pancreatic ribonuclease in 1957 [41]. Two-dimensional NMR was introduced by Richard Ernst in 1976, who discovered that transient signals could be converted to a normal spectrum by

Fourier transformation [42]. This led the way to the introduction of multidimensional techniques ten years later [43].

One remarkable advantage of NMR over X-ray crystallography is that it can be used to obtain dynamic information of macromolecules. Figure 1.3 shows 20 conformers of the N-terminal docking domain in NRPS subunit solved by NMR. But NMR spectroscopy has limited applications for small proteins. As the molecular weight increases, the width of resonance also increases, which eventually makes it impossible to assign resonance because of spectral overlap.

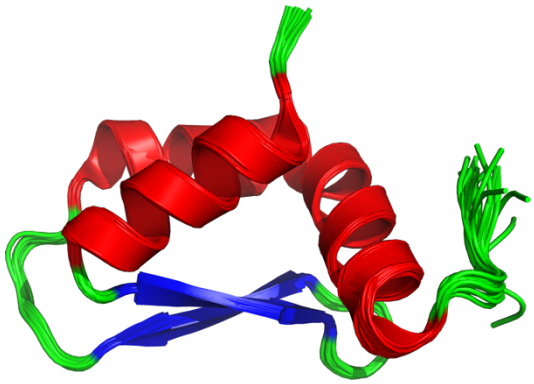


Figure 1.3: Example of protein structure solved by NMR spectroscopy. This figure shows 20 conformers of N-terminal docking domain in NRPS subunit (PDB ID: 6ewu [44]). Color code for secondary structures are the same as Figure 1.2.

1.2.3 Electron microscopy

cryo-electron microscopy (cryo-EM) is a powerful tool for the determination of macromolecular structures in multiple conformations in their native environment. In this method, molecules are trapped in random orientations in amorphous ice by plunge-freezing. Images of these molecules, in their different orientations, are recorded as projections in the electron microscope at low temperature and averaged to generate enhanced image, which are then used to reconstruct the 3D structure. Figure 1.4 shows one example of protein structure solved by cryo-EM. The map file shows the electron density value at each voxel in 3D space.

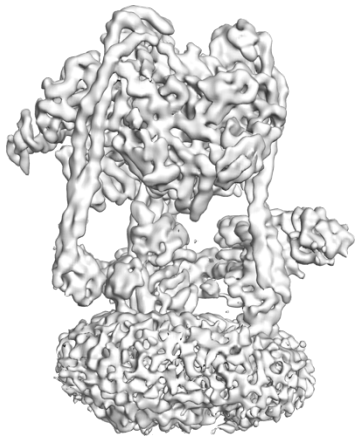


Figure 1.4: Example of protein structure solved by cryo-EM. This structure is for the V-ATPase:SidK complex in yeast (EMD-8724 [45]).

Dubochet and coworkers introduced a unique sample preparation method in 1980s to preserve specimens at near native condition within a thin amorphous ice film [46, 47]. This marks the beginning of modern cryo-EM. There are two major innovations that lead to recent advances in cryo-EM: the employment of direct electron detector (DED), and improvement of image processing methods and microprocessor performance. These two technologies have led to a tremendous success in improving the resolution of EM structures [48]. Cryo-EM can be used for structure determination of macromolecular complexes across a wide molecular mass range from tens kilodaltons to several hundreds of megadaltons [49-52]. The lower molecular weight limitation is expected around 38 kDa, which is calculated from estimates of the signal-to-noise ratio (SNR) considering the limiting dose rate [53].

Unlike X-ray crystallography and NMR spectroscopy, cryo-EM requires a much smaller amount of sample and it accepts a larger variation of specimen types. It does not require crystallization of samples. As biological molecules generally form an intact structure in fully hydrated or in partly hydrated form, cryo-EM is an ideal tool to observe such structures in their near-native environment.

1.2.4 Electron tomography

Cells are complex environments that are populated with millions of proteins [54]. As introduced in previous sections, structural biologists can solve the structure of isolated proteins or complexes using X-ray crystallography, NMR spectroscopy, and cryo-EM. To retrieve structural

information from within the cell or organelle environment directly, researchers have developed a number of techniques to ensure that the macromolecular complex is studied in its intact state. Fluorescence microscopy is the most commonly used technique [55], but the level of detail is limited as the signal comes from the fluorophore used for detection rather than from the target protein. An alternative method is cryo-electron tomography (cryo-ET), which provides a more static image but at higher resolutions. With the development of direct detectors and the associated resolution revolution, cryo-ET can visualize the cellular architecture and the structural details of macromolecular complexes three-dimensionally. Figure 1.5 shows a 2D slice of the tomogram that describes the dorsal closure event in fly embryos.

The biological samples imaged by cryo-ET are sensitive to beam-induced radiation, which limits the maximal resolution of individual tomograms. One common approach to increase resolution is to average volumes of particle, which requires correct localization and identification of specific particles in the first place. Manual localization and classification are rarely feasible, due to the low signal-to-noise ratio and the large amount of data. Instead, researchers have developed many machine learning approaches for automated particle localization and classification within a tomogram [56-60].

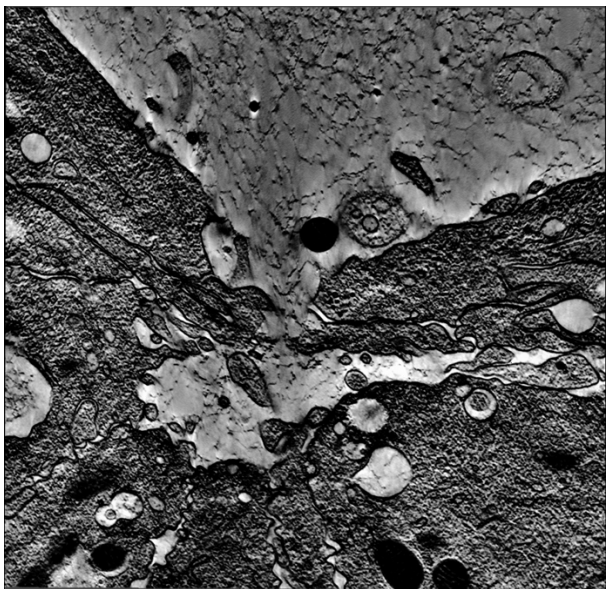


Figure 1.5: Example of structure solved by cryo-ET. This image shows a 2D slice of the tomogram that describes the dorsal closure event in fly embryos (EMD-2610 [61]).

1.3 Existing methods for protein structure comparison

The problem of quantifying the similarity between two protein structures is nontrivial and continues evolving. Proteins are flexible molecules, which admits flexible motions related to their biological functions. Consequently, quantifying the structural differences in a sensible way becomes essential. With the rapid increase in the amount of protein structures in the Protein Data Bank (PDB) [30] and the Electron Microscopy Data Bank (EMDB) [62], it is also necessary to develop methods to quantify structure similarity and allow efficient structure-based search against the entire database in real-time.

1.3.1 Atomic structure comparison

Protein sequence alignment is a standard technique in bioinformatics to find the best-matching piecewise alignments of two query protein amino acid sequences. Computational approaches to sequence alignment can be of two types: global alignment and local alignment. Global alignment tries to find the alignment that spans the entire length of both query sequences, while local alignment identifies local similar regions within long sequences. The classical global sequence alignment method is the Needleman-Wunsch algorithm, which is based on dynamic programming. The Smith-Waterman algorithm is a general local alignment method also based on dynamic programming.

As protein structure is more conserved than protein sequences, protein structure comparison is much more important than protein sequence comparison and plays an important role in predicting functions of novel proteins [63]. Protein structure comparison methods can be divided into alignment methods and non-alignment methods. Some commonly used alignment methods include TM-Align [11], MM-Align [8], CE method [9], DALI [64], and FATCAT [65]. Root-mean-square deviation (RMSD) is the most commonly used metric to quantify the similarity of two superimposed atomic structures, but it is strongly affected by the most deviated fragments in two proteins and also requires prior assignment of atom correspondences from structure alignment. As the structural alignment procedure is time consuming, it is unfeasible for real-time search against the entire protein structure database.

The second type of protein structure comparison methods is non-alignment methods, where protein structure similarity is quantified by the similarity of the descriptors of their molecular

shapes without any alignment. Our lab has applied 3D Zernike descriptor (3DZD) to capture the global shape information in protein atomic structures [4, 5, 66]. 3DZD is a vector derived from a series expansion of a 3D function, which describes protein structure in a compact and rotation-invariant fashion. Similarity between two proteins is quantified by the Euclidean distance of their 3DZD vectors. Omokage search adopts a different descriptor to quantify the protein structure similarity. In Omokage search, each protein is instead converted to four types of one-dimensional (1D) profiles [6]. It firstly converts each protein structure into a set of representative 3D points using vector quantization. Four types of 1D profiles are then calculated from the 3D point models: P_{30}, P_{50}, P_{o25}, and P_{PCA}. The first three profiles are incremental distance rank (iDR) profiles. The numbers of 3D points are set to 30 and 50 for P_{30} and P_{50}, respectively. P_{o25} is calculated by the outermost 25 points among the 50 points. The last profile, P_{PCA}, derives from the principal component analysis (PCA) of the 50 point set and describes the standard deviations along the first, second, and third principal axes. Comparison of protein structure similarity is performed using the similarity of those four types of 1D profiles.

The descriptors discussed above treat proteins as rigid bodies. Some researchers have also developed shape descriptors that take flexibility of protein structures into consideration. Liu *et al.* developed the Inner Distance Shape Signature (IDSS), which measured the length of the shortest path between landmark points within the molecular shape [67]. As IDSS is sensitive to shape deformation of molecules with topological changes, they later developed Diffusion Distance Shape Descriptor (DDSD), which calculated the average length of paths connecting two landmark points [68]. Wang *et al.* proposed Local Average Distance (LAD) based on either geodesic distances or Euclidean distances for pairwise flexible protein structure comparison [69].

1.3.2 Electron microscopy map comparison

With the recent development in cryo-EM technology, there is a rapid accumulation of EM maps in the Electron Microscopy Data Bank [62]. This creates a demand for better methods to analyze the data, including improved scores for comparison, classification and integration of data at different resolutions. Different from the atomic structures stored in PDB, EM maps do not provide the 3D coordinates of each atom in the protein. Instead, the 3D shape is represented as electron density for each voxel in 3D space. Therefore, it is unfeasible to directly apply the atomic structure comparison methods to comparison of EM maps.

There are a set of scoring functions that compare EM maps [70]. Those scoring functions can be further divided into three categories: density-based scores, surface-based scores, and overlap-based scores. Density-based scores include cross correlation and mutual information. Cross correlation can be calculated for either all voxels in two maps or only the overlap region. Mutual information measures statistical relationship between the two binned densities based on their joint entropy [71, 72]. Similar to cross correlation, mutual information can also be calculated for all voxels or for the region of overlap. Surface-based scores include the Chamfer Distance and Normal Vector score. Chamfer Distance is the average Euclidean distance between nearest surface points from two maps [72, 73]. Normal Vector score is calculated as the average angle between the normal vectors at aligned surface points [72, 74]. Overlap-based score relies on quantifying the overlap regions between the two maps. The Overlap score is calculated as the fraction of overlapping voxels within the contour level with respect to the smaller of the two volumes.

EM map superimposition is required prior to the calculation of density-based scores, surface-based scores, and overlap-based scores. Same as alignment of atomic structures, EM map alignment is also a time-consuming process and unsuitable for real-time database search. Researchers have also developed non-alignment methods for EM map comparison. Our lab has developed a compact fingerprint representation of EM maps based on the 3D Zernike descriptor, which offers five options for contour shape representation [7, 66]. Omokage search also supports the similarity search for EM maps using four types of 1D profiles.

1.4 Contribution

The major contribution of this work is to expand both the breadth and the depth of computational techniques related to protein structure comparisons. Some previous studies have mapped protein structures into low-dimensional space using either structural similarity or FragBag vector [75-77]. As protein-protein interactions are established by surface residues, analysis of protein universe in shape level is more relevant to the molecular environment in the cell. Viewing protein universe in terms of global surface is also valuable to protein design field, which has wide applications in therapeutics [78, 79]. To design inhibitors of arbitrary targets, shape complementarity is a critical consideration [80]. Chapter 2 presents a novel mapping of protein shapes that represents the variety and the similarities of 3D shapes of proteins and their assemblies. This mapping provides various novel insights into protein shapes including determinant factors of

protein 3D shapes, which enhance our understanding of the design principles of protein shapes. The mapping will also be a valuable resource for artificial protein design as well as references for classifying medium- to low-resolution protein structure images of determined by cryo-electron microscopy and tomography.

Protein global surface shape information was captured by 3D Zernike descriptors in Chapter 2. As 3D Zernike descriptors is inappropriate for local structure matching, it is necessary to develop a new algorithm that is capable of matching protein subunit into the complete protein structure. Comparison of EM maps also differs from the comparison of atomic structures. As atom coordinate information is not provided in EM maps, we cannot directly apply traditional structure alignment methods to superimpose the subunit onto the complete structure. Chapter 3 introduces a local vector-based algorithm named VESPER for global and local matching of EM maps. Unlike existing methods that merely matches density values and the shape of maps, VESPER captures similarity of underlying structures embedded in maps by taking local gradient directions into consideration. Compared to existing methods, VESPER achieved substantially more accurate global and local alignment of maps as well as database retrieval based on benchmark datasets. Chapter 4 demonstrates the performance of our protein shape comparison methods on the community-wide assessment of 3D-shape retrieval algorithms. We have integrated 3D Zernike descriptors with neural networks to improve the retrieval performance of protein atomic structures. In addition, we propose a new approach for localization and classification of particles in cryo-electron tomograms.

With the rapid accumulation of protein structures in PDB and EMDB, it is necessary to develop tools to search the structure database efficiently. Finally, in Chapter 5, we introduce two web-based tools, 3D-SURFER and EM-SURFER, for real-time comparison and analysis of protein atomic structures and EM maps. The structure database in both servers is updated every week to incorporate information of new structures. Taking an atomic structure or an electron microscopy map of a protein or a protein complex as input, the 3D Zernike descriptors of a query protein is computed and compared with the 3DZD of all other proteins in PDB or EMDB.

Proteins are the major molecules involved in almost all cellular processes. Comparison of protein structures allows the transfer of knowledge about known proteins to a novel protein and thus plays an important role in predicting functions of novel proteins. The research described in

these chapters deepens our understandings of protein structure universe and increases the capabilities protein structure matching via conceptual and practical developments.

CHAPTER 2. GLOBAL MAPPING OF PROTEIN SHAPE SPACE USING 3D ZERNIKE DESCRIPTORS[1]

Proteins are involved in almost all functions in a living cell, and functions of proteins are realized by their tertiary structures. Obtaining a global perspective of the variety and distribution of protein structures lays a foundation for our understanding of the building principle of protein structures. In light of the rapid accumulation of low-resolution structure data from electron tomography and cryo-electron microscopy, here we map and classify three-dimensional (3D) surface shapes of proteins into a similarity space. Surface shapes of proteins were represented with 3D Zernike descriptors, mathematical moment-based invariants, which have previously been demonstrated effective for biomolecular structure similarity search. In addition to single chains of proteins, we have also analyzed the shape space occupied by protein complexes. From the mapping, we have obtained various new insights into the relationship between shapes, main-chain folds, and complex formation. The unique view obtained from shape mapping opens up new ways to understand design principles, functions, and evolution of proteins.

2.1 Background

Proteins are the primary workers in a living cell, involved in transportation, catalysis, signaling, energy production, and many other processes. Classification of protein structures provides fundamental information for our understanding of the principles that govern and determine protein structures, which is one of the essential goals of structural biology and protein bioinformatics. Understanding the repertoire of protein structures is also of practical importance for artificial protein design, which has broad applications in therapeutics such as designing inhibitors [82] and small peptide drugs [80], as well as the development of biomaterials [78].

Conventionally, protein structures have been classified based on their main-chain conformations and evolutionary history [83-85]. Such classifications led to several important observations including the number of different protein folds in nature [86-88], distributions of folds in genomes [89, 90], and the relationship between sequence and structure conservations [91]. The discovery of the limited number of folds yielded stimulating discussions on the mechanism

[1]Portions of this chapter have been previously published [81]

behind it [92, 93]. Furthermore, such studies contributed to the birth of a very successful paradigm of threading [94] and more recent fragment-based approaches [95] in protein structure prediction.

Some recent studies mapped protein structures into a low-dimensional space to reveal high-level organization of the variety of protein structures. Kim and his colleagues computed structural similarity with DALI, a residue-contact map-based structure comparison method [96], and mapped representative proteins into a 3D space using multidimensional scaling [76, 77]. Osadchy and Kolodny represented protein structure domains as a vector indicating the occurrence of fragments in the structure [75]. In both works, the maps exhibited a trend where structures formed clusters according to their fold classes, α, β, α/β, and $\alpha+\beta$, and others, which is reasonable but expected.

Here, we present a global mapping of 3D surface shapes of single proteins and complexes. In contrast to the previous works [75-77] that considered main-chain conformation to define the structural similarity, the use of surface shape representation led to findings of previously undescribed relationships between protein shape, fold class, and assemblies. We perform a thorough analysis of surface shapes in consideration of the rise of medium- to low-resolution structures determined by electron tomography [97] and cryo-electron microscopy (cryo-EM) [98]. Classifying protein structures by shape would be more relevant to functional classes of proteins than using conventional main-chain conformations since protein functions such as binding and catalysis occur at the surfaces of proteins. As shown in our previous study [99], functionally related proteins often share similar global surface but with low sequence and backbone conformation similarity. An illustrative example is DNA topoisomerase I from human and *E. coli.* Despite their low sequence identity and structure similarity, both of them share a characteristic pore to encircle DNA double strand. This function similarity can be easily captured by shape descriptors, but not captured by conventional main-chain conformation approach.

Protein surface shapes were represented with 3D Zernike Descriptors (3DZD), mathematical moment-based invariants of 3D functions [99]. 3DZD has been demonstrated efficient for various biomolecular structure comparisons [100], including comparisons of EM maps [101]. Another critical difference between the current study and the previous works is that we analyzed protein complexes in comparison with single proteins. The shape mapping of single-chain and complex protein structures with 3DZD yielded a unique landscape of protein structure space that was not explored before. Dominant features that characterize protein shape are the eccentricity, which is the degree of elongation of shapes, and the number of domains. Symmetry groups are another

feature that affects the shape in the case of protein complexes. A detailed analysis of the principal axis corresponding to the elongation of protein shape has suggested that proteins are required to form multimers if their shape is elongated over a certain degree. Overlapping the shape space occupied by single proteins and complexes identified shapes that are only possible in complexes. The unique view obtained from the current shape mapping leads to a more comprehensive understanding of building mechanisms, evolution, and design principles of proteins.

2.2 Methods

2.2.1 Single-chain dataset

The representative set of single-chain protein structures was selected from a PISCES culled list with a resolution cutoff of 2.2 Å, an R factor cutoff of 0.2, and a pairwise sequence similarity cutoff of 25% [102]. From 7,260 chains in the list, we removed short chains with less than 40 amino acids. We have also removed proteins that have a large spatial gap, i.e. structures having more than one cluster when C_α atoms were clustered with a 9 Å cutoff. We further removed 82 chains were further removed from the list because their sequences had more than 25% sequence similarity to other chains. This process yielded a dataset of 6,841 non-redundant protein structures.

From this dataset, we prepared another dataset by pruning structures that include less than 95% of residues relative to the whole chain length. The protein lengths were obtained from UniProt [103]. There are 2,366 chains in this high-coverage single chain dataset. For each chain, fold class was assigned following CATH. Also, by referring to PISA [104], we assigned biological unit information.

2.2.2 Complex dataset

From PDB, we identified structures that exist as a complex as defined in PISA and downloaded the first biological unit (BU). The same resolution, R factor, and length cutoffs as in the single chain dataset were applied. A complex is considered as redundant if there is another complex with the same number of chains and corresponding chains between them have over 25% sequence similarity. Among redundant complex entries, we chose the one with the highest resolution and the lowest R factor. This procedure yielded 5,326 complexes. Symmetry information for complexes was obtained from PDB if the BU of the complex considered has the

same composition as in PDB. Out of the 5,326 complexes, 2,876 of them acquired symmetry information.

2.2.3 Protein surface shape representation

We used 3DZD, mathematical rotation-invariant moment-based descriptors, to represent the surface shape of single-chain proteins and complexes. For a protein structure, a surface was constructed using the MSMS program [105] and then mapped to a 3D cubic grid of the size of N^3 (N was set to 200). Protein size is not explicitly considered in 3DZD calculation. But in our previous study [23], we have shown that it is rare for proteins with very different sizes to share similar global surface. MSMS failed to generate surface for two cases each in the single-chain dataset and the complex structure dataset, for which we used the MSROLL program [106] instead. Each voxel (a cube defined by the grid) is assigned either 1 or 0; 1 for a surface voxel that locates closer than 1.7 grid interval to any triangle defining the protein surface, and 0 otherwise. This 3D grid with 1s and 0s was considered as a 3D function $f(\mathbf{x})$, for which a series is computed in terms of the Zernike-Canterakis basis [107] that is defined by the collection of functions

$$Z_{nl}^{m}(r,\theta,\Phi) = R_{nl}(r)Y_{l}^{m}(\theta,\Phi) \tag{2.1}$$

with $-l \le m \le l$, $0 \le l \le n$, and $(n-l)$ is even. $Y_{l}^{m}(\theta,\Phi)$ are spherical harmonics. R_{nl} are radial functions defined by Canterakis, constructed so that $Z_{nl}^{m}(r,\theta,\Phi)$ are homogeneous polynomials when written in terms of Cartesian coordinates. 3D Zernike moments of $f(x)$ are defined as the coefficients of the expansion in this orthonormal basis, *i.e.* by the formula

$$\Omega_{nl}^{m} = \frac{3}{4\pi}\int_{|x|\le 1} f(x)\,\bar{Z}_{nl}^{m}(x)dx \tag{2.2}$$

To achieve rotation invariance, the moments are collected into $(2l+1)$-dimensional vectors $\Omega_{nl} = (\Omega_{nl}^{l}, \Omega_{nl}^{l-1}, \Omega_{nl}^{l-2}, \Omega_{nl}^{l-3}, \ldots, \Omega_{nl}^{-l})$, and the rotationally invariant 3D Zernike descriptors F_{nl} are defined as norms of the vectors Ω_{nl}. Thus

$$F_{nl} = \sqrt{\sum_{m=-l}^{m=l} (\Omega_{nl}^{m})^2} \tag{2.3}$$

Index n is called the order of the descriptor. The rotational invariance of 3D Zernike descriptors means that calculating F_{nl} for a protein and its rotated version would yield the same

result. We used 20 as the order because it gave reasonable results in our previous works on protein 3D shape comparison [5, 99, 108, 109]. A 3DZD with an order n of 20 represents a 3D structure as a vector of 121 invariants [99]. The similarity between two proteins X and Y was measured by the Euclidean distance d_E between their 3DZDs, $d_E = \sqrt{\sum_{i=1}^{121}(X_i - Y_i)^2}$, where X_i and Y_i represent the ith invariant for protein X and Y, respectively.

To illustrate the characteristics of 3DZDs, we compare it against two other structure similarity measures, the Procrustes distance [110] and TM-Score [111]. The Procrustes distance is a root-mean square deviation (RMSD) between corresponding points in two objects after an appropriate optimization of translation, rotation, and scaling. The smaller the Procrustes distance, the more similar the shape are. On the other hand, TM-Score is one of the common measures of the similarity of the main-chain conformations of proteins. TM-Score ranges from 0 to 1, with 1 for identical protein structures. Proteins within the same fold usually have a score above 0.5. The Euclidean distance of 3DZD is usually below 10 for proteins of the same shape [99, 108].

In Figure 2.1, the Euclidian distance of 3DZD and the Procrustes distance were compared in two datasets. Panel A compares pairs of 20 ellipsoids with increasing eccentricities, while panel B shows results on 1,278 single-chain protein pairs that have the same number of vertices in the surface representation. The two measures correlated well with a correlation coefficient of 0.9784 for the ellipsoid dataset (Figure 2.1A), because surface points were systematically distributed in the same fashion for all the ellipsoids and thus corresponding points are easily matched for aligning two ellipsoids. On the other hand, the two measures often have very different distances in protein shape cases (Figure 2.1B), which typically happened when point correspondences do not even allow appropriate scaling of the two structures. In Figure 2.1B, there are many protein pairs that have different surface shapes with a 3DZD Euclidean distance of over 10 but with a small Procrustes distance of around 0.2. Figure 2.1C and 2.1D show such protein pairs. As shown, proteins in these pairs have very different shapes, which indicates that 3DZD performs more reasonably for comparing protein shapes. Indeed, for protein shape comparison, The Procrustes distance has difficulty because corresponding surface points in two proteins need to be determined prior to the distance computation, which are not available in general for protein surface comparison. This is more difficult when two proteins have a different number of surface points to be compared. Apparently, 3DZD does not have such a problem because it does not align points to points.

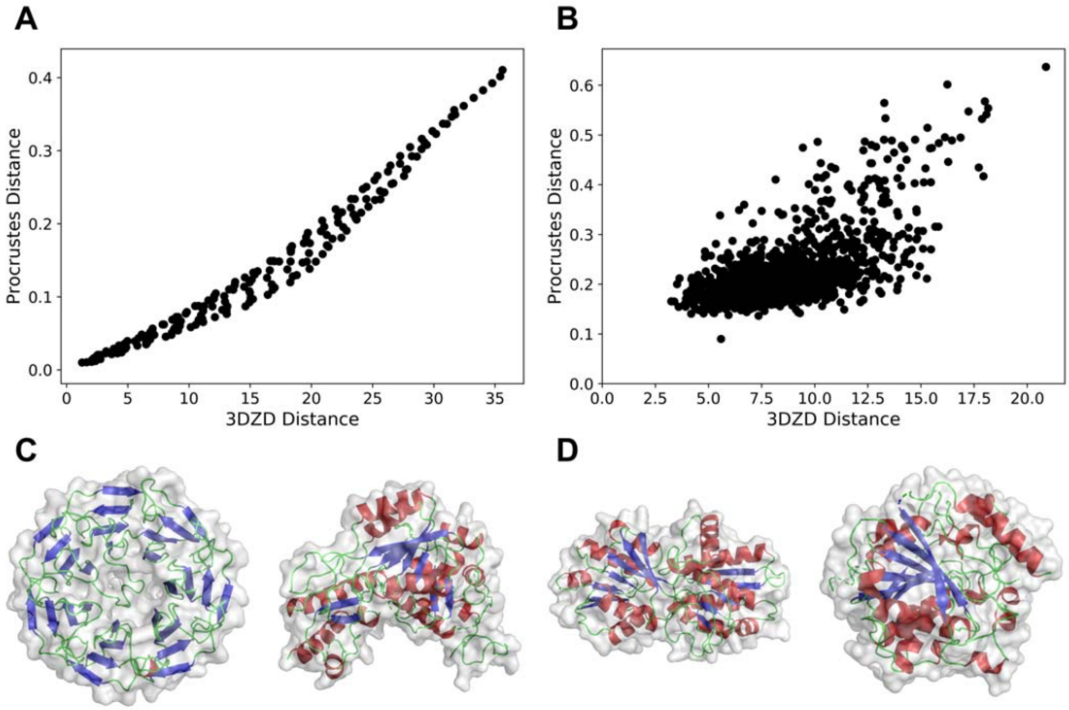


Figure 2.1: Comparison between 3DZD and the Procrustes distance. (A), Comparison of the Euclidian distance of 3DZD and the Procrustes distance for all the pairs of 20 ellipsoids with increasing eccentricity values from 0.0 to 0.92. On each ellipsoid 2500 points were sampled uniformly on the spherical coordinates. The two angles, θ (0 to π) and φ (0 to 2π) were evenly divided into 50 intervals and a point was placed on the ellipsoid surface for each combination of θ and φ. (B), The 3DZD and the Procrustes distances were compared for 1,278 single-chain protein pairs that have the same number of vertices in the surface triangle mesh representation. For computing the Procrustes distance for a protein pair, the closest surface point pairs from the two proteins were matched using the coherent point draft algorithm. (C), an example of protein pairs that have a large 3DZD distance and a small Procrustes distance. 2bwrA (CATH code: N/A) and 3ke3A (CATH: 3.40.640.10, 3.90.1150.10, there are two CATH codes because this is a two-domain structure). The 3DZD distance: 13.88; the Procrustes distance: 0.19. (D), another such example of protein pairs. 4gnrA (CATH: 3.40.50.2300, 3.40.50.2300) and 3ga7A (CATH: 3.40.50.1820). The 3DZD distance: 13.13; the Procrustes distance: 0.18.

Figure 2.2A and 2.2B show the comparison between 3DZD and TM-Score. As shown, these two measures have virtually no correlation. The correlation coefficient was -0.1735 for these two measures. Panel B shows the density of the two measures. The highest density (yellow) was observed at around 3DZD distance of 5 to 10 and TM-score of 0.3, which is the score range for proteins with similar surface shape but with different main-chain fold. There are cases that proteins of the different fold class have a small 3DZD Euclidian distance. Figure 2.2C and 2.2D shows two such examples, where two structures have a similar surface shape to each other according to 3DZD but have a very large difference in their main-chain conformations. These results are consistent with our earlier work where we extensively compared 3DZD with conventional protein structure comparison methods [99].

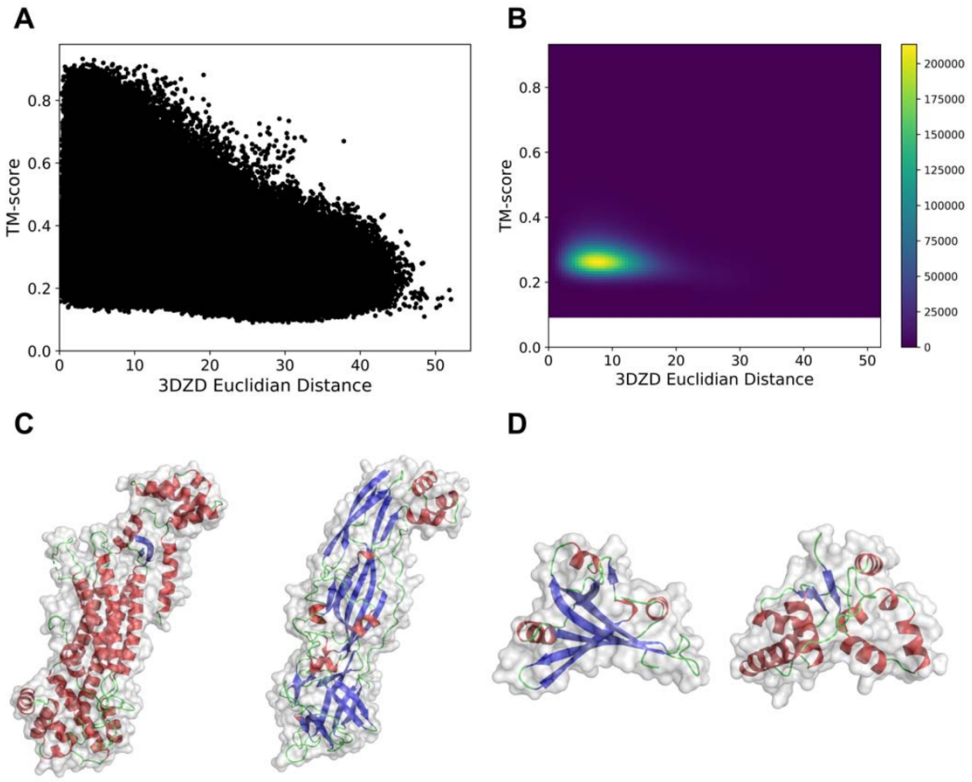


Figure 2.2: Comparison between 3DZD and the TM-Score on the single-chain dataset. (A), each point represents a protein pair. (B), the same data are represented with the density information. (C), an example of protein pairs that has a small 3DZD Euclidian distance but from different fold classes, the α class and the β class. Left, PDB ID: 1c3cA; CATH code: 1.10.276.10. Right, 4jp0A, 2.80.10.50. The Euclidian distance of 3DZD was 2.4, while the TM-score was 0.265. (D), another example of protein pairs with a small 3DZD Euclidian distance but from different fold classes, the β class and the αβ class. Left, 3a6rA; 2.30.110.10. Right, 3h87A, 3.40.50.1010. The Euclidian distance of 3DZD was 2.4, while the TM-score was 0.254.

2.2.4 Mapping structures

We used principal component analysis (PCA) to project 3DZDs of 121 value vectors of protein structures into 3D. Three eigenvectors were chosen for the mapping because the scree (Figure 2.3) showed that adding more eigenvalues does not contribute much to explaining data variance, and also to be consistent with the previous related works [75-77]. The three eigenvalues explained 52.64% and 47.76% of the total variation in the single-chain and the complex structure datasets, respectively.

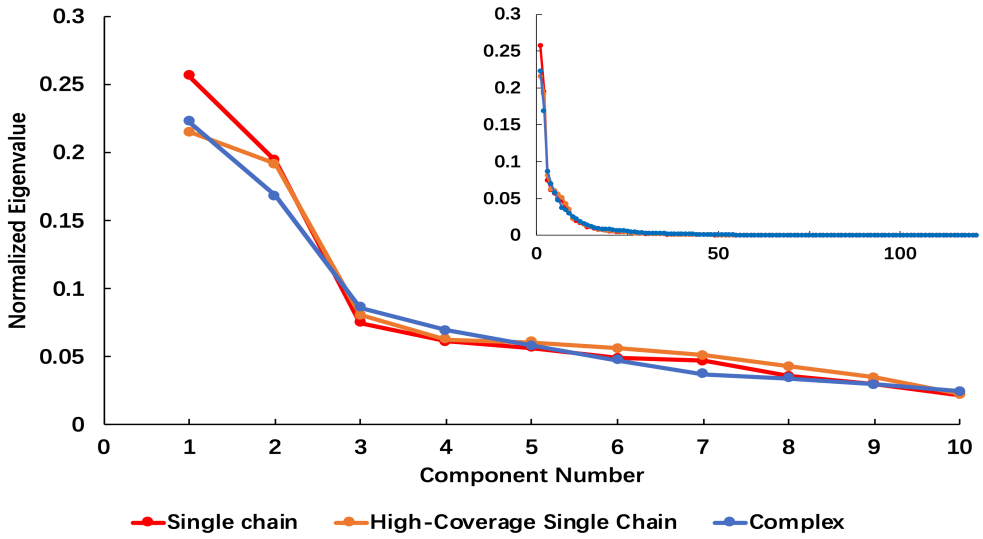


Figure 2.3: Scree plots of single-chain and complex datasets. The figure shows top 10 eigenvalues of the covariance matrix sorted in the descending order. The insert shows all 121 eigenvalues. Eigenvalues of single-chain, high-coverage single-chain and complex datasets are colored in red, orange and blue, respectively. The sharp drop up to the third eigenvalue indicates that adding fourth and more eigenvalues do not add substantially more information.

2.2.5 Eccentricity of a protein shape

In order to quantify how elongated a structure is, we have defined the term eccentricity, which is calculated from the minimum volume enclosing ellipsoid (MVEE) of a structure. Given all atoms in a structure, protein MVEE is the ellipsoid with minimum volume that encloses all atoms. From MVEE, the eccentricity is defined as $\sqrt{(2 - b^2/a^2 - c^2/a^2)/2}$, where a, b, and c are the length of longest, the second longest, and the third longest semi-principal axes of the

ellipsoid, respectively. Elongated structures have an eccentricity close to 1, while spherical structures have an eccentricity close to 0.

2.2.6 Protein volume computation

The volume of proteins was computed using MSROLL with a probe radius set to 0. For 42 cases in the single-chain dataset and 82 cases in the complex dataset where the MSROLL failed, we used the ProteinVolume program [112] instead. The volume values computed by these two programs were very consistent; the difference of volume values for ten randomly selected protein structures was on average 1.04%. The convex hull of a protein structure and its volume was computed using the ConvexHull function in the scipy.spatial package [113].

A pocket on a protein surface was identified and its volume was computed with VisGrid [114]. The average size of the pocket volume in the single-chain proteins was 6,302.9 $Å^3$. We analyzed the location of proteins with a large pocket whose size is within the top 10% (12,219 $Å^3$ or larger) in the single-chain protein surface space.

2.2.7 The genus number

Donut-shaped structures were identified by first screening structures with genus > 0 and then with the conditions of $0.9 \leq b/a \leq 1.0$ and$0 \leq \sqrt{(c^2/a^2 + c^2/b^2)/2} \leq 0.6$, where a, b, and c are the parameters of MVEE of the structures. Then, structures that passed the criteria were visually examined. The genus number was computed with the Euler-Poincaré Formula, which states the following relationship between the number of vertices (V), edges (E), faces (F), loops (L), shells (S), and genus (g) of a manifold: V + F – E – (L – F) = 2 (S – g). To obtain these values of a protein surface, we used triangular meshes computed by EDTSurf [115]. L is equal to F for triangle meshes since triangular faces have exactly 1 loop. S was computed as the number of disconnected groups of faces.

2.3 Results

2.3.1 Shape space of single chains

Figure 2.4 overviews the 3D space mapping of 6,841 representative single-chain protein shapes. The surface shape of each protein was represented with the 3DZD, a rotation-invariant mathematical descriptor of 3D protein surface shape, and mapped to a 3D space using principal component analysis (PCA). 3DZD is based on a series expansion using 3D basis functions, which represents the target 3D shape by a weighted combination of the basis functions. The rotation-invariance is achieved by computing a norm of the coefficient values that are assigned to the basis functions. PCA locates similar protein shapes close to each other in the space. The color of points indicates the eccentricity of the shapes, which quantifies how much a shape deviates from a sphere, with a higher value (red) assigned for more elongated structures (the maximum value is 1) and 0 for a perfect sphere (blue).

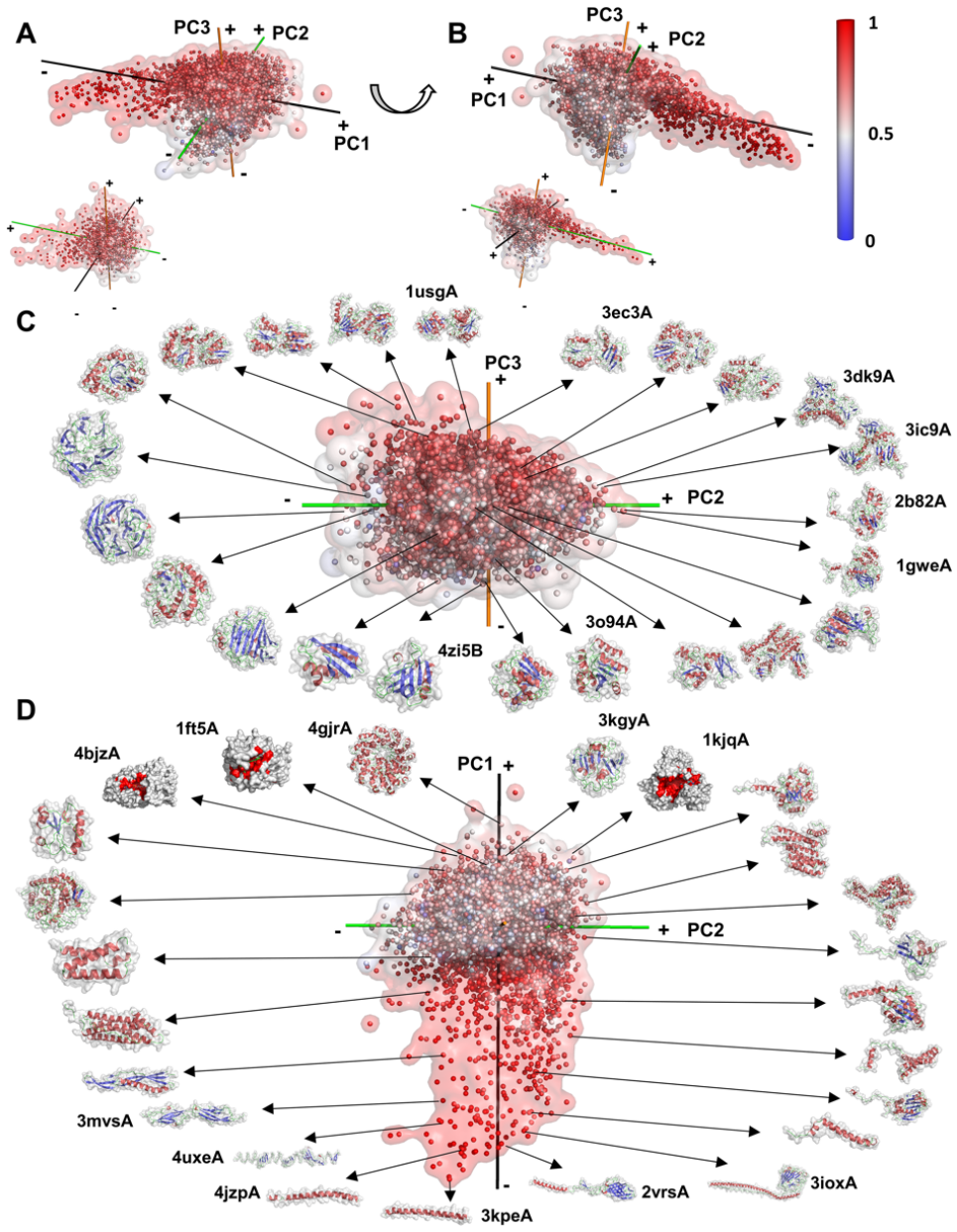


Figure 2.4: The 3D shape space of single-chain proteins. Each point corresponds to a protein. The distance between points represents the similarity of the corresponding protein shapes. The color indicates the eccentricity (the degree of elongation of a shape) from blue to red for 0.0 (sphere) to 1.0 (elongated shape). Shapes close to perfectly spherical (blue data points) do not exist in the single-chain dataset but exist in the complex structure dataset we discuss later. See Methods for the definition of the eccentricity. (A) and (B), the 3D shape space of single-chain proteins viewed from two different angles. The first, second, and third principal (PC1, PC2, and PC3) axes are shown in black, green, and orange, respectively. The positive and negative ends of an axis are labeled with + and -, respectively. The inset (a small figure of the shape space placed at bottom left) shows the distribution of high-coverage structure dataset, where a structure covers 95% or larger part of the entire protein. (C) and (D) show examples of protein shapes in the distribution on the PC2-PC3 plane (C) and on the PC1-PC2 plane (D).

Many entries of the Protein Data Bank [26] contain only a fraction of the whole structure; thus, we thought it may be possible that the distribution we see in Figure 2.4A may be biased toward surface shapes of structure fragments. For comparison, we also show in the inset figure of Figure 2.4A the distribution of 2,366 almost complete protein structures, which have at least 95% structure coverage of the whole proteins. The projection was made with PCA independently for this high-coverage dataset. As shown, the distribution of the high-coverage protein dataset is very similar, indicating that partial structures do not bias the distribution of the single-chain dataset.

2.3.2 Shape transition in the mapping space

The overall distribution (Figure 2.4A, 2.4B) shows that many proteins are on or close to the plane defined by the second and the third axes (the PC2-PC3 plane) with a characteristic thin layer of "tail" region, which expands on the PC1-PC2 plane along the first axis. Proteins located in the tail region and expanded towards the negative end of the first axis have elongated shapes (colored in red). Figure 2.4D confirms this observation on the tail region by showing representative structures in a 2D projection. Structures located at the negative end of the first axis are single α-helices (e.g. 4jzpA, 3kpeA), which are elongated and have high eccentricity values. Next to these long α-helical proteins are proteins of elongated shapes with more secondary structure elements, including β class (e.g., 3mvsA, 4uxeA) and αβ class structures (e.g. 3ioxA, 2vrsA). On the opposite (positive) end of the first axis more spherical shapes can be found (e.g. 4gjrA, 3kgyA) colored in white to blue (Figure 2.4D). The average eccentricity decreases along the first axis as shown in Figure 2.5, which has a correlation coefficient of -0.8704. Thus, the eccentricity is the primary factor for characterizing single-chain protein shapes.

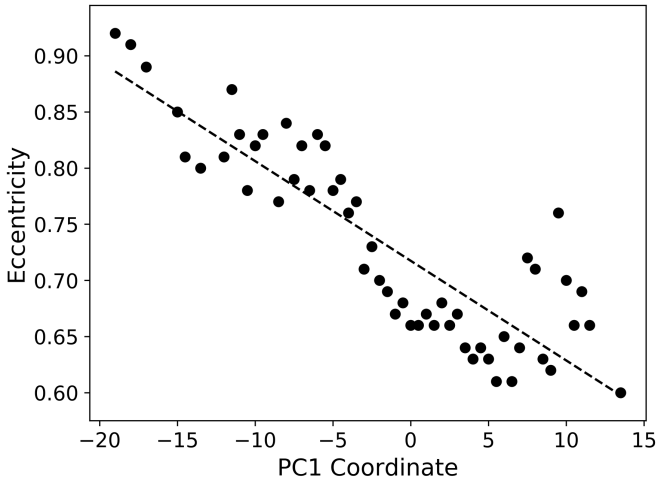


Figure 2.5: Structure transition along the first axis. The average eccentricity along the PC1 axis. Eccentricity of protein shapes are averaged at an interval of 0.5 along the axis, using shapes that locate in a sliding cylinder of a radius of 2.0 and a height of 0.5. The dashed line is the linear regression, (eccentricity) = -0.0089 * (PC1 coordinate) + 0.7174. The correlation coefficient between the eccentricity and the axis coordinate is -0.8704.

There are other noticeable trends in the mapping. Two-domain structures (e.g. 1usgA, 3ec3A) are dominant on the positive end of the third (orange) principal axis (Figure 2.4C) whereas the negative end contains more spherical single-chain shapes (e.g. 3o94A, 4zi5B). The positive end of the second (green) axis has shapes with multiple domains (e.g. 3dk9A, 3ic9A). Figure 2.6A and 2.6B plot the number of multi-domain proteins along the second axis and the third axis, respectively. The biases of observing multi-domain structures the positive side of the two axes were both statistically significant (p-value < 0.05 with χ^2 test). Further, Figure 2.7A and 2.7B visualize the number of domains (defined in the CATH database [83]) in the protein shape space. Associated with the trend of multi-domain proteins in the shape space, the positive ends of the second and third axes tend to contain long proteins (Figure 2.7C, 2.7D). We observed weak correlations between the average protein length and the coordinates of the second and the third axes, with correlation coefficient values of 0.3770 and 0.6185, respectively (Figure 2.6C, 2.6D).

It was also observed that the positive end of the first and the second axes accumulate proteins with relatively large and deep pockets. The bias of proteins with a top 10% largest pocket being on the positive side of the two axes was statistically significant (p-value < 0.05 with χ^2 test). Figure 2.4D includes three such examples, 3-hydroxybenzoate 6-hydroxylase (PDB ID: 4bjzA), which binds flavin-adenine dinucleotide (FAD), cytochrome c554 (1ft5A), which binds heme

(HEM), and glycinamide ribonucleotide transformylase (1kjqA), which binds adenosine 5'-diphosphate (ADP). The pockets of these three proteins are colored in red in the figure.

Overall, proteins with similar shapes are positioned close to each other in the mapping space, and transitions of the shapes are noticeable along each axis.

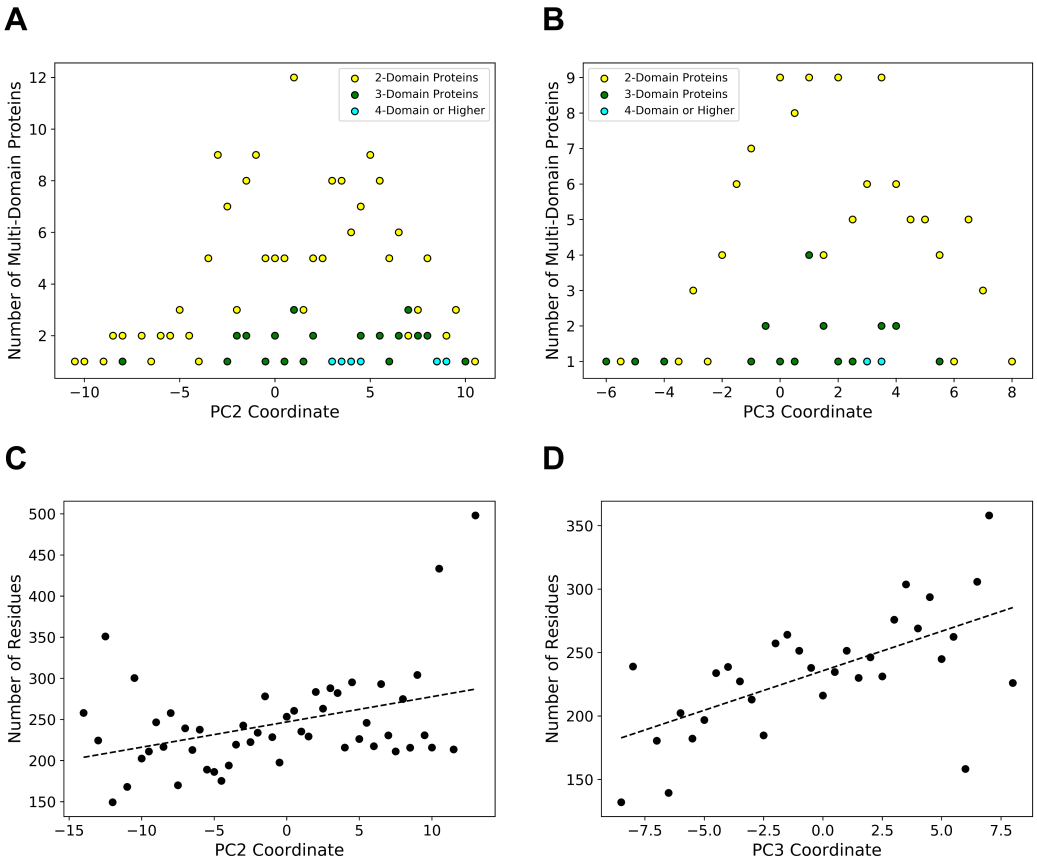


Figure 2.6: Structure transition along the second and the third axes. (A), the number of multi-domain proteins along the PC2 axis. The same sliding cylinder was used as in Figure 2.5. Proteins with two domains, three domains, and four or more domains are shown in yellow, green, and cyan, respectively. (B), the number of multi-domain proteins along the PC3 axis. (C), the average protein length along the PC2 axis. The same sliding cylinder was used as in Figure 2.5. The linear regression shown in the dashed line is (number of residues) = 3.073*(PC2 coordinate) + 246.93. The correlation coefficient is 0.3770. (D), the average protein length along the PC3 axis. The linear regression: (number of residues) = 6.226*(PC3 coordinate) + 235.61. The correlation coefficient is 0.6185.

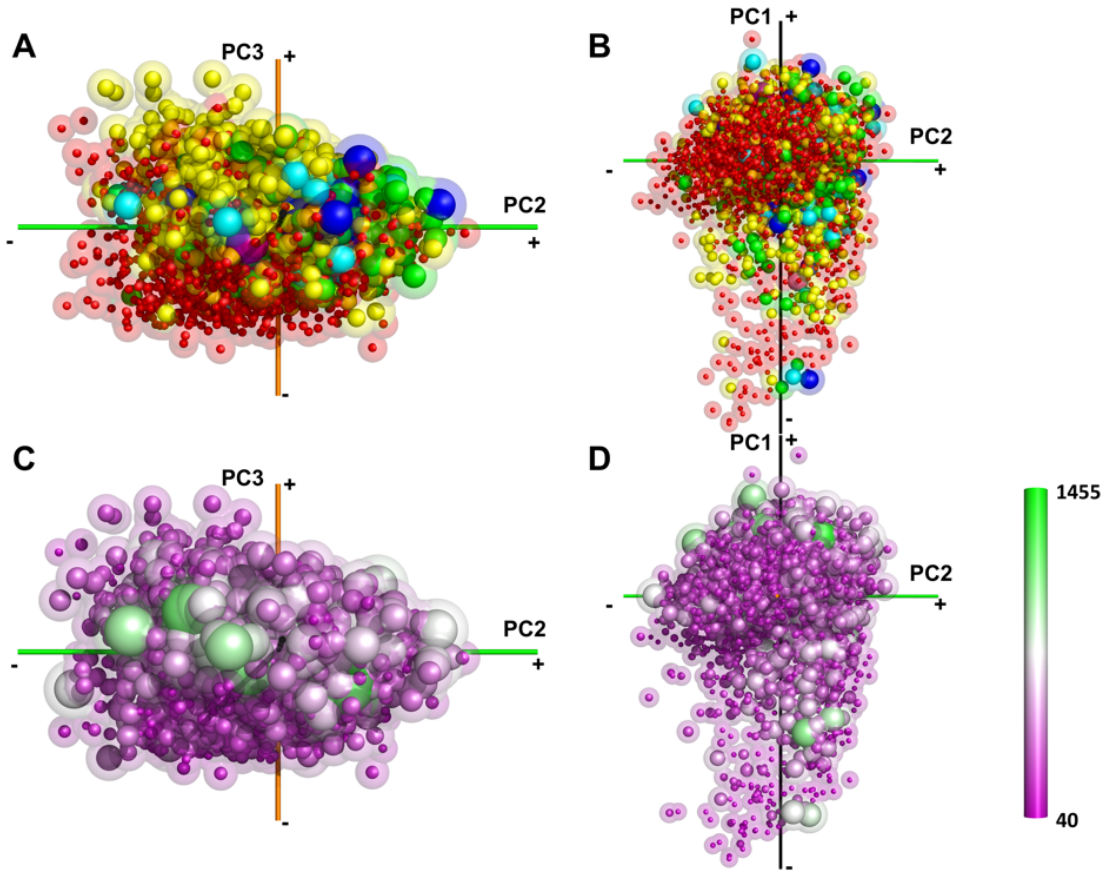


Figure 2.7: The distribution of the chain lengths and the number of domains in the single-chain shape space. (A) and (B), the number of domains in the proteins as defined by CATH. Red, yellow, green, cyan, blue, pink, and purple correspond to 1, 2, 3, 4, 5, 6, 8 domains, respectively. 6,109 proteins (89.3%) have CATH annotations. (C) and (D), the color code that ranges from purple to green shows the length (i.e. the number of amino acids) in proteins from short to long. The lengths were classified into twelve bins, 40-140, 140-240, and so on up to 1140-1540.

2.3.3 Monomer proteins and complex-forming proteins

A single-chain may either exist as a monomer or form a complex with other proteins in a cell. Is there any shape difference between these two classes of proteins? In Figure 2.8, proteins are colored in orange if they form a complex according to the biological unit information in the PISA database [104]. There are 2,259 (33.0%) monomers and 3,665 (53.6%) complex-forming proteins in the entire single-chain dataset (the remaining 13.4% do not have information in PISA). 754 (70.3%) out of 1,072 elongated-shape proteins (with an eccentricity of 0.8 or higher) are also indexed in PISA as forming complexes. On the other hand, for more spherical proteins (an eccentricity less than 0.5), the fraction of complex-forming proteins was 45.4%. The fractions of monomers and complex-formers in both elongated and spherical shapes are significantly different from the overall distribution in the entire single-chain dataset (p-value < 0.05 by χ^2 test). Therefore, the first principal axis, which showed a gradual shift from spherical to elongated shapes, also represents the transition from monomers to complex-forming proteins.

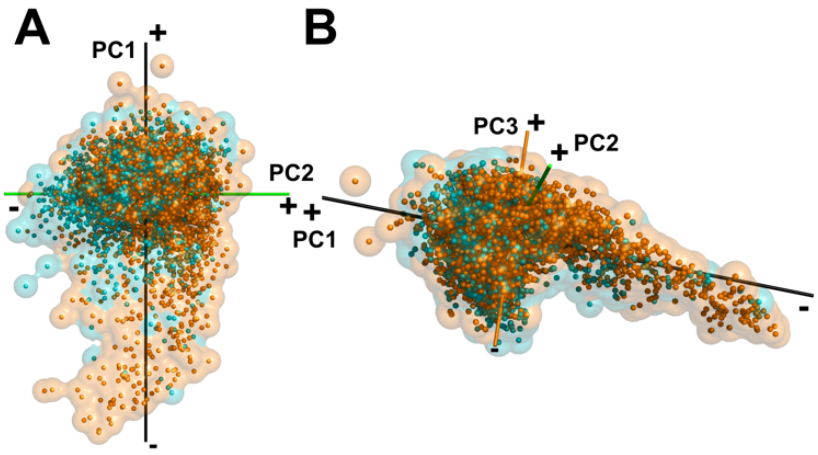


Figure 2.8: The distribution of monomers and complex forming proteins. (A), the distribution is shown on the 1-2 plane, same as the orientation in panel D in Figure 2.4. (B), the distribution is shown in the same orientation as panel B in Figure 2.4. Cyan, monomers; orange, complex-forming proteins.

The dataset includes 318 elongated proteins (with an eccentricity over 0.8) which PISA indicates monomers as their biological unit, not agreeing with the general trend. However, most of them (82.4%) turned out to be a part of a full structure, and if not, they interact with other proteins or nucleotides for their biological function. Examples include a ribosomal protein L22

(PDB ID: 1bxeA), which interacts with ribosomal RNAs and *Listeria monocytogenes* phage PSA endolysin (1xov) that binds to cell walls of host bacteria.

2.3.4 Protein main-chain folds in the surface mapping space

The protein fold spaces presented previously by the other groups [75-77] showed clear separation of structures of the α, β, and α/β classes in the projection space. In contrast, the protein shape space of the current work shows a very different view of the protein universe. Proteins with similar protein folds (i.e. main-chain conformations) are placed close to each other locally in the protein shape space as shown in Figure 2.4; however, in a larger picture there is no clear separation between different structure classes. Table 2.1 and Figure 2.9 show there are a substantial number of proteins from different classes that share similar global surface shape. Table 2.1 shows the number of protein pairs from various fold class combinations whose distances fall within the top closest pairs. Figure 2.9 shows the 3DZD distance of proteins from different fold classes (e.g. the α class and the β class) have very similar distribution as protein pairs from the same fold class (e.g. both from the α class). The results imply that there may be various completely different main-chain conformations building the same protein surface shape.

Table 2.1: Structure pairs from different CATH classes in single chain dataset

Fold class pairs	Top 0.1% *	Top 1% *	Top 5% *
α vs. α	713	5,416	23,529
β vs. β	1,035	8,849	37,581
α + β vs. α + β	6,714	63,882	311,314
SS vs. SS †	5	20	54
α vs. β	1,291	12,164	54,910
α vs. α + β	2,941	29,236	146,991
α vs. SS	39	215	905
β vs. α + β	4,329	41,595	195,848
β vs. SS	9	144	738
α + β vs. SS	30	322	1,843
Others ‡	2,608	33,121	191,561
Total	19,714	194,964	965,274

*Number of structure pairs belonging to each specific CATH class combination within certain percentage of all structure pairs in single chain dataset. Here structure pairs are sorted by their distance in projection space, with close ones at the top. †"SS" is short for "Few Secondary Structures". ‡Chain with multiple domains vs. Chain with one domain or multiple domains.

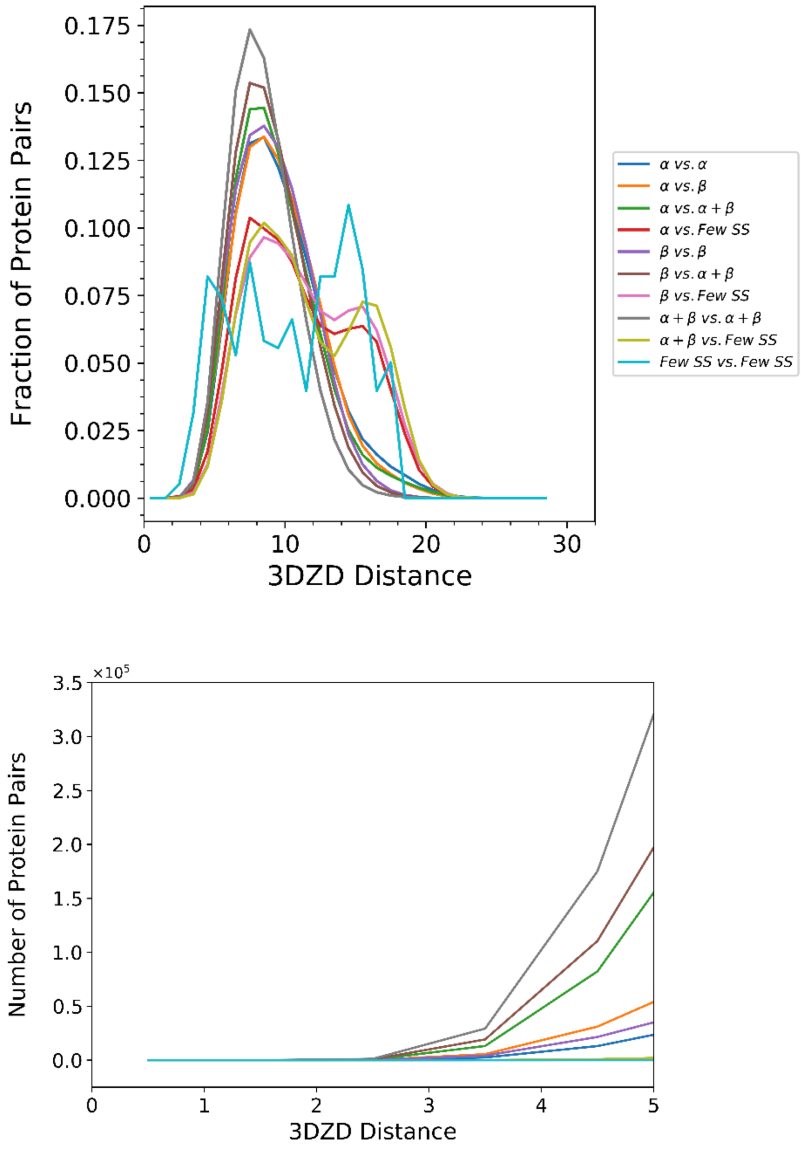


Figure 2.9: Distribution of 3DZD distances of protein pairs from different fold classes in the single-chain protein dataset. Top, the histogram of the 3DZD distances of proteins from different combinations of fold classes. Fold class information was obtained from the CATH database. The y-axis shows the fraction of pairs that falls into each distance bins. Two peaks are observed for pairs that involve the few secondary structure (ss) class. There are only 28 chains in the few ss class. Those chains have roughly two kinds of shapes, either elongated, or relatively spherical. The peak at a relatively small distance corresponds to pairs within each category, while the peak at a relatively large distance corresponds to pairs across two categories. Bottom, the 3DZD distance distribution of up to a bin of 4.0-5.0. The y-axis is now the actual number of protein pairs.

2.3.5 Shape space of protein complexes

Next, we discuss the shape space of protein complexes (Figure 2.10). The dataset of protein complexes contains 5,326 non-redundant structures. We obtained the biological units of complexes from PISA. As in Figure 2.4, the color indicates the eccentricity of shapes. The complex shape space is overall very similar to the single-chain shape space (Figure 2.4), with the majority of structures located around the globular region near the origin of the axes and a tail region dominated by elongated shapes (the region with many red points). On the other hand, some differences were observed between the complex and single-chain distributions. The protein complexes have more spherical shapes than the single-chain distribution (data points in dark blue in the mapping) (Figure 2.10A, 2.10B). The eccentricity histograms for the single-chain and complex datasets (Figure 2.11) verify this observation, which shows that the complex dataset contains highly spherical shapes with a low eccentricity. While there are no single-chain proteins with an eccentricity below 0.2, the complex dataset includes 72 such cases.

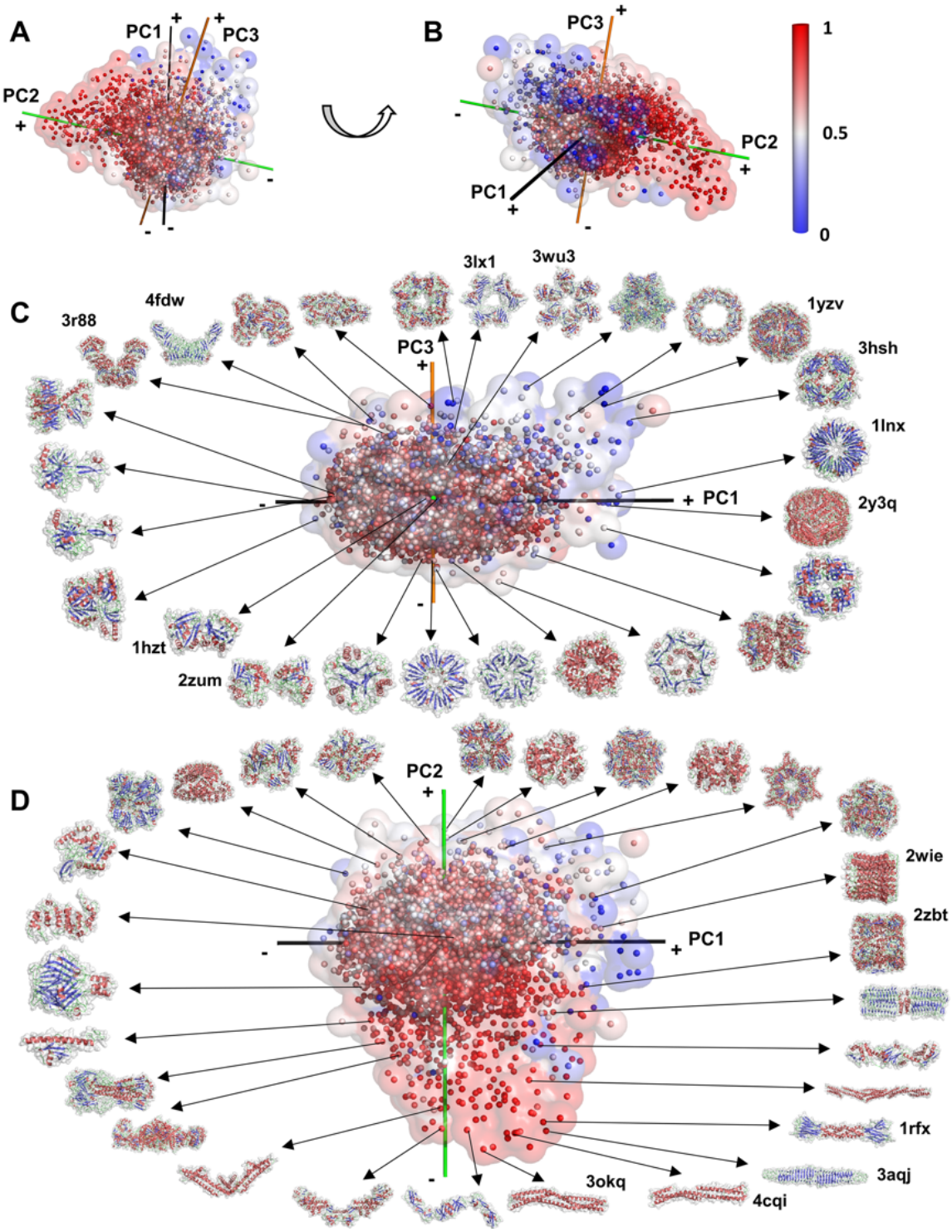


Figure 2.10: The overview of the complex shape space. 5,326 representative complex shapes are represented as points in the space. Points are colored by the eccentricity. (A) and (B), the shape space is viewed from two different angles. The color codes of axes and the eccentricity scale are the same as in Figure 2.4. (C) and (D) show examples of protein shapes in the distribution.

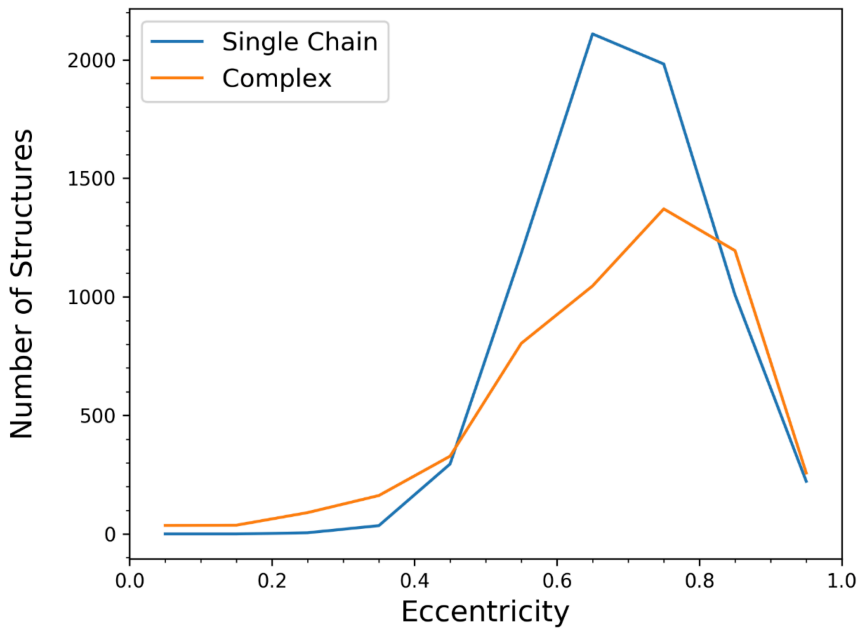


Figure 2.11: Histograms of eccentricity of the single-chain and complex datasets. The blue line is for the single-chain protein dataset, while the orange line is for complex dataset.

The differences between the shape spaces of the single-chain proteins and complexes become apparent when they are superimposed (Figure 2.12). To compare the size of the spaces occupied by the two datasets, the space was segmented into cubes of 1 axis unit edge length, and cubes were counted if they were occupied by the proteins in the datasets. Among all the cubes (3,895 cubes) that were occupied by at least one protein, 24.5% were filled by both single-chain and complex structures while 26.1% and 49.4% were occupied by only single-chain proteins and complex structures, respectively. Thus, the complex structure dataset occupies a larger space than the single-chain protein dataset. Figure 2.12C shows two example structures each from single-chain specific and complex-specific areas in the shape space. In the single-chain dataset, structures with a flexible tail (e.g. 3gzrA) were observed. Another example shown is 3e7kA, a narrow, elongated shape with a single helix, which is obviously very unique in single-chain proteins. On the other hand, highly spherical or symmetrical shapes are unique in protein complexes. 1yzv shown in Figure 2.12C has a spherical shape with the octahedral symmetry and 4ldm has a two-layer tube-like structure. The wide spread of complex structures suggests that assembling subunits into complexes can increase the range of attainable structures.

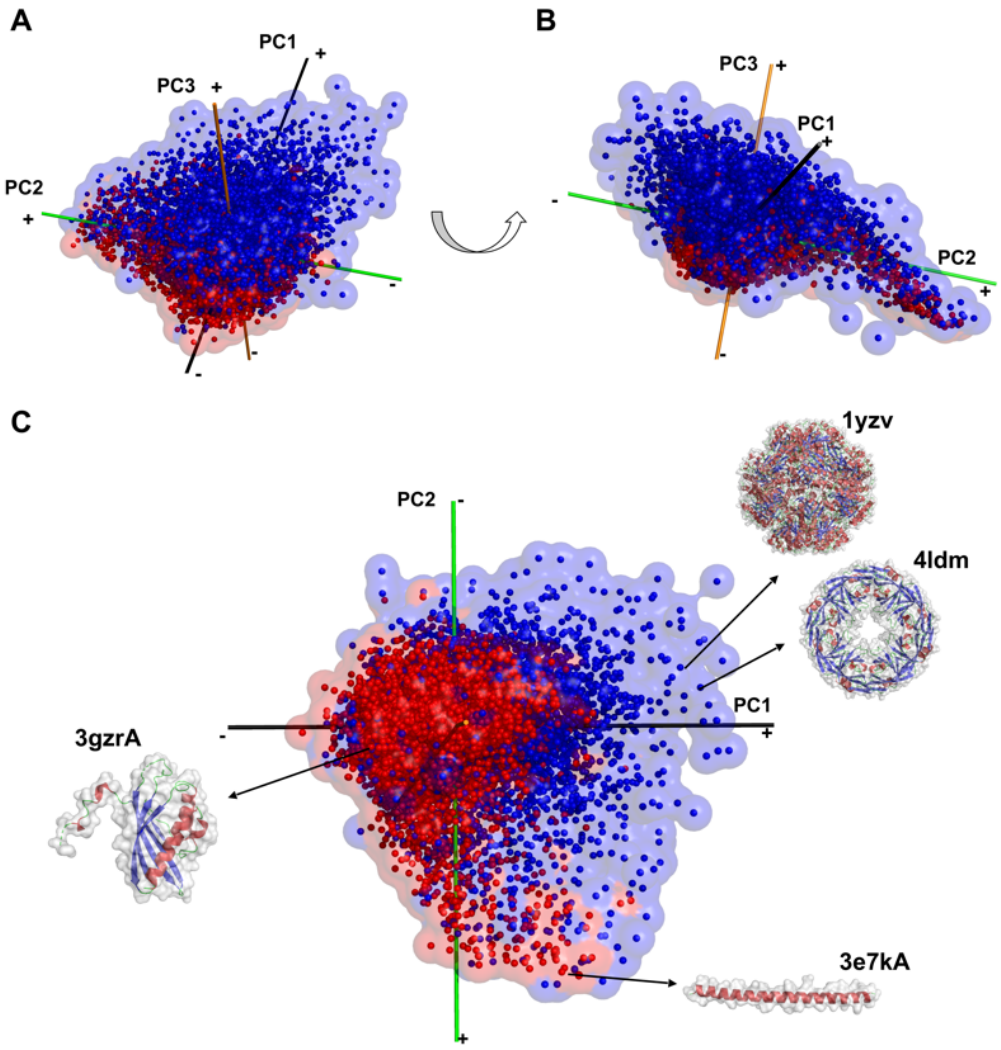


Figure 2.12: Superimposition of the single-chain and complex protein shape spaces. PCA was performed on the combination of the two datasets. Red, single-chains; blue, complex structures. (A) and (B) show the spaces in two different orientations. (C), examples of structures that locate in the single-chain specific (3e7kA and 3gzrA) and complex-structure specific (1yzv and 4ldm) areas in the protein shape space. 1yzv has octahedral symmetry and 4ldm has D4 symmetry.

Figure 2.10C and 2.10D annotate representative structures in the complex shape space. The outskirts of the distribution in the first quadrant (i.e. top right) in Figure 2.10C includes shapes of the almost perfect sphere (e.g. 1yzv, 2y3q), two layers of circular ring-like arrangements (e.g. 1lnx), and cube-like shapes (e.g. 3hsh). In the second quadrant (top left) several symmetrical "spiky" shapes with multiple protrusions are observed (e.g. 4fdw, 3r88). Close to the origin (0, 0, 0), dimeric complexes (e.g. 1hzt, 2zum) are observed. Figure 2.10D views the complex shape mapping from a different direction, showing the tail region occupied by structures with elongated shapes. They include protein structures of different fold classes, e.g. long α helices (e.g. 4cqi, 3okq), β structures (e.g. 3aqj), mixtures of them (e.g. 1rfx), and tube-like shapes (e.g. 2wie, 2zbt).

2.3.6 Shape symmetry

We further examined the symmetry of complex structures (Figure 2.13). Almost all complex structures have a specific symmetry type. Consistent with Figure 2.10C, we observed many complexes with dihedral symmetry (blue spheres, e.g. 1lnx) in the first quadrant. We also found that complexes in the first quadrant with higher-order symmetry, i.e. tetrahedral, octahedral (purple), and icosahedral symmetry (large orange sphere). These complexes are highly spherical and colored in blue in Figure 2.10.

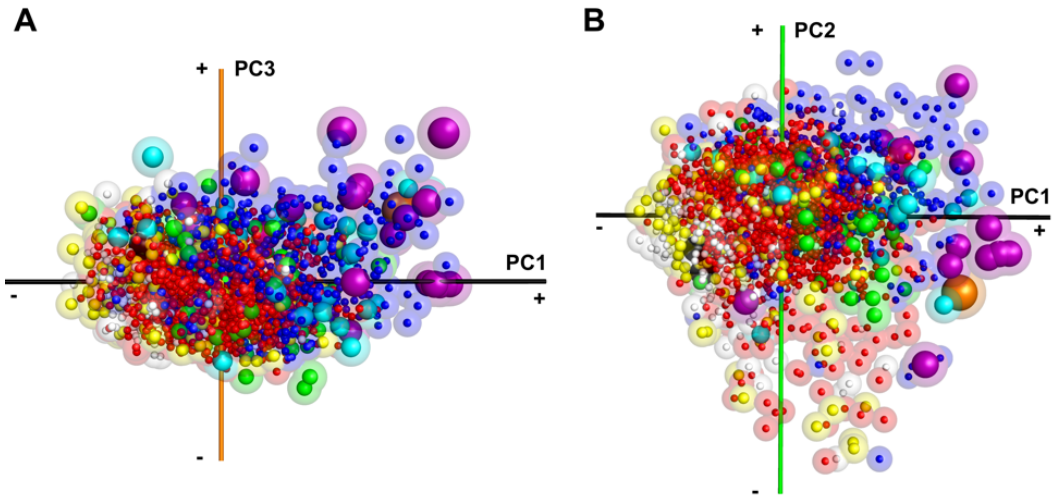


Figure 2.13: The structural symmetry of protein complexes. The protein complex shape space was colored by the structural symmetry. There were 24 symmetries in our complex dataset. Asymmetric structures, white; C2, red; C3, yellow; C4-C5, green; C6-C15, cyan. All dihedral symmetries (D2-D7) are colored in blue. Tetrahedral and octahedral, purple; icosahedral, orange; and helical, black. The radius of spheres reflects the symmetry number with a larger radius used for structures with a larger number. The distribution is shown in two orientations, A and B, which are the same as panel C and D in Figure 2.10.

2.3.7 Structures with holes

There are structures with large pockets or penetrating holes, which can be identified by comparing the volume of the structure itself and that of its convex hull (convex envelop that covers the volume). In the current analysis, buried cavities were not included because inner surface of buried cavities was not considered when the protein surface was generated. Also, tunnels in channel proteins were not effectively considered because such tunnels were too narrow to survive as cavities when the surface was constructed, and there are only three complete channel structures in the dataset in the first place.

Figure 2.14A shows the distribution of the ratio of the protein volume (Vp) to that of its convex hull (Vc). Complex structures tend to have a smaller Vp/Vc ratio, which is partly attributed to penetrating holes in structures. The relative abundance of structures with holes in complex structures can also be confirmed by computing a topological parameter, genus, using the Euler-Poincaré Formula. 93.1% of the single-chain structures have genus 0, which indicates that the structures do not have a hole, whereas the fraction decreases to 70.9% for complex structures. Figure 2.14B is an example of single-chain proteins that have large holes in the surface. The protein

is a subunit of a heteromeric complex, and the holes are formed by loop regions, which provide a binding space for other subunits. Figure 2.14C is an example of complex shapes. It is a homo-trimeric ring-shaped complex of proliferating cell nuclear antigen (PCNA), which encircles DNA at the hole in the middle of the structure and is involved in chromosomal DNA replication [116]. In the complex dataset, there were 71 other donut-shaped complexes, which have large penetrating holes in their centers.

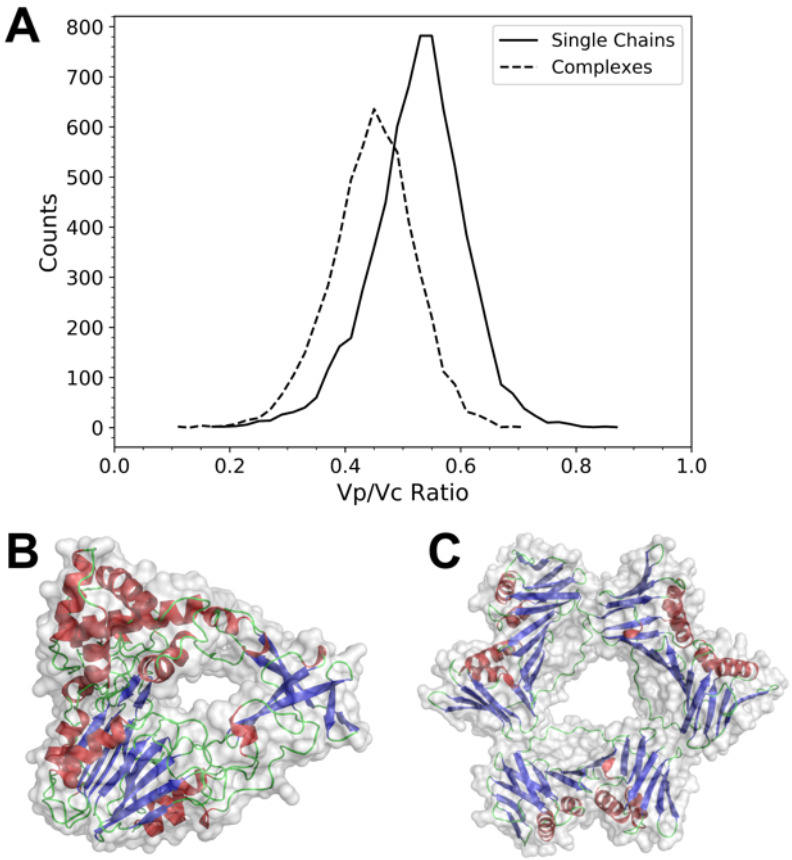


Figure 2.14: Protein structures with holes. (A) The distribution of the ratio of the protein volume (Vp) to the volume of its convex hull (Vc). Solid line, the single-chain dataset; and dashed line for the complex structure dataset. (B) glutaryl-7-aminocephalosporanic acid acylase b-chain (PDB ID: 4hst-B). The Vp/Vc Ratio is 0.408. (C) Active proliferating cell nuclear antigens (PCNAs), trimer (3lx1). The Vp/Vc ratio is 0.399.

2.3.8 Length dependency of structural features

In Figure 2.15, we examined how the eccentricity, the size of pockets, and the Vp/Vc ratio distribute relative to the number of amino acids for protein structures in the single-chain and the complex structure datasets. The first panel (Figure 2.15A) shows that very low eccentricity, i.e. highly spherical shapes, are achieved only by complex structures, which confirms the observation in earlier sections. Complex structures tend to have larger pockets as shown in Figure 2.15B. Naturally, larger protein complexes are capable of having larger pockets. Furthermore, a closer look at the plot around the protein length of up to 1,000 residues indicates that complex structures tend to have larger pockets than single-chains even when proteins of the same size are compared. Figure 2.15C examines the Vp/Vc ratio, the ratio of the protein volume relative to the convex hull of the protein. Overall, single-chain proteins and complex structures show similar distributions, but there are more complex structures observed in the lower end of the Vp/Vc ratio. Panels D, E, F illustrate the difference of shapes with a small Vp/Vc ratio between the two datasets. In the case of single-chains, a small Vp/Vc ratio occurs for flexible proteins such as 3ag3I (Figure 2.15D) while for complexes typical such shapes are symmetrical ones with protrusions (Figure 2.15E) and shapes with a large hollow inside as shown in Figure 2.15F.

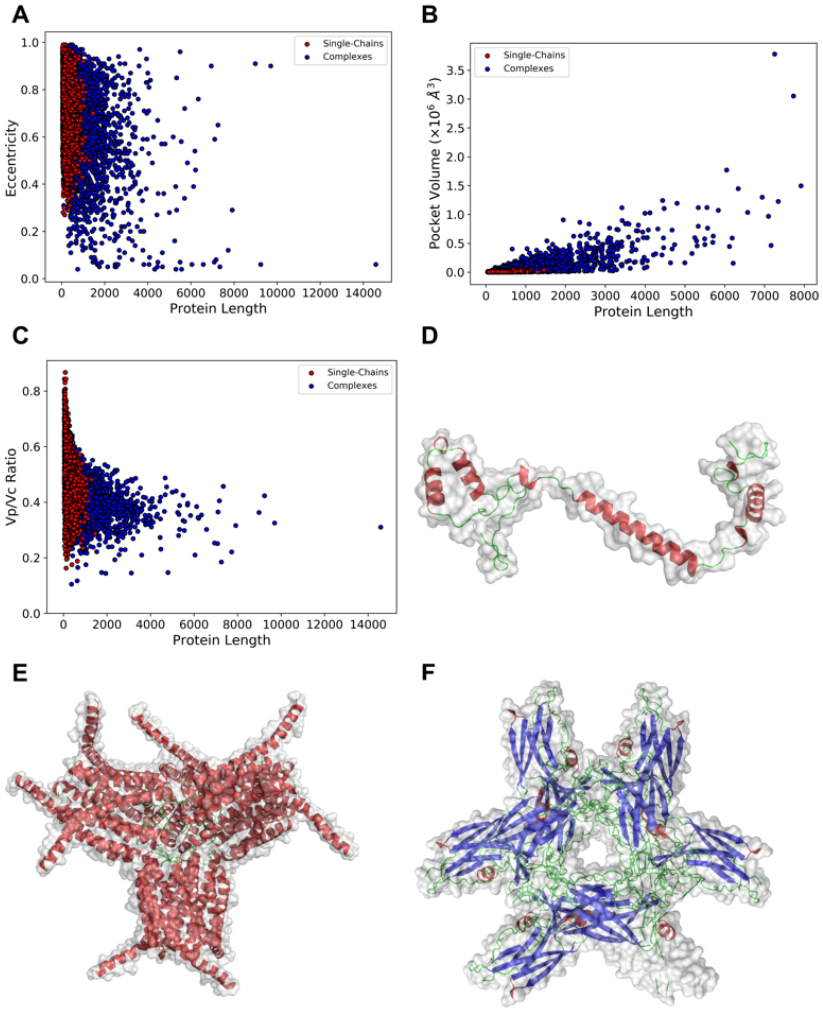


Figure 2.15: The eccentricity, the pocket size, and the Vp/Vc ratio relative to the protein length. (A), the eccentricity of proteins was plotted relative to the protein length. Red, single-chain proteins; blue, complex structures. (B), the pocket volume (Å^3) relative to the protein length. (C), The Vp/Vc ratio relative to the protein length. (D), an example of single-chain proteins that have a small Vp/Vc ratio. 3ag3I, a 72 residue-long protein, which has a Vp/Vc ratio of 0.29. (E), An example of complex structures with a small Vp/Vc ratio. 3pcv, a complex with 12 chains with a total of 1,752 residues. The Vp/Vc ratio is 0.147. (F), another example of complex structures with a small Vp/Vc ratio. 3p69, a complex with 12 chains with a total of 1,524 residue long. The Vp/Vc ratio is 0.295.

2.4 Discussion

In this study, we have constructed a mapping of the protein structure space for the first time by considering the overall surface shape of both single-chain and complex proteins. The shape space visualized in this work would give an impression that the protein shape space is continuous, but this is not specific to the protein surface shape representation. Indeed, earlier works that mapped protein structures considering main-chain conformations also show continuous structure distributions [75-77, 96]; and moreover, there exists active discussion on the continuity [117] or the many-to-many similarity relationship [118] of the protein structure space. Analogous to well-established protein main-chain structure classifications, such as SCOP [84] and CATH [83], this work will lead to a new classification for protein shapes at a medium to low resolution, which are being accumulated at an increasing pace by cryo-electron tomography and cryo-EM. By establishing the classification from the distribution of the protein shapes, for example, we will be able to take a census of protein shapes, that is, to count the number of specific protein shapes in organisms and compare across different organisms [119].

The observed variety of protein shapes in this work will also be useful for designing protein representations used in a cell-scale physical simulation of biomolecules [120]. Rather than using an overly simplified molecular representation, as is usual for such a simulation, one could diversify protein shapes in the simulation box by sampling structures from different locations in the shape space (Figure 2.4 and Figure 2.10).

Last but not least, this work has strong implications for protein design. Our study indicates that a protein shape can be realized with utterly different backbone conformations that even belong to different fold classes as shown in Table 2.1. Also, the shape mappings of single chains and complexes revealed regions in the shape space that are not occupied by either of them, or are occupied only by complex shapes (Figure 2.12). Shapes that correspond to the former may be difficult to construct with proteins, and other materials such as DNAs or polysaccharides may be required, while those in the latter region may be better designed using complexes rather than a single-chain protein.

In the coming age of medium- to low-resolution biomolecular structures, protein design needs a novel way of viewing biomolecular shapes. We expect that this work makes a unique and significant contribution by providing a foundation of understanding the protein shape universe.

CHAPTER 3. LOCAL DENSITY VECTOR BASED ALGORITHM FOR EM MAP ALIGNMENT

In Chapter 2, I described the application of 3D Zernike descriptors in capturing protein global surface shape information. Similarity between two protein atomic structures can be measured by the Euclidean distance of their 3DZD vectors. Besides comparison of protein atomic structures, 3DZD can also be applied to global matching of electron microscopy (EM) density maps. But as 3DZD focuses on global shape of a protein structure, it cannot perform partial matching of a protein subunit to the complete protein complex. In this chapter, I will present a method to perform both global and partial matching of EM maps.

3.1 Background

In light of recent technological advancements in cryo-electron microscopy (cryo-EM) [98, 121, 122], there are more and more macromolecular structures deposited in the Electron Microscopy Data Bank (EMDB) [62]. Currently, EMDB holds over 10,000 entries, and the number of entries is increasing rapidly. To take full advantage of these 3D macromolecular structures, it is necessary to have the tool to efficiently search against the entire database and superimpose maps at the correct position. Based on the completeness of the query structure, the EM map retrieval task can be further divided into two subtasks: global matching and partial matching. In global matching, each query map is considered as a complete structure, which is matched to all other maps in the database. For example, if ribosomal 50S subunit is used as the query, the goal is to retrieve all other maps for 50S subunit from the EMDB database. In partial matching, each query map is considered as an incomplete structure, which is used to retrieve maps for similar complete structures from the database. In this case, if 50S subunit is the query, the goal is to retrieve both 50S subunits and 70S ribosomes from the database, as 50S subunit is part of the complete 70S ribosome.

Previously, our lab has developed EM-SURFER, a web-based tool for real-time global matching and analysis of EM maps [7, 66]. In EM-SURFER, 3D Zernike Descriptors (3DZD) is utilized for the efficient comparison of EM map isosurfaces. 3DZD is based on a mathematical series expansion of a given 3D function [123], which gives a compact and rotation-invariant

representation of EM maps. Given a query map, EM-SURFER can successfully retrieve related entries of the same molecules at the top. However, EM-SURFER does not offer the partial matching functionality. It does not give the best superimposition between two maps either.

There are two existing methods for fitting EM maps, which are gmfit and fitmap. gmfit represents each map using a Gaussian mixture model (GMM), which is linear combinations of several Gaussian functions [6, 124]. To find the optimal superimposition of maps, gmfit randomly generates initial configurations of one map and performed a steepest-descent method to minimize the fitness energy. Similarities between two GMMs are evaluated using the correlation coefficient of the two distribution functions of GMM over all space. To use gmfit, user needs to specify the number of Gaussian distribution functions (GDFs) to approximate each EM map. This parameter also affects the fitting performance. High resolution maps require a larger number of GDFs to achieve a given value of correlation coefficient. Chimera also offers a function named fitmap for superimposing a pair of EM maps [125]. In fitmap, users can specify the number of initial placements N. Then N initial placements of one map within the other are generated randomly, then subjected to local optimization to maximize the correlation between two maps. As fitmap finds the local maxima rather than a global optimum, the best fit from fitmap heavily depends on the initial placements. If none of the initial placements is close to the correct superimposition, fitmap may not be able to identify a good fit after local refinement.

To solve those problems, here we have developed a new method named VESPER for aligning EM maps. VESPER is short for VEctor-based local SPace ElectRon density map alignment. To align two maps, VESPER firstly converts each map into a set of unit vectors using mean shift algorithm, where each voxel is represented by a unit vector pointing to the local dense region. Compared to cross-correlation based methods that use electron density values directly, this vector representation captures the local density neighborhood information for each voxel and is thus able to achieve better performance in both global matching and partial matching. After conversion of both maps into unit vectors, VESPER applies fast Fourier transform (FFT) [126] to search for good rotations and translations that maximize the summation of dot products of matched vectors (DOT score). Compared to EM-SURFER, VESPER has several advantages: (1) VESPER offers the partial matching functionality, which makes it possible to search for a complete structure given an incomplete query; (2) VESPER has better global matching performance compared to EM-SURFER; (3) VESPER gives the best superimposition of two maps and also gives a fitness

score to show how good the match is at each location; (4) a pool of superimpositions are produced, which gives the user several options to compare, check and choose from. The DOT score used in VESPER focuses on the matched region between two maps and adds no penalty for unmatched regions. Thus, compared to gmfit, VESPER can achieve better performance in partial matching. As FFT is used to search all rotation and translation combinations, VESPER identifies the best superimposition globally rather than the local maxima found by fitmap.

We have compared the retrieval performance between VESPER and 3DZD, gmfit and fitmap, in both global matching and partial matching. VESPER's performance is also compared to the cross-correlation based matching (CC), where the FFT search process maximizes correlation coefficient instead of the DOT score.

3.2 Methods

3.2.1 Unit vector conversion with mean shift

The first step of the algorithm is to convert each map into unit vectors, where each vector points to the direction of local dense region. Mean shift algorithm, a non-parametric clustering approach, is used for this task. First, the grid spacing of a map is converted to 7 Å. Then, a unit vector will be placed at each grid point x_i (i = 1, …, N) with a density value that is no less than the author-recommended contour level Φ_{thr} in an EM map. The unit vector located at x_i is $\frac{\overrightarrow{(y_i - x_i)}}{|y_i - x_i|}$, where y_i is computed as follows:

$$y_i = \frac{\sum_{n=1}^{N} k(x_i - x_n)\Phi(x_n)x_n}{\sum_{n'=1}^{N} k(x_i - x_{n'})\,\Phi(x_{n'})} \tag{3.1}$$

where $k(p)$ is a Gaussian kernel function and $\Phi(x_n)$ is the density value of the grid point x_n. To obtain a unit vector, the vector at each grid point is further normalized by the vector length. The Gaussian kernel $k(p)$ is defined as

$$k(p) = \exp(-1.5\frac{p^2}{\sigma^2}) \tag{3.2}$$

where σ is the bandwidth of the Gaussian kernel and set to 8.0 in our calculations.

3.2.2 Exploration of parameter combinations

We have explored several different combinations of voxel spacing and angle spacing to determine the optimal parameter combinations for VESPER. Smaller voxel spacing would give us more detailed density information in the map, but it would also increase the computation time significantly. Same case for angle spacing. Smaller angle spacing increases the precision in fitting but results in slower computations. To explore the retrieval performance of different parameter combinations, we have chosen three levels of voxel spacing and 4 levels of angle spacing. For each combination of voxel and angle spacing, we have searched 3 maps (EMD-3661, EMD-8724, and EMD-1203) against all other maps in the global matching dataset. Here we chose to use 3 maps with relatively small volume to reduce the total amount of computations. Table 3.1 shows the average computation time in CPU hours for each parameter combination. The use of smaller voxel spacing and angle spacing would increase the total computation time tremendously. For example, p = 3 Å and A = 10° would take more than 200 times of the computation time of p = 7 Å and A = 30°, the setting we would use in remaining sections of the paper.

Table 3.1: The average CPU hours for combinations of voxel and angle spacing settings

Angle Spacing	Voxel Spacing (Å)		
	3	7	10
10°	9568.3	319.5	91.4
30°	506.1	41.6	19.2
60°	267.5	30.7	16.5
90°	216.0	29.3	16.1

Besides the intention to control the total computation time, we also found that the retrieval performance of p = 7 Å and A = 30° is comparable to the ones from smaller voxel spacing and angle spacing. To evaluate the retrieval performance, we compared the average nearest neighbor (NN), first tier (FT) and second tier (ST) of different parameter combinations (Figure 3.1). We observed a linear drop in all three metrics as we increase the angle spacing. But the retrieval performance of p = 7 Å and A = 30° is only slightly worse compared to the best performance from other combinations. Therefore, we chose to use this parameter combination in following analyses to make the fitting both fast and accurate.

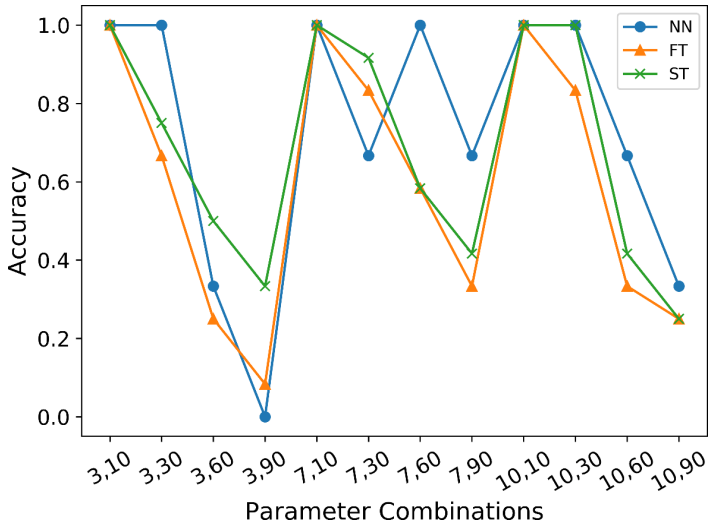


Figure 3.1: Global map retrieval performance using different voxel and angle spacing combinations. The average fraction of correct maps within the nearest neighbor (NN, i.e. top hit, blue), within the first tier (FT, orange), and the second tier (ST, green) for three query maps, EMD-3661, EMD-8724 and EMD-1203, were plotted. Along the x-axis, combinations of (voxel, angle) spacing values used are shown.

3.3 Results

3.3.1 Overview of the VESPER procedure

Figure 3.2a illustrates the overview of VESPER workflow. To identify the best superimposition of two EM maps, each map is firstly converted to a set of unit vectors using the mean shift algorithm. The voxel spacing of both maps is converted to 7 Å. This idea is inspired by the approach we took to identify local dense points (LDPs) in MAINMAST, a fully automated de novo structure modelling method developed by our lab [127, 128]. MAINMAST achieves good performance in building Cα models on data sets of both simulated and experimental maps. In MAINMAST, the locations of seed points are updated each iteration using the mean shift algorithm until convergence. Different from MAINMAST, the locations of seed points are not updated in VESPER. But instead, we represent each seed point as a unit vector pointing to the direction of the movement calculated from the mean shift algorithm. This vector representation gives us information about the local dense region around each voxel. Given a pair of EM maps, the goal is to find the transformation that maximizes the agreement of the local neighborhood around all voxels, which is captured by the directions of their unit vectors. After conversion of maps into unit vectors, the best superimposition of two maps is identified using the fast Fourier transform (FFT).

For each rotation sampled, a translation scan is performed using FFTs to optimize the summation of dot products of matched vectors (DOT score). The dot product of a pair of matched vectors ranges from -1 to 1: 1 for a perfect match, 0 if two vectors are perpendicular, and -1 if two vectors are in the opposite direction. With this unit vector representation, the vector at each voxel is given equal weight. Therefore, maximization of the DOT score is same as maximization of the overall agreement of all voxels in both maps. Lastly, for each of the top 10 models from FFT search, VESPER would perform 5° local refinement along each axis and then write top 10 models after the refinement into the output. In following sections, we will only focus on the best model from VESPER.

VESPER gives a fitness score for each vector in each of the top 10 models. The fitness score is the dot product of one vector to corresponding vector in the other map. By coloring the superimposition by the fitness score, users can tell the goodness-of-fit between two maps at different regions. We have shown the matching of V_o region to the complete V-ATPase as an example in Figure 3.2a. In the third panel on the left, vectors from the V_o region are colored by their fitness score, which range from -1 to 1. As shown in the third panel on the right, the helix from the V_o region at the left bottom corner does not match well to the helix from the complete V-ATPase. Vectors from that helix (blue) are thus in opposite direction to corresponding vectors from the complete structure (cyan), resulting in fitness scores worse than -0.5. For the helices and loops on the right side, as they match well in two structures, vectors from the V_o region (red) are almost in the same direction as the vectors from the complete structure (magenta) with fitness score better than 0.5.

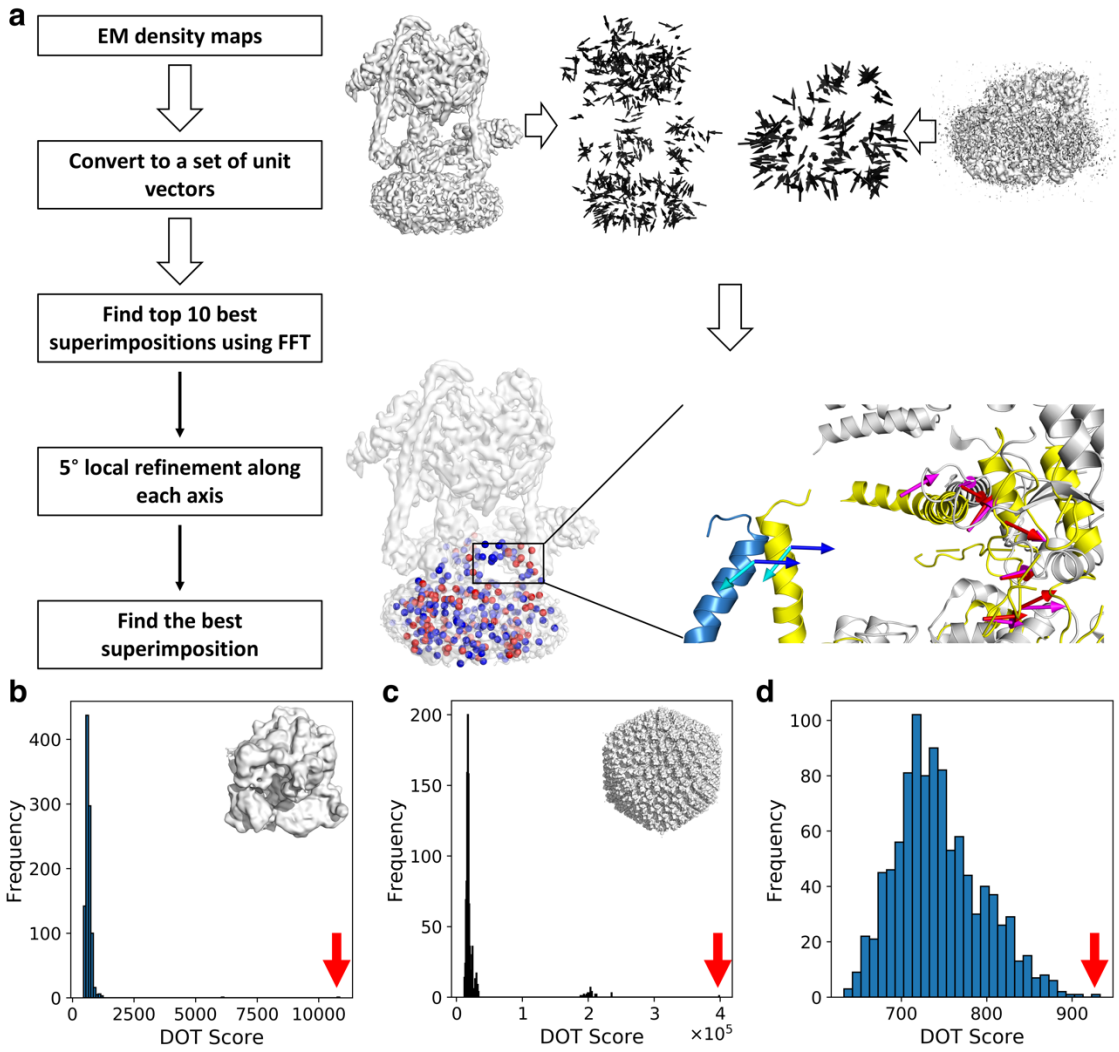


Figure 3.2: Overview of VESPER. a, Flowchart of VESPER. Steps of VESPER are illustrated in the right panel with an example of a map alignment between the V_o region of the V-ATPase (EMD-8409, 3.9 Å; right) and the complete V-ATPase (EMD-8724, 6.8 Å; left). First, a set of unit vectors are computed using the mean shift algorithm for each map. Next, the two maps are matched using FFT to maximize the sum of the dot products of matched vectors. Then, each of the top 10 models from the FFT search undergoes a local angle refinement with a 5° interval. The best scoring superimposition is shown at the bottom of the right panel. In the superimposed maps, vectors with non-positive and positive DOT scores are colored in blue and red, respectively. The bottom right in 1a shows the superimposition of the PDB structures from V_o region (PDB ID: 5tj5, colored in yellow) and the complete V-ATPase (PDB ID: 5vox, colored in gray). The helix in the complete V-ATPase that does not match well to the V_o region is colored in light blue. Vectors from the V_o region are colored in blue and red while those from the complete V-ATPase are colored in cyan and magenta. b, 70S ribosome (EMD-2978, resolution: 11.6 Å) matched to itself. DOT score: 10841; Z-score: 101.62. c, Human adenovirus 5 capsid (EMD-3004, resolution: 12.5 Å) matched to itself. DOT Score, 398169; Z-score: 94.31. d, Alignment between human adenovirus (EMD-3004) and 70S ribosome (EMD-2978). DOT Score, 943; Z-score: 3.97.

3.3.2 Benchmark procedure

For a database search for a query map, we use a normalized score (Z-score) instead of the raw DOT score because the DOT score has a dependency on the size of maps. We compute the Z-score as follows: The query map is placed in a rotational pose with an angle interval (the default value is 30°), and for each rotational pose, the query map is translated by the translational interval (7 Å is default). Then, the largest DOT score among all the translations for a rotational pose is stored.

Examples of the distribution of the largest DOT score from each rotational pose are shown in Figure 3.2b-d for self-comparison of 70S ribosome (EMD-2978), self-comparison of human adenovirus 5 capsid (EMD-3004), and a comparison between the two maps, respectively. From each of the top 10 scoring poses found, a further local rotational refinement with an interval of 5° is performed, from which the largest DOT score for two maps is identified. Then, the Z-score for the largest DOT score (indicated with a red arrow in Figure 3.2b-d), which is defined as (DOT_Score - mean)/standard_deviation, is computed from the DOT score distribution. Since a DOT score distribution can be biased when two symmetrical maps, such as virus capsids, are compared as shown in Figure 3.2c, the DOT scores in the distribution are first clustered with single-linkage clustering with a cutoff of 20% of the difference between the maximum and minimum DOT scores, and scores in the largest cluster is used for computing the Z-score. For the comparison of virus capsids in Figure 3.2c, this clustering process eliminates a bias to the score distribution for the Z-score computation, which is introduced from a small peak that locates at around 200,000. This peak comes from all rotational poses at the correct translation position of the two maps. The clustering process does not affect to usual cases of comparison with asymmetric maps (e.g. Figure 3.2b). The Z-score computed for the three map comparisons are 101.62, 94.31, and 3.97 for Figure 3.2b-d, respectively. Thus, self-comparisons (Figure 3.2b, 3.2c) had a very high Z-score apparently indicating that the compared maps are similar (actually identical as they are self-comparison) while maps with different shapes have insignificant Z-score (Figure 3.2d).

3.3.3 Dataset construction

To evaluate the performance of VESPER, we constructed a dataset of EM density maps from EMDB as follows: First, maps that do not have contour level information were excluded. Then,

remaining maps were grouped by the name of the macromolecules of the maps. Groups were inspected manually. A group was removed if it only contains low-resolution maps with a resolution of 20 Å or worse or if it contains less than five maps. From each group, five maps were randomly selected. This process yielded a dataset with 129 groups with 645 maps in total. Finally, groups that share the same partial structures are merged into a class. For example, the group for the complete V-ATPase and the group for the V_o domain of V-ATPase were merged in the same class. The resulting dataset with 105 classes was used for evaluating partial map matching performance. The number of maps in a class ranged from 5 to 50.

For global map matching, one group was randomly selected from each class to form a dataset of 82 classes with 410 maps. Thus, each class consists a single group with five maps. This is to have each class distinct from each other to prevent a query map to have a correct partial match with maps from different but related groups.

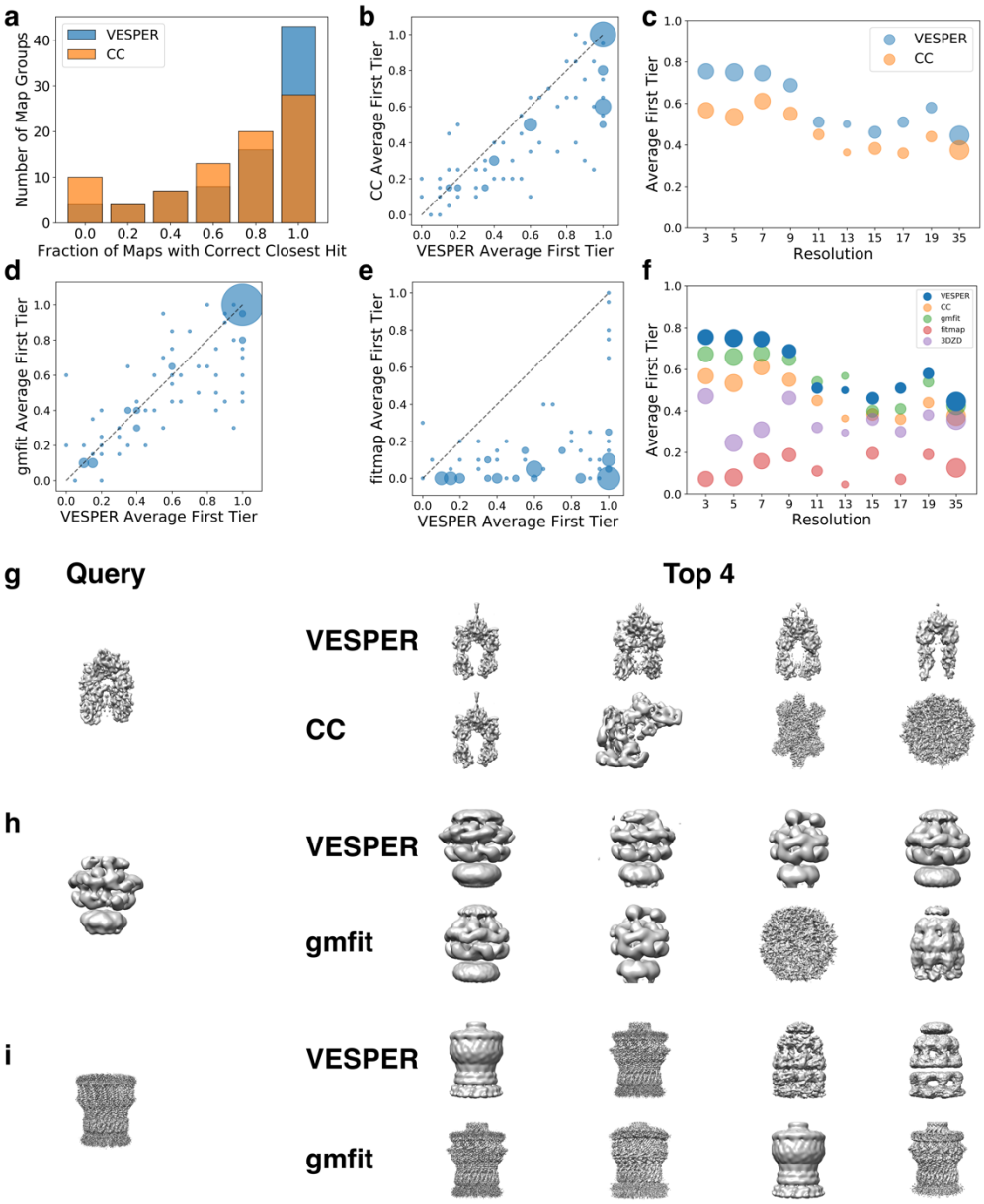
Using this dataset, we evaluated a method's ability to retrieve a map in the same class within the top, the first tier, and the second tier. The first tier is defined as the ranks up to the number of other maps in the same class with the query, and the second tier is double of it. Thus, for a global map search, the first and the second tier is within the 4th and the 8th ranks. For a partial map search, the ranks of the first and the second tiers depends on the number of maps in the same class.

3.3.4 Global map search

First, we examined the global map search performance of VESPER. We primarily used an angle interval of 30° and a translational interval of 7 Å unless noted otherwise because this setting showed a reasonable balance between the accuracy and the speed among other settings tested (See Methods). In Figure 3.3a-c, we compared results using the DOT score with cross-correlation (CC), which is a commonly used metric to evaluate the fitting of two EM maps [70, 125, 129]. For a query map in the global map matching dataset, the rest of the maps were compared with the query and ranked by the Z-score of the DOT score or CC. We examined if a map in the same group (a correct map) was retrieved as the closest, or within the first or the second tier. To evaluate map retrieval results for a group, the fraction of query maps in the group that found a correct map within a cutoff rank was computed. The overall performance of a method is computed by the average over all the groups.

Figure 3.3a shows the histogram of the fraction of maps in each group that found a correct map as the closest hit. As shown, VESPER with the DOT score (blue bars) had more groups (43 groups) that achieved 1.0 than using CC (orange bars; 28 groups). In Figure 3.3b, the retrieval performance of each map group is plotted considering the first tier. It is apparent that the performance of the DOT score was better than CC for the majority of the map groups. VESPER with the DOT score had a higher correct map fraction for 54 groups, while CC was better for 12 groups. Both methods tied for 16 groups. The same trend was observed when up to the second tier was considered (Figure 3.4a). The superior performance with the DOT score was observed consistently across all resolution bins from 2 to 50 Å (Figure 3.3c and Figure 3.4b). Figure 3.3g is an example where VESPER had a better retrieval performance than CC. For the query map of PKS module 5 (PikAIII) from the pikromycin pathway (EMD-5664 [130]), VESPER retrieved all other four maps in the same group, while CC retrieved only one map in the same group.

Figure 3.3: Performance on global map search. a, Number of map groups with different fractions of maps with a correct top hit. VESPER with the DOT score (blue) and CC (orange). b, Average fraction of correct hits within the first tier for each group. the x-axis, VESPER with the DOT score; the y-axis, CC. The area of a data point is proportional to the number of groups at that data point. c, Comparison of VESPER and CC on maps at different resolutions. The average fraction of correct hits within the first tier was considered. The resolution of the query map was considered. d, Comparison between VESPER and gmfit on the average fraction of correct hits within the first tier for each map group. e, Comparison between VESPER and fitmap on the first tier hit fraction. f, Average first tier hit fraction for maps in each resolution bin for VESPER (blue), CC (orange), gmfit (green), fitmap (red), and EM-SURFER (3DZD; purple). The resolution of the query map was considered. g, Example of a query map where VESPER performed better than CC in the global map retrieval. The query map is the PKS module 5 (PikAIII) from the pikromycin pathway (EMD-5664, resolution: 7.8 Å). The top 4 retrieved maps by VESPER were all from PikAIII: EMD-5649 (resolution: 7.8 Å), EMD-5663 (resolution: 7.9 Å), EMD-5651 (8.6 Å), and EMD-5666 (resolution: 11 Å), in this order. On the other hand, only 1 out of the top 4 retrieved maps by CC were PikAIII: EMD-5649 (PikAIII), EMD-6443 (Tetrahymena telomerase; 8.9 Å), EMD-6635 (bovine glutamate dehydrogenase; 3.3 Å), EMD-5145 (bovine TriC; 4.7 Å), in this order. The maps were visualized at the author-stated contour level in EMDB. h, Example of map retrieval where VESPER performed better than gmfit. The query is a map of ClpB bound to ClpP (EMD-2558; resolution: 21 Å). All the four maps retrieved in the first tier by VESPER were ClpB-ClpP complex: EMD-2557 (resolution: 17 Å), EMD-2556 (21 Å), EMD-2560 (25 Å), EMD-2559 (20 Å) in this order. With gmfit, only two within the top four retrieved maps were the ClpB-ClpP complex: EMD-2559 (ClpB-ClpP complex), EMD-2560 (ClpB-ClpP complex), EMD-5145 (bovine TriC; 4.7 Å), EMD-2327 (GroEL-GroES complex; 15.9 Å). i, Example of map retrieval where gmfit performed better than VESPER. The query is a 3.04 Å map of secretin GspD of the type II secretion system (EMD-6675). VESPER retrieved only two correct maps among the top four retrieved maps: EMD-1763 (secretin GspD, resolution: 19 Å), EMD-6676 (secretin GspD; 3.26 Å), EMD-2325 (GroEL-GroES complex; 8.9 Å), and EMD-1203 (GroEL-gp31 complex; 12 Å) in this order. All four retrieved maps by gmfit were all from secretin GspD: EMD-6676, EMD-8779 (4.2 Å), EMD-1763, and EMD-6677 (4.22 Å) in this order.


a
VESPER
CC
Number of Map Groups
Fraction of Maps with Correct Closest Hit
b
CC Average First Tier
VESPER Average First Tier
c
VESPER
CC
Average First Tier
Resolution
d
gmfit Average First Tier
VESPER Average First Tier
e
fitmap Average First Tier
VESPER Average First Tier
f
VESPER
CC
gmfit
fitmap
3DZD
Average First Tier
Resolution
g
Query
Top 4
VESPER
CC
h
VESPER
gmfit
i
VESPER
gmfit

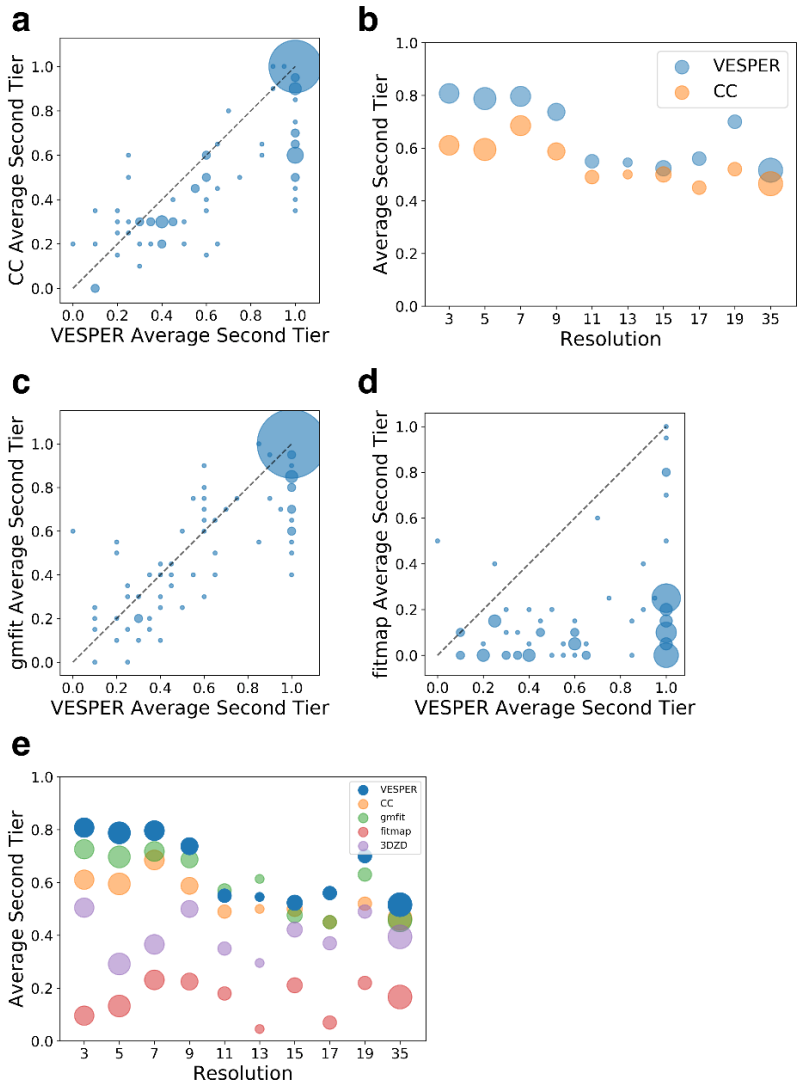


Figure 3.4: Performance of global map search in terms of correct hits within the second tier. Corresponding results considering the first tier are shown in Figure 3.3. a, Average fraction of correct hits within the second tier for each map group. the x-axis, VESPER with the DOT score; the y-axis, CC. The area of a point is proportional to the number of groups at that data point. b, Comparison of VESPER and CC on maps at different resolutions. The average fraction of correct hits within the first tier was considered. Blue, VESPER with the DOT score; orange, CC. c, Comparison between VESPER and gmfit in terms of the average fraction of correct maps within the second tier for each map group. d, Comparison between VESPER and fitmap in terms of the average fraction of correct maps within the second tier for each map group. e, Comparison of the average fraction of correct maps within the second tier at each resolution bin for VESPER (blue), CC (orange), gmfit (green), fitmap (red), and 3DZD (EM-SURFER) (purple).

We further compared the performance of VESPER using the DOT score and CC with three existing methods, gmfit and fitmap, and EM-SURFER that uses 3DZDs for the map shape search [7, 66]. gmfit was run with 20 Gaussian distribution functions, which is the setting in the Omokage map search web server [6], which uses gmfit for map superimposition. In fitmap, the number of initial placements was set to 100. Table 3.2 summarizes the map retrieval performance within the first and the second tier. For the global map search, VESPER had the best average correct map fractions within the first and the second tier, which were 5% points and 6.1% points higher than the second-best method, gmfit (the left half of Table 3.2). The direct comparison with gmfit (Figure 3.3d) and fitmap (Figure 3.3e) also shows that VESPER performed better for more map groups in the first tier. The same trend was shown when the second tier was considered (Figure 3.4c, 3.4d).

Table 3.2: Average fraction of correct maps retrieved within the first and the second tier

	Global		Partial	
	FT	ST	FT	ST
VESPER	**0.613**	**0.670**	**0.592**	**0.657**
CC	0.479	0.551	0.456	0.515
gmfit	0.563	0.609	0.479	0.551
fitmap	0.124	0.164	0.101	0.123
EM-SURFER	0.350	0.398	0.285	0.339

Global and partial map search results are shown. FT, the fraction of correct maps within the first tier (top |C|-1 maps, where |C| represents the number of maps in the same class as the query map); ST, the fraction in the second tier (top 2*(|C|-1) maps). The number in bold shows the best performance for each metric.

Figure 3.3f and Figure 3.4e show the map retrieval performance for maps at different resolutions. VESPER had the highest fraction of correct maps for most resolution bins. gmfit was the second for most of the resolution bins and the best for the resolution bin of 12 to 14 Å. Figure 3.3h is an example of map search from a query map of the ClpB-ClpP complex (EMD-2558 [131], resolution: 21 Å) where VESPER performed better than gmfit in map retrieval. While VESPER found all the other four maps of the same complex in the first tier, gmfit retrieved two unrelated maps that have somewhat similar overall shape at the third and fourth ranks, which happened perhaps due to the low resolution of the query map. Figure 3.3i is the opposite case, where VESPER's retrieval result was worse than gmfit. gmfit's retrievals were all correct in the first tier for the query map of the secretin GspD (EMD-6675 [132]) while VESPER's third and the fourth

retrievals were incorrect, both from GroEL. For this query, the GroEL maps had relatively high score because they have overall similar shape and also because these maps are largely hollow inside, and thus inconsistency inside the maps were not much penalized.

3.3.5 Global map alignment accuracy

Next, we examined the global map alignment accuracy. For this test, we randomly selected three pairs of maps each from the resolution range of better than 5 Å, 5-10 Å, and over 10 Å, thus in total of nine pairs, which have a fitted protein structure in PDB (Table 3.3). The ground truth of the superimposition for the map pairs were computed by aligning underlining protein structures of the maps using MM-align [8]. Table 3.3 summarizes the root mean square deviation (RMSD) of the best-scoring superimposition by VESPER with the DOT score in comparison with CC, gmfit, and fitmap. For VESPER, four parameter combinations of a shifting distance and a rotational angle were examined.

For seven out of nine map pairs, VESPER showed the lowest RMSD using one of the parameters used among the methods compared (Table 3.3). Comparing VESPER and CC, we see that VESPER had more cases with a smaller RMSD, indicating that the DOT score performs better than scoring with CC. Examining VESPER's results for maps with less (better) than 5 Å resolution, RMSDs achieved became lower as finer shifting distances were used. This implies that the DOT score was able to distinguish small differences in the alignments. The same trend still held but was less obvious for maps with worse resolutions. Overall, VEPSER showed best performance among the methods compared for both global map database search and alignment.

Table 3.3: Global map alignment by VESPER, CC, gmfit, and fitmap

Res. Range	Protein Name	Map 1 IDs	Map 2 IDs	RMSD (Å)									
				1 Å, 10°		3 Å, 10°		5 Å, 10°		7 Å, 30°		gmfit	fitmap
				VES	CC	VES	CC	VES	CC	VES	CC		
< 5 Å	γ-secretase	3240/5fn5	2677/5a63	**2.21**	2.38	3.61	3.92	3.25	3.25	8.87	8.87	2.63	2.90
	TRPML ch.	8881/5wpq	8764/5w3s	**1.12**	**1.12**	2.05	2.05	2.05	2.05	2.05	2.05	1.19	70.17
	$Ca_v1.1$ comp.	9515/5gjw	6475/3jbr	**2.31**	3.12	2.88	6.65	5.51	4.18	5.76	5.76	2.95	97.48
5 - 10 Å	Hsp104	8744/5vy8	8267/5kne	**0.86**	**0.86**	1.74	2.93	2.01	2.01	2.01	2.01	2.30	73.67
	V-ATPase	6284/3j9t	8724/5vox	2.79	2.67	3.67	2.89	3.67	3.67	3.67	3.67	5.05	**1.04**
	HCC comp.	3342/5fwm	3341/5fwl	**0.56**	**0.56**	**0.56**	**0.56**	**0.56**	**0.56**	**0.56**	**0.56**	3.60	4.98
> 10 Å	TRiC	1961/4a0v	1962/4a0w	**4.74**	5.39	4.94	6.11	4.94	4.94	4.94	4.94	8.17	7.61
	ClpB-ClpP	2557/4d2u	2559/4d2x	3.27	**2.85**	3.95	3.02	6.09	6.09	8.64	8.64	7.38	52.85
	70S rib.	9759/6j0a	9778/6j45	**2.71**	2.94	**2.71**	**2.71**	**2.71**	**2.71**	**2.71**	**2.71**	13.38	3.52

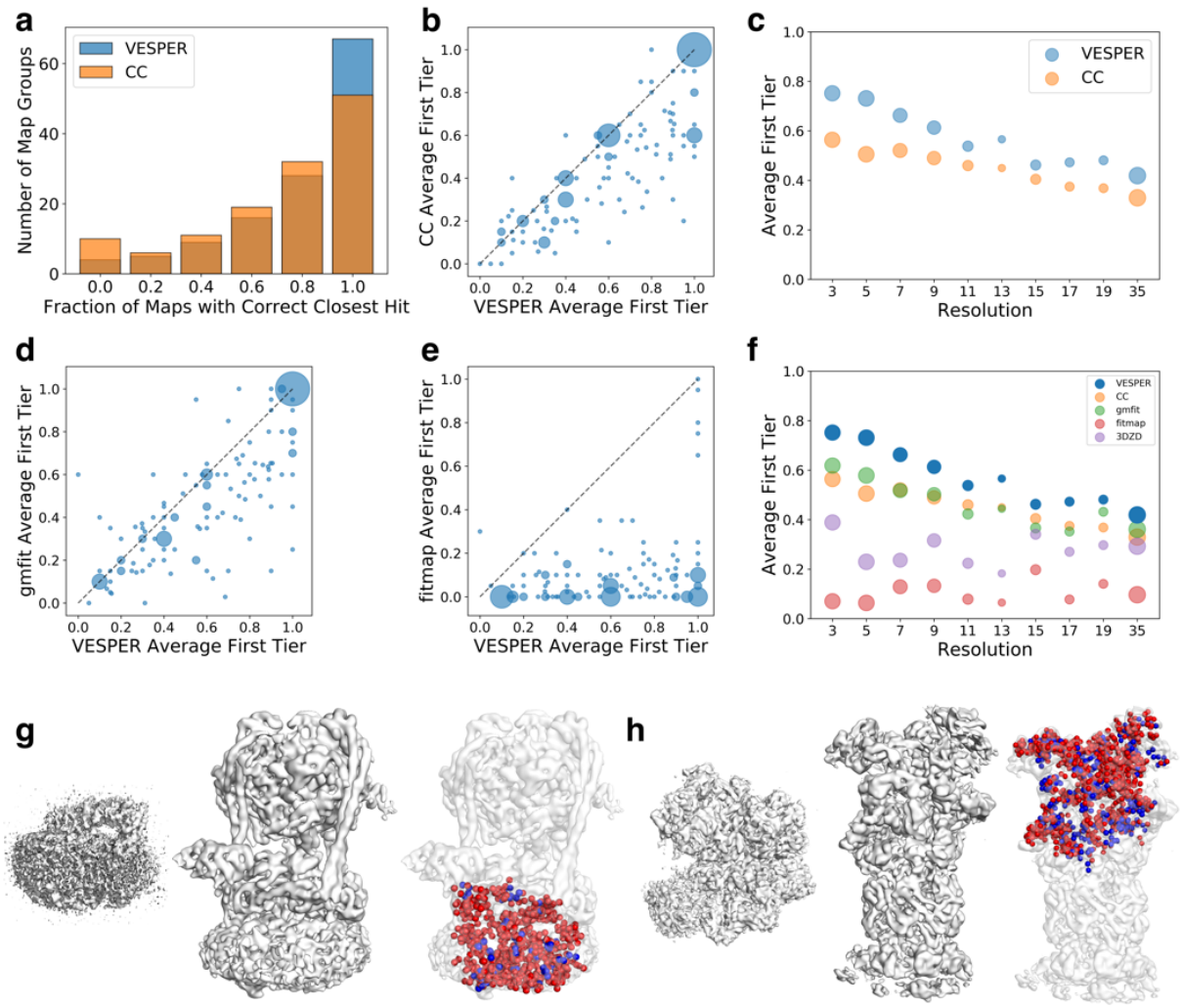
Res. Range, resolution range. In the protein name column, some names are abbreviated. They are TRPML channel (ch.), voltage-gated calcium channel $Ca_v1.1$ complex (comp.), Hsp104 disaggregase, Hsp90-Cdc37-Cdk4 (HCC) complex (comp.), chaperonin TRiC, ClpB-ClpP complex, 70S risobome (rib.). Map 1(2) IDs columns list EMDB and PDB IDs of the two maps and the associated PDB entries. The RMSD columns show the results from the top-scoring alignment by each method. For VESPER (VES) and Cross-correlation (CC), four different shifting and angle interval combinations were used. In this comparison, 3D-SURFER was not used as it does not provide map alignment. The smallest RMSD value for each map pair is highlighted in bold.

3.3.6 Partial map search

Next, we discuss VESPER's performance in partial map search. The partial map search is aimed at finding maps in a dataset, which contains common macromolecules with the query map. The results are summarized on the right half of Table 3.2 and Figure 3.5. Table 3.2 shows that VESPER achieved the highest average success rates when retrievals within the first and the second tier were considered. When the top hit was considered, VESPER had more map groups (67 groups) than CC (51 groups) that had a 100% successful retrieval (Figure 3.5). When individual map groups were considered, VESPER was more accurate than CC for the majority of map groups (Figure 3.5b), and the VESPER's advantage was consistent over all the resolution range (Figure 3.5c). When compared to gmfit (Figure 3.5d) and fitmap (Figure 3.5e), it was clear that VESPER performed better in the map retrieval for more map groups. Comparison for maps of different resolutions (Figure 3.5f) shows that VESPER was the best for all resolution bins. The same trend was observed when different criteria were used for evaluation (Figure 3.6).

Two examples of local matches are shown in Figure 3.5g, 3.5h. The first example is a search from a map of the V_o domain of V-ATPase (EMD-8409), which found a map of the complete V-ATPase (EMD-8726) at the third rank (Figure 3.5g). This retrieval at the high rank contributed to a substantially higher first tier (FT) success rate by VESPER in comparison with the other methods. VESPER's FT success rate was 0.57 while CC, gmfit, and fitmap (0.36, 0.36, and 0.21, respectively). As shown in the right panel of Figure 3.5g, the majorities of vectors from the two maps have positive dot product, which yielded the high retrieval rank. The RMSD of the local alignment by VESPER was 6.05 Å, which was sufficient to capture the map similarity. The match by CC and gmfit were not successful as shown as very large RMSD values of the alignments of 132.45 Å and 140.23 Å. In terms of RMSD, fitmap had a better RMSD for this map pair, 2.27 Å, but this match was ranked as low as 67 in the search. The second example (Figure 3.5h) is from a search from the proteasome regulatory particle (EMD-8675), which is aligned with 26S proteasome (EMD-3537). Although the alignment was not highly accurate (an RMSD of 11.32 Å), it was sufficient to rank the full 26S proteasome map as the top rank in the search. The other three methods had a completely wrong alignment with an RMSD over 100 Å and could not retrieve this map within a high rank (see the figure caption).

Figure 3.5: Performance on partial map search. a, Number of map groups with different fractions of maps with a correct top hit. VESPER with the DOT score (blue) and CC (orange). b, Average fraction of correct hits within the first tier for each of 129 groups. The x-axis, VESPER with the DOT score; the y-axis, CC. The area of a point is proportional to the number of groups at that data point. c, Comparison of VESPER and CC on partial map retrieval at different resolutions. The average fraction of correct hits within the first tier was considered. The resolution of the query map was considered. d, Comparison between VESPER and gmfit on the average fraction of correct hits in partial map search within the first tier for each map group. e, Comparison between VESPER and fitmap on the first tier hit fraction in partial map search. f, Average first tier hit fraction for maps in each resolution bin for VESPER (blue), CC (orange), gmfit (green), fitmap (red), and EM-SURFER (3DZD; purple). The resolution of the query map was considered on the x-axis. g, V_o domain of V-ATPase (left, EMD-8409, resolution; 3.9 Å) matched to the complete V-ATPase (middle, EMD-8726, resolution: 7.6 Å). Colored dots in the right panel shows the dot product of matched vectors, with red being a positive score and blue for a negative score. For this query, the first tier success rates of VESPER/CC/gmfit/fitmap were 0.57/0.36/0.36/0.21, respectively. The ranks of this hit (EMD-8726) from the query by VESPER/CC/gmfit/fitmap were 3/524/66/67 and the RMSD values of the match computed with the underlying protein subunit were 6.05/132.45/140.23/2.27 Å, respectively. h, Proteasome regulatory particle (left, EMD-8675, resolution: 6.1 Å) matched to 26S proteasome (middle, EMD-3537, resolution: 7.7 Å). The first tier success rates of VESPER/CC/gmfit/fitmap were 0.89/0.32/0.37/0.11, respectively. The ranks of this hit (EMD-8726) from the query by VESPER/CC/gmfit/fitmap were 1/507/184/473 and the RMSD values of the match computed with the underlying protein subunit were 11.32/111.40/138.33/131.44 Å, respectively.


a
VESPER
CC
Number of Map Groups
Fraction of Maps with Correct Closest Hit
b
CC Average First Tier
VESPER Average First Tier
c
VESPER
CC
Average First Tier
Resolution
d
gmfit Average First Tier
VESPER Average First Tier
e
fitmap Average First Tier
VESPER Average First Tier
f
VESPER
CC
gmfit
fitmap
3DZD
Average First Tier
Resolution
g
h

.

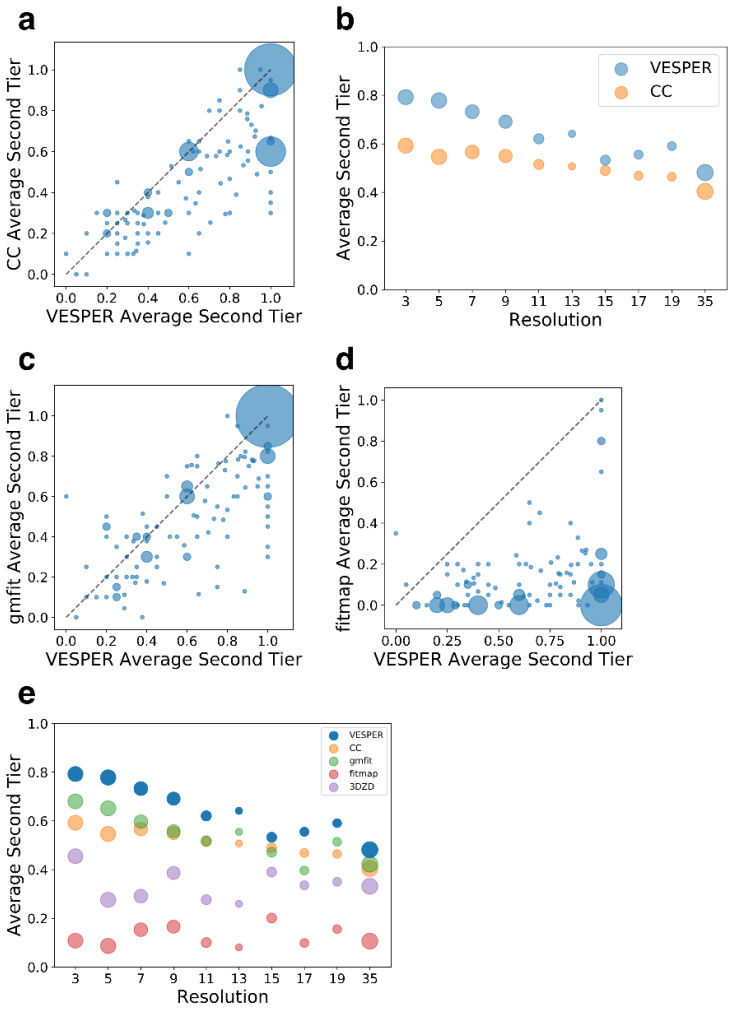


Figure 3.6: Performance of partial map search in terms of correct hits within the second tier. Corresponding results using the first tier are shown in Figure 3.5. a, Average fraction of correct hits within the second tier for each map group. the x-axis, VESPER with the DOT score; the y-axis, CC. b, Comparison of VESPER and CC on maps at different resolutions. The average fraction of correct hits within the first tier was considered. Blue, VESPER with the DOT score; orange, CC. c, Comparison between VESPER and gmfit in terms of the average fraction of correct maps within the second tier for each map group. d, Comparison between VESPER and fitmap in terms of the average fraction of correct maps within the second tier for each map group. e, Comparison of the average fraction of correct maps within the second tier at each resolution bin for VESPER (blue), CC (orange), gmfit (green), fitmap (red), and 3DZD (EM-SURFER) (purple).

3.3.7 Partial map alignment accuracy

In the last section, we discuss local map alignment accuracy of the methods. Nine maps were selected for a test set, three each from resolution ranges of better than 5 Å, 5-10 Å, and over 10 Å (Table 3.4). These maps have an associated PDB entry of a protein structure that covers most of the region of the maps and do not contain nucleic acid structures. Therefore, an alignment of maps can be evaluated in terms of RMSD of the inferred superimposition of the protein structures. From each map, the density region of each chain was manually segmented out using Zone in Chimera, which was used as a query for searching against the original map.

Table 3.4: EM maps used in partial map alignment evaluation

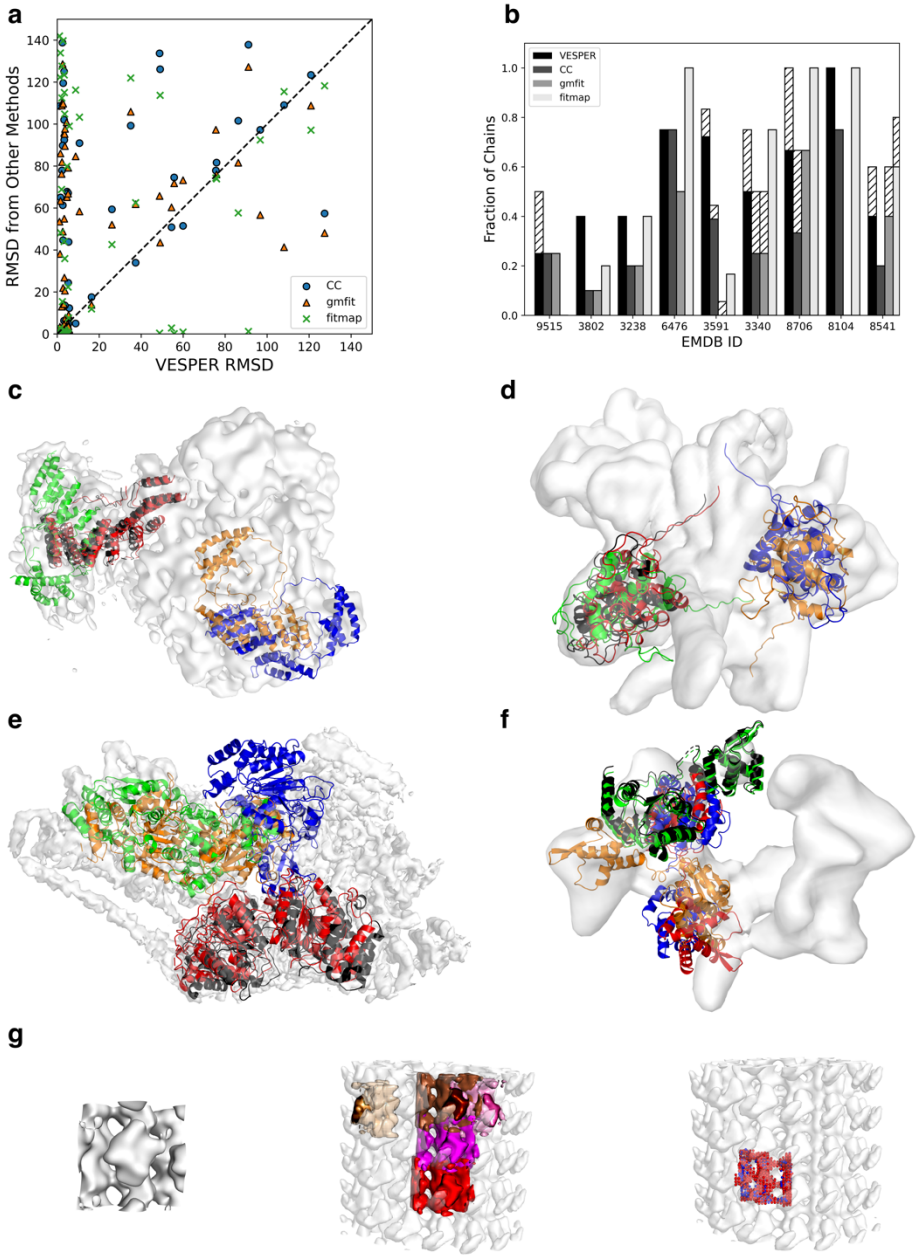
Resolution Bins	Protein Name	Map ID	Number of Chains	Chain IDs
< 5 Å	Voltage-gated calcium channel	9515/5gjw	4	A, C, E, F
	Transcription factor IIH	3802/5of4	10	A, B, D-H, X-Z
	γ-secretase	3238/5fn3	5	A-D, G
5 -10 Å	Voltage-gated calcium channel	6476/3jbr	4	A, B, E, F
	RNA polymerase I-Rrn3-CF complex	3591/5n5z	18	A-R
	Hsp90-Cdc37-Cdk4 complex	3340/5fwp	4	A, B, E, K
> 10 Å	Dynein-Lis1 complex	8706/5vlj	3	A-C
	NMDA receptor	8104/5ipt	4	A-D
	Origin recognition complex	8541/5ujm	5	A-E

Map ID shows the EMD-ID and the associated PDB ID of the macromolecules. The Chain IDs show the chains that were used as queries of the local alignment evaluation and the number of chains indicates the number of query chains. The name of the four entries that were shown in Figure 3.7 are underlined.

Results are summarized in Figure 3.7. The first panel, Figure 3.7a, shows the RMSD values of the top-scoring alignment of 57 queries computed by the four methods. To characterize the performance of VESPER, the plot compares VESPER with each of the other methods. Among the 57 queries, VESPER aligned more query maps, 33 (57.9%) and 39 (68.4%) within an RMSD of 5.0 Å and 10.0 Å, than the other three methods, as shown in the plot. CC, gmfit, and fitmap had 20/22, 10/13, 24/25 maps within an RMSD of 5.0/10.0 Å, respectively. The same plots for individual maps are provided in Figure 3.8. Figure 3.7b is a breakdown of the alignment results for each EM map in Table 3.4. It counted the fraction of query chains that were aligned within 5.0

Å and 10.0 Å by each method. It is interesting to see that the performance of all the methods for the first three maps, EMD-9515, EMD-3802, and EMD-3238, which have a relatively higher resolution of less than 5.0 Å, were worse than the rest of the maps. This is probably because maps at a high resolution have more detailed gradients in the density, which could cause an alignment to be locally trapped.

Figure 3.7: Performance of local map alignment. a, Comparison of RMSD of the top-scoring map alignment by VESPER with CC, gmfit, and fitmap. EM maps in the dataset are listed in Table 3.4. Blue circles, comparison against CC; orange triangles, gmfit; green crosses, fitmap, respectively. For EMD-3802, chain E and F are considered similar and both chain locations were considered as similar (RMSD: 4.20 Å over 160 residues) and thus an additional correct position for each chain and the RMSD was considered as such. Similarly, for EMD-3340, chain A and B are considered as similar enough to be counted as an additional correct alignment position (RMSD: 1.23 Å over 616 residues). The same plot for individual maps is provided in Figure 3.8. b, the fraction of query chains for each map that had the top-scoring alignment with an RMSD of 5.0 Å or less (solid gray bars) and 10.0 Å or less (including hatched bars). Black bars, VESPER; dark gray, CC; medium gray, gmfit; pale gray, fitmap. The same type of plot that considers the lowest RMSD alignment within the top five scoring alignments is provided as Figure 3.9. c, Alignment of Rrn11 (PDB ID: 5n5zR) with the RNA polymerase I-Rrn3-CF complex (EMD-3591, 5n5z). The correct position of Rrn11 is shown in black. Best-scoring alignment by VESPER, CC, gmfit, and fitmap is shown in red, blue, orange and green, respectively. RMSD of these four alignments by VESPER, CC, gmfit, and fitmap is 3.27 Å, 125.07 Å, 95.88 Å, and 44.08 Å, respectively. d, Cdk4 (5fwpK) aligned with the Hsp90-Cdc37-Cdk4 kinase complex (EMD-3340, 5fwp). The correct position of Cdk4 is shown in black. Color code of chains for the methods is the same as the panel c. RMSD of the aligned poses of the four methods (the same order as panel c) is 5.15 Å, 67.18 Å, 66.85 Å, and 22.34 Å, respectively. e, Alignment of the XPB subunit (5of4A) with the density map of human transcription factor IIH (EMD-3802, 5of4). The color code of the chains is the same as in panel c. The RMSD of the four methods is 4.63 Å, 67.87 Å, 65.19 Å, and 79.91 Å. f, Alignment of the subunit 2 (5ujmB) with the complete origin recognition complex (EMD-8541, 5ujm). The RMSD by the four methods is 54.43 Å, 50.85 Å, 60.35 Å, and 2.85 Å, respectively. g, kinesin-5 motor domain attached to microtubule (left, EMD-2541, resolution: 25 Å) matched to a map of the complete microtubule (middle and right, EMD-1026, resolution 25 Å). The colors in the middle panel indicate the five top-scoring positions VESPER identified. The score was higher in the following order: red, brown, magenta, pink (on the right), and light yellow (on the left). The panel on the right visualizes the dot product with blue and red for vectors with negative and positive scores, respectively.


a
RMSD from Other Methods
VESPER RMSD
CC
gmfit
fitmap
b
Fraction of Chains
VESPER
CC
gmfit
fitmap
9515
3802
3238
6476
3591
3340
8706
8104
8541
EMDB ID
c
d
e
f
g

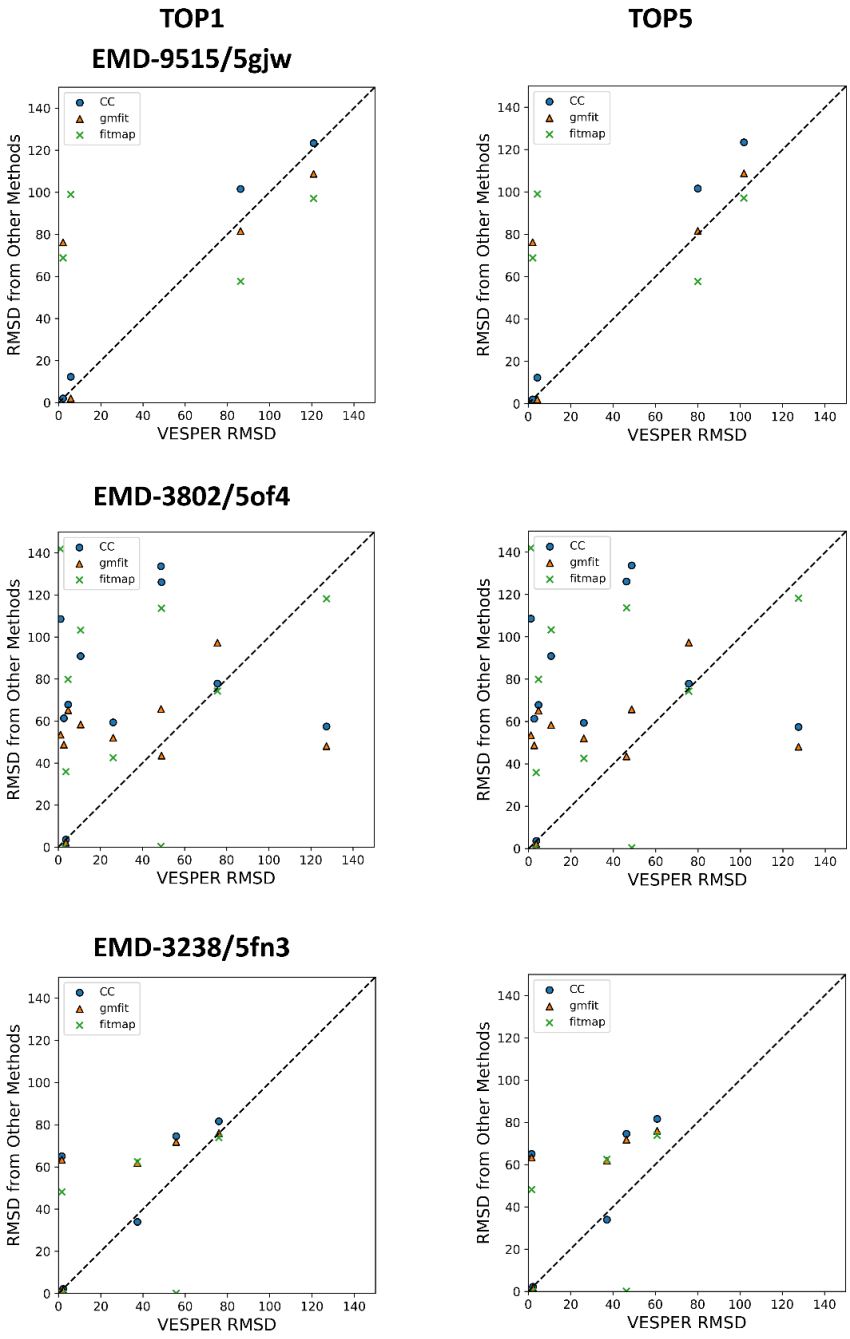


Figure 3.8: Performance of local map alignment of individual maps. Comparison of RMSD of the top-scoring map alignment (in the left column) and best (smallest) RMSD among the top 5-scoring maps by VESPER with CC, gmfit, and fitmap. EM maps in the dataset are listed in Table 3. The summary plot from the nine maps is provided as Figure 3.7.

Figure 3.8 continued

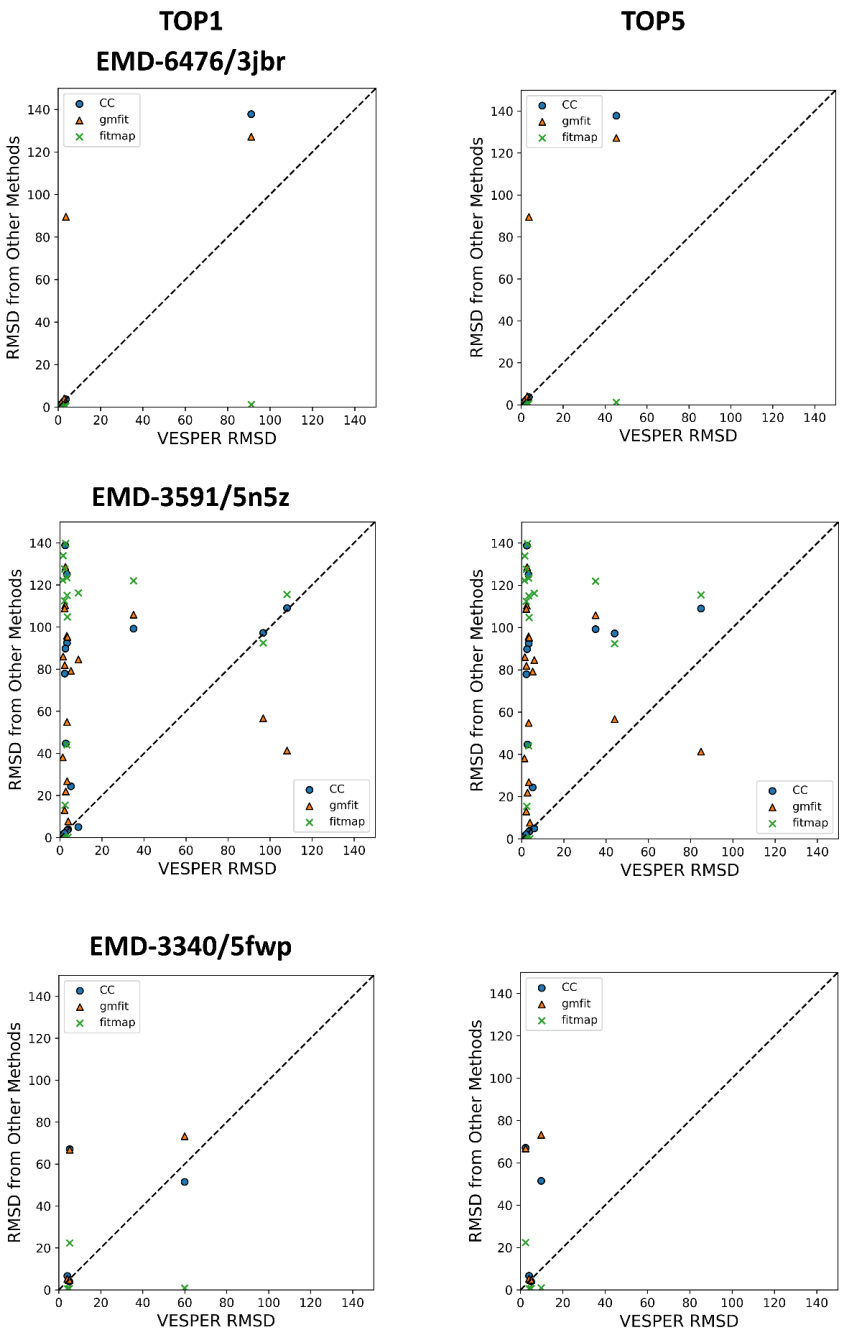
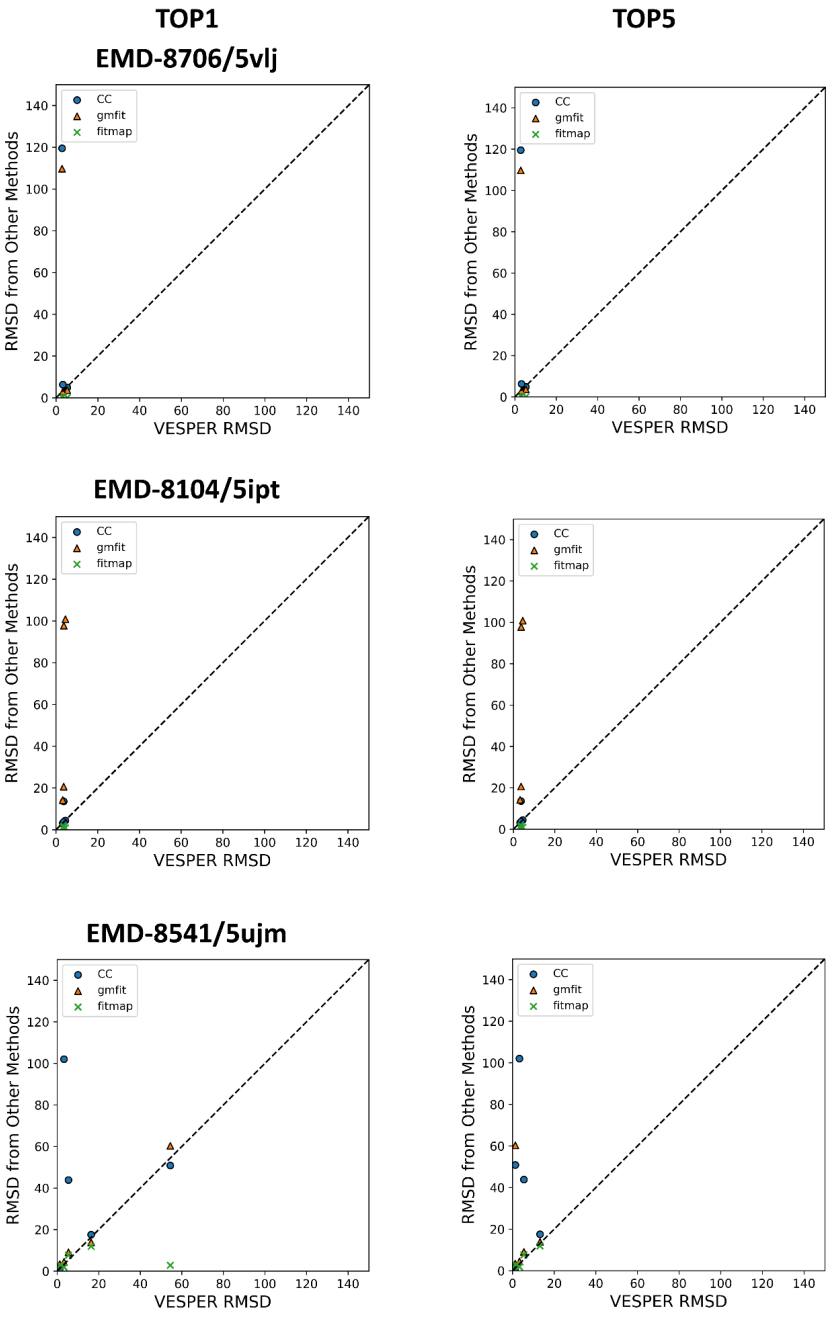


Figure 3.8 continued



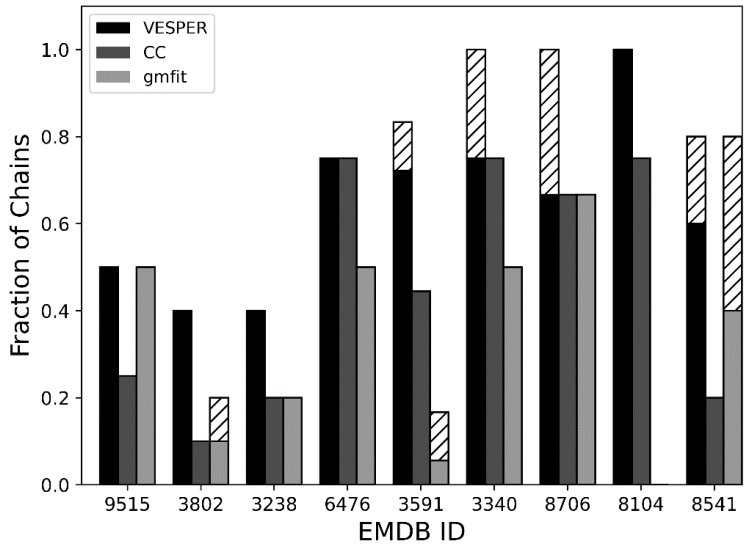


Figure 3.9: The fraction of query chains for each map that had an alignment with an RMSD within 5.0 Å (solid bars) and 10.0 Å (hatched bars) among the top five scoring alignments. Black, VESPER; dark gray, CC; medium gray, gmfit. The same type of plot that considers the top-scoring alignments is provided as Figure 3.7b**.** This plot does not include results by fitmap because fitmap only outputs one alignment.

3.4 Discussion

We have developed VESPER, an exhaustive vector-based FFT search method to identify the best superimposition of EM maps. When benchmarked on the global matching dataset and partial matching dataset, VESPER showed better performance in comparison with four other methods: CC, gmfit, fitmap and 3DZD. In global matching, performance from VESPER is close to but slightly better than gmfit. This is because the vector information extracted from VESPER can be similar to the GMM representation used in gmfit, but without the restriction of the number of Gaussian distribution functions. In the vector representation, each unit vector points to the direction of the local dense region, which can be close to the center of the Gaussian distribution that grid point belongs to. This is why the global retrieval performance from VESPER is very close to gmfit. Compared to other methods, fitmap has the worst performance in both global and partial matchings. fitmap heavily depends on the initial placements, as it only performs local refinement around each initial placement. If none of the initial placements is close to the correct superimposition, fitmap cannot identify a good match between two maps.

When we analyzed VESPER's performance, we have focused on the best superimposition between two maps. But VESPER can output multiple alternative superimpositions after local refinement. It allows users to compare, examine and choose a proper fitting from multiple possible positions. One application is to fit the map of the asymmetric unit or subunit to the complete symmetric structure, which has multiple copies of the subunit. For example, if we have the map for the asymmetric unit of a microtubule cylinder, VESPER is able to identify possible positions of the asymmetric unit on the complete microtubule cylinder. Another advantage of VESPER is that, for each unit vector, it will report a fitness score ranging from -1 to 1 to show how good the match is. By coloring the vectors after the fitting process, users can identity regions where two maps agree well to each other and regions where one map is very different from the other map.

VESPER may also help in fitting biomolecular atomic structures and segmenting EM maps. It can be hard to fit atomic structures to the map if the resolution of the map is relatively low [133]. One potential solution is to convert each subunit into electron density map, and use VESPER to identify the location and orientation that best matches the atomic structure to the map. Similarly, by fitting the map of each subunit into the complete map, we can segment the complete map into regions for each subunit. Alternatively, if the map of some subunits is already solved by

other researches before, VESPER can find the correct region for those subunits in a new map and facilitate the annotation process. Therefore, we expect VESPER will serve as an indispensable addition to the structural biology toolbox, allowing users to obtain a good fit between EM maps.

CHAPTER 4. 3D SHAPE RETRIEVAL CONTEST (SHREC) OF PROTEIN SHAPE AND TOMOGRAM CLASSIFICATIONS[2]

In Chapter 2, I demonstrated that 3DZD can represent the surface shape of proteins and is thus utilized to construct the protein shape universe for both single-chain and complex proteins. Besides analysis of protein atomic structures, 3DZD is also a valuable resource for interpreting medium- to low-resolution protein structure images of determined by cryo-electron microscopy and tomography. As introduced in Chapter 3, previously our lab has developed a computational method based on 3DZD to efficiently search low-resolution EM maps in EMDB. In this chapter, I report on the performance of our group's protein shape comparison method in SHREC, a community-wide evaluation of 3D-shape retrieval algorithms.

4.1 Background

SHREC (3D Shape Retrieval Contest) was established in 2005 to evaluate the effectiveness of 3D-shape retrieval algorithms. It is organized in conjunction with the Eurographics Workshop on 3D Object Retrieval, where the evaluation results will be presented and included in the workshop proceedings. Every year, there are multiple tracks in SHREC that focus on different applications of shape retrieval. Organizers of each track propose the task, build the dataset, and evaluate the performance of all participants in that track. Our group has participated in SHREC 2017 in the track of classification of protein shapes, and SHREC 2019 in the track of protein shape retrieval and the track of classification of cryo-electron tomograms.

Proteins are complex macromolecular molecules constituted of hundreds to millions of atoms. They are usually classified according to their function in the cellular environment. There are multiple approaches to quantify the similarity of protein structures. The most intuitive way is to calculate the root mean square deviation (RMSD), which requires residue correspondence between two proteins. Another useful way to represent proteins is to compute their surfaces, typically representing their solvent excluded surface (SES) as defined by Connolly. Detection of protein surface shape similarity has many applications, including drug discovery pipelines, adverse drug event prediction, and characterization of molecular processes and diseases. However, it is

[2] Portions of this chapter have been previously published [134-136]

still very challenging to detect and characterize protein surface shape similarity because of protein conformation changes when they bind to other proteins and ligands. Therefore, in SHREC 2017 and SHREC 2019, the organizers have proposed a track for protein shape comparison to evaluate the effectiveness of existing methods.

There is a resolution gap in knowledge of cellular life between the molecular level and the cellular level. At molecular level, protein structures can be determined by techniques such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy. At cellular level, the structures are typically obtained by light microscopy techniques. cryo-electron tomography (cryo-ET) is the technique that has the potential to bridge this gap by simultaneously visualizing the cellular architecture and macromolecular structures. Due to the low signal-to-noise ratio of the tomograms and the large amount of data, manual localization of the particles by experts is rarely feasible. Machine learning approaches have been successfully applied to cryo-ET. Support vector machines have been used for both detection and classification [56]. With ever increasing amount of data captured by cryo-EM and -ET methods [58], deep learning methods are gaining popularity. Models were proposed for localization [59], classification [57], end-to-end segmentation [60] and structural mining [137], providing potentially faster, reference-free, and often more accurate results than template matching. To evaluate the performance of existing computational methods, the organizers have proposed a task of localization and classification of particles in the cryo-electron tomogram volume in SHREC 2019.

4.2 Methods

4.2.1 Method for protein shape retrieval in SHREC 2017

In SHREC 2017 track: Protein Shape Retrieval, the goal is to find top 200 models from 5,856 shapes in the database for each query shape. The query set is composed of 10 proteins which are obtained from the Protein Data Bank [30] (Table 4.1). In order to ensure structural diversity, the 10 proteins were manually picked from the special collection, the molecule of the Month, which summarizes the structure and function of one important protein molecule each month by David Goodsell [138]. The organizers construct the target set from a subset of PDB models, composed of 13,182 non-redundant protein structures. These structures are abstracted to shape models that fit in a sphere of the same radius (30 Å). The atoms are mapped to 3D grid points separated by 1

Å. These operations ensure that the biological features are removed to a good extent, so that the technology in computational shape comparison can be applied. In the second stage, for each query structure, the organizers used the exhaustive model matching method to rank the 13,182 structures, and then chose the top 1,000 shape models. Because same model can be present in top 1000 for different queries, the final dataset contains 5,854 unique models after removing repeated models. This is the target set for shape retrieval.

Table 4.1: The list of query molecules in SHREC 2017

Protein	PDB ID
Lysozyme	2LYZ
Antibody	4MMV
HIV reverse transcriptase	3HVT
Insulin	2HIU
HSP90	2CG9
Bacteriophage	1CD3
G protein	1GG2
Ribosome	4V4J
Penicilin-binding protein	1HVB
Zika virus	5IRE

This table is adapted from [134].

The specific task of retrieval is to select top 200 models from 5,854 model structures for each query shape, and rank the models by their similarity to the query. To generate the rank list for each query, I collaborated with Genki Terashi and developed our method that integrated results from two parts: global surface comparison using 3D Zernike descriptors (3DZD) [99, 100], and structure similarity comparison using MM-align [8]. I firstly generated simulated EM maps at resolution 6 Å from the PDB file for all queries and model structures. 3D Zernike descriptors (3DZD) was then calculated for each simulated map at contour level 0. For each query, to generate the retrieval list, I have sorted all models by the Euclidean distance between their 3DZD and the 3DZD for the query, and removed models whose sizes are very different from the query (number of residues is above 1.25 times or below 0.8 times the number of residues in the query). Then we have added information from MM-align to further refine the rank lists. Genki have aligned each query structure to all model structures with similar sizes, and only kept models whose TM-score

is equal to or better than 0.5. I then added those models to the top of the retrieval list for each query.

There are four teams who have participated in the Protein Shape Retrieval track in SHREC 2017. The organizers have used the results from 3D Zernike moments-based method (3DZM) as the ground truth for this track [134]. Given two protein structures, the organizers have applied Fast Fourier Transform (FFT) to sample all rotation and translation combinations, and calculated 3D Zernike moments for each protein at each orientation [139]. They have then calculated the correlation coefficient (CC) between two proteins at certain orientation using Equation 4.1:

$$CC = \frac{< \rho_1 \rho_2 > - < \rho_1 >< \rho_2 >}{\sigma(\rho_1)\sigma(\rho_2)} \tag{4.1}$$

where ρ_1 and ρ_2 are 3D Zernike moments for the first protein and the second protein, respectively. The similarity between two protein structures is measured by the maximum CC among all orientations. For each query protein, top 200 models with the largest CC from this 3DZM method are considered as true positives in the following evaluation.

The retrieval performance from each team was evaluated using seven metrics: consistency with the ground truth, Nearest Neighbor (NN), First tier (FT), Second tier (ST), Precision-Recall plot, E-measure, and Mean Average Precision (MAP). Consistency with the ground-truth is measured by the average CC between the query and top N models. Here N was set to 1, 3 and 5. NN, FT and ST check the ratio of models that belong to the same class as the query. NN considers only the first match, excluding the query itself. FT and ST consider top |C| -1 and 2 * (|C| - 1) matches, respectively, where |C| is the number of models in the query's class. In the Precision-Recall plot, Precision P measures the ratio of retrieved models from the query's class C among all retrieved models, while Recall R measures the ratio of retrieved models from class C compared to |C|. E-measure is a combination of both Precision and Recall, which is calculated by:

$$E - measure = 1 - \frac{2}{\frac{1}{P} + \frac{1}{R}} \tag{4.2}$$

Mean Average Precision measures the mean of average precision of all queries. The average precision of a query is the average of all precision values computed when each relevant domain is found.

4.2.2 Method for protein shape retrieval in SHREC 2019

In SHREC 2019 track: Protein Shape Retrieval, instead of comparing the shape similarity of complete protein structures, the goal is to retrieve evolutionary classification of domains based on their surface meshes only. The organizers have used the Structural Classification of Proteins – extended (SCOPe) database [25] to generate a dataset of 5,298 protein domains. All domains are used as a query against the whole dataset. The specific task is to produce a distance-to-the-query dissimilarity matrix. The retrieval performance was evaluated at the *species* level and the *proteins* level.

When constructing the dataset of domains, protein structure flexibility was an important consideration. Proteins display various motions reflecting both the relative motion of their atoms and small to large conformational changes to perform their cellular activities. To incorporate the protein motion information, the organizers only kept SCOPe entries: 1) from NMR structures whose conformers display the same number of atoms; 2) from three Classes: All alpha proteins, Alpha and beta proteins (a+b), and Alpha and beta proteins (a/b); 3) with at least four ortholog proteins [135]. To reduce the dataset size, they have randomly selected 5,298 domains, representing 241 SCOPe entries from 211 PDB structures.

To generate the dissimilarity matrix of 5,298 domains, I have collaborated with Genki Terashi and Tuan M. Lai, and applied three approaches, all of which are based on 3D Zernike descriptors (3DZD). The idea behind the proposed framework is to represent the protein global surface information with 3D Zernike Descriptors and quantify the similarity between 3DZDs by either Euclidean distance or similarity score from neural network. To calculate 3DZD for each protein, the surface triangulation of solvent excluded surface was mapped onto a 3D cubic grid, where each voxel (a cube defined by the grid) was assigned either 1 or 0: 1 for a surface voxel that locates closer than 1.7 grid interval to any triangle defining the protein surface, and 0 otherwise. This 3D grid with 1s and 0s was considered as a 3D function $f(x)$, from which 3DZD is computed.

In the first approach, the global surface similarity between two proteins was quantified by Euclidean distance of their 3DZDs. Small distance values indicate that two proteins share similar global surface. In this calculation, we took the triangulated surface (.off file) for each of the 5,298 domains as the input to 3DZD computation and calculated the 121-dimensional vector for each domain. Euclidean distances between one query domain against all other 5,297 domains were

calculated and put into each row in our first distance matrix. I have performed the 3DZD and Euclidean distance calculations for this approach.

In the second approach, we have built a deep learning based model to quantify the similarity between protein structures. The model was trained on all proteins in the SCOPe 2.07 database. Solvent excluded surface of each protein was generated using the EDTSurf software [115, 140]. The triangulated surface was then taken as the input to 3DZD computation, which produced 121-dimensional vector for each protein. On a high level, given a pair of protein structures, the deep learning model outputs a score between 0 and 1 indicating their similarity (the higher the score, the more similar the structures). The model consists of an encoder whose role is to compute key features from a 3DZD vector. The encoder is a feed forward neural network consisting of three hidden layers. Each layer uses ReLU as the activation function. Intuitively, each hidden layer of the encoder computes a new level of representation of the original 3DZD vector. Given two protein structures as input, the model uses the encoder to compute new features for each protein's 3DZD vector. The computed features and the original 3DZD vectors of the two structures are then compared using various operations such as Euclidean distance, Cosine similarity, element-wise absolute difference, and element-wise product. The comparison results as well as additional features such as the difference in number of vertices and the difference in number of faces are together fed into a final feed forward neural network that outputs a score between 0 and 1. We used techniques such as batch normalization and dropout to improve the training process. In the training data, if two protein structures have the same protein level, they are considered as being similar.

The third approach is similar to the second approach, except that in this case we trained the model to consider two protein structures to be similar only when the two structures have the same species level. In the second approach, if two protein structures have the same protein level but different species level, we still consider them as being similar. In the second and third approaches, Genki has downloaded all protein coordinates from SCOPe 2.07 database. I have calculated 3DZD vectors for each protein in the dataset. Tuan has then trained the neural network using 3DZD vectors as input and generated the dissimilarity matrix.

There are five teams who participated in the Protein Shape Retrieval track in SHREC 2019. The retrieval performance from each team was evaluated at both the *proteins* level and the *species* level, as defined in the SCOPe database. At the *proteins* level, it evaluates the ability to retrieve a

conformation from ortholog proteins; while at *species* level, it evaluates the ability to retrieve a conformation from the protein of a given species. At each level, the performance from each team was evaluated by six metrics: Nearest Neighbor (NN), First tier (FT), Second tier (ST), Precision-Recall plot, E-measure, and Mean Average Precision (MAP).

4.2.3 Method for classification in cryo-electron tomograms in SHREC 2019

In Classification in Cryo-Electron Tomograms track in SHREC 2019, the focus is on the localization and classification of particles in the cryo-electron tomogram volume. The organizers provide 10 simulated cryo-electron tomograms. Each tomogram is filled with on average 2,500 proteins from 12 different classes. The PDB IDs of those 12 protein complexes are: 1bxn, 1qvr, 1s3x, 1u6g, 2cg9, 3cf3, 3d2f, 3gl1, 3h84, 3qm1, 4b4t, 4d8q. The dataset construction started with creating the original density maps (grandmodels). The protein complexes were placed into the grandmodel at random locations in random orientations without overlapping each other. The space in-between proteins was filled with vitreous water, which was subjected to structural noise (stdev = 0.05). The organizers have then created a series of projection images of the grandmodel with a signal-to-noise ratio of 0.02, applied a contrast transfer function correction to each projection image, added shot-noise, and done a weighted back-projection reconstruction. The resulting reconstructions have a resolution of 1nm/voxel and have a size of 512*512*512 voxels. The center coordinates and class label were offered for all particles in 9 out of 10 tomograms for learning-based methods. Participants were asked to submit the center coordinates and class label for particles in the test tomogram.

To predict the location and label of particles in the tomogram, our lab has used a method based on a deep learning model. It predicts whether a specific location in the tomogram is a center of a protein and if so, it predicts the class label of the protein. Given a specific point of interest, the model took three 2D image slices as input whose centers are at the point's location and that are parallel to the XY, XZ, and YZ-planes. The size of each 2D input slice was selected to be 32 by 32. Each input slice was encoded into a vector consisting of 128 numbers using a convolutional neural network whose architecture resembles ResNet [141]. After this step, we concatenated the three encoded vectors for the three slices into one vector consisting of 384 numbers in total. We then fed the new vector into a feed forward neural network consisting of two hidden layers. At the end, the feed forward neural network outputs 13 probability scores. The first score corresponds to

the predicted probability of the location of interest is not the center of any protein. Each of the remaining 12 scores corresponds to the predicted probability of the point is the center of a particular type of protein. The sum of the 13 probability scores is equal to 1. In this track, I have calculated the size information for each of those 12 protein complexes. Tuan then decided the optimal 2D slice size given this information and trained the neural network.

In order to train the proposed deep learning model, we used the ground truth information provided in the first 9 tomograms to generate training examples. Each example consists of three 2D slices whose centers are located at the same point and the correct protein label for the point. In order to generate positive examples, we checked the locations of the proteins in the tomogram from the information files provided by the organizers. In order to generate negative examples, we randomly sampled points in the tomogram that are not too close to any protein. In the end, we generated about 20,000 positive examples and about 20,000 negative examples.

During inference (i.e., when generating the final predictions for the final tomogram), we first ran our neural network for each location in the tomogram (using a stride of 2). Note that if a particular point was predicted to be the center of some protein, other points that are next to it will be likely to have high probability scores of being the center of the same protein. In order to avoid having too many predicted centers at essentially the same locations, we did a final simple post-processing to alleviate the problem. Essentially, we looked at the neighborhood locations of a point and use a majority voting to determine the final label for the point. For example, suppose a point is predicted by the deep learning model to be the center of a protein, but if the majority of the neighborhood points are predicted to be negative, the final label for the point of interest will be changed to negative.

To evaluate the performance of the results from each participant, the organizers firstly built a “hitbox” volume using an automatic script based on ground truth information. This volume consists of bounding boxes that can be traced back to corresponding ground truth particle. Next, for each predicted particle in the submitted result, the organizers analyzed whether it lies within any bounding box, and recorded statistical information, including whether the predicted class is correct, how far the predicted particle center is from the real center. Five metrics were used to evaluate the performance: Precision, Recall, F1 score, and False Negative Rate (FNR). F1 score is the harmonic mean of Precision and Recall. FPR measures the percentage of results that yield negative test outcomes.

4.3 Results

4.3.1 Performance in protein shape retrieval in SHREC 2017

Among the 10 query shapes, there are two queries (q_11 and q_14) whose identical models are also included in the target database. Our 3DZD based approach (named "Kihara") can retrieve the identical model as the best matched model for both queries, whereas the methods from two other groups retrieve the identical model for q_11 at rank 2 and rank 37, respectively. It shows that our approach provides a stable and effective performance to find a protein in a large dataset.

The organizers have calculated the average correlation coefficient between queries and their top 1, top 3, and top 5 models for each submitted method in this track (Table 4.2). 3DZM is the ground truth approach proposed by the organizers. 3DZD is a new method proposed by the organizers and is compared to methods from other groups as well as the ground truth 3DZM. The 3DZD method uses the correlation coefficient of 3D Zernike descriptors to quantify the similarity of two protein structures. As both 3DZM and 3DZD maximize the correlation coefficient in their search process, those two methods have achieved the best performance in terms of average correlation coefficient. Our "Kihara" approach shows the third-best performance.

Table 4.2: The average correlation coefficients between queries and their top N models

Method	Top 1	Top 3	Top 5
3DZM	0.769	0.718	0.701
Kihara	0.715	0.641	0.619
MS-DEM-MAD	0.625	0.608	0.598
MS-DEM-RMSD	0.629	0.614	0.597
PCAS	0.646	0.619	0.607
3DZD	0.723	0.657	0.643

This table is adapted from [134].

Figure 4.1 shows the evaluation curves for all methods in this track. As shown in Figure 4.1a, 3DZM and 3DZD show the best performance in average correlation coefficient, as this is the score they maximize when they rank the models for each query. This result is also consistent with the observation shown in Table 4.2. The Precision-Recall plot indicates that at low recall, our Kihara method is able to achieve the second-best precision. This means that our method can

correctly retrieve more relevant models at the top. Table 4.3 shows the evaluation metrics for all five methods. Our method shows the third-best performance in NN and MAP.

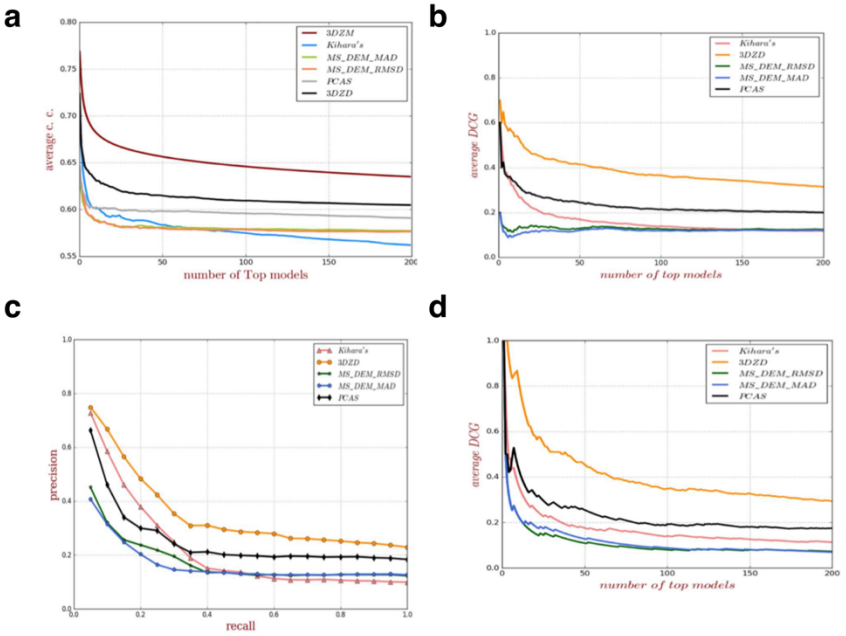


Figure 4.1: Evaluation Curve. a, The average correlation coefficients for all queries. The plots indicate that the top 10 models can be considered to be correct for most methods, indicated by the sharp transitions to the plateaus. b, The average DCG for all queries. The ranking results from 3DZM method was used as the 'ground truth' for DCG calculation. c, Precision-Recall Curve for all the queries. d, The average DCG for queries q_11 and q_14. The exact shape models for these two queries are present in the target set. Adapted from [134].

Table 4.3: Statistics of evaluation parameters in SHREC 2017

Method	NN	FT	ST	MAP	E
Kihara	65.29	18.40	9.20	20.34	20.3
MS-DEM-MAD	31.40	24.35	12.15	15.16	26.21
MS-DEM-RMSD	34.92	23.92	11.95	16.02	25.85
PCAS	68.41	35.00	17.50	24.94	29.19
3DZD	74.44	26.70	22.20	34.44	38.30

This table is adapted from [134].

4.3.2 Performance in protein shape retrieval in SHREC 2019

Figure 4.2 shows the Precision-Recall plot for each submitted method at both *proteins* level and *species* level. All 13 submitted methods have shown better performance at the *proteins* level than at the *species* level. Our methods named as "3DZD" have shown similar performance to HAPT, the method from another group which characterizes protein shapes with the Histograms of Area Projection Transform Both methods have achieved better retrieval performance compared to methods from other groups. Even for high recall values, our method can still display good precision. The two deep learning based models, 3DZD2 and 3DZD3, have improved the precision significantly at medium and high recall values.

Table 4.4 and Table 4.5 summarizes the evaluation metrics at the *species* level and at the *proteins* level, respectively. Consistent with the observations in Figure 4.2, all methods have achieved higher values at the *proteins* level than at the *species* level. At the *proteins* level, the number of available domains is bigger, thus the learning-based methods have more data to learn from. The protein structure differences are also bigger at the *proteins* level. Those are two possible explanations for the better performance at the *proteins* level. The retrieval performance of our 3DZD approaches are among the best of all 13 submitted methods. At the *proteins* level, more than 98% of first retrieved domains is from the same level as the query. Our methods have also shown higher FT and ST values. The training in 3DZD2 and 3DZD3 has improved all evaluation metrics. 3DZD2 and 3DZD3 always show better retrieval performance compared to 3DZD1. Thus, instead of using Euclidean distance, the more complicated combination of components in 3DZD vector learnt in 3DZD2 and 3DZD3 is able to pick up more correct structures at the top.

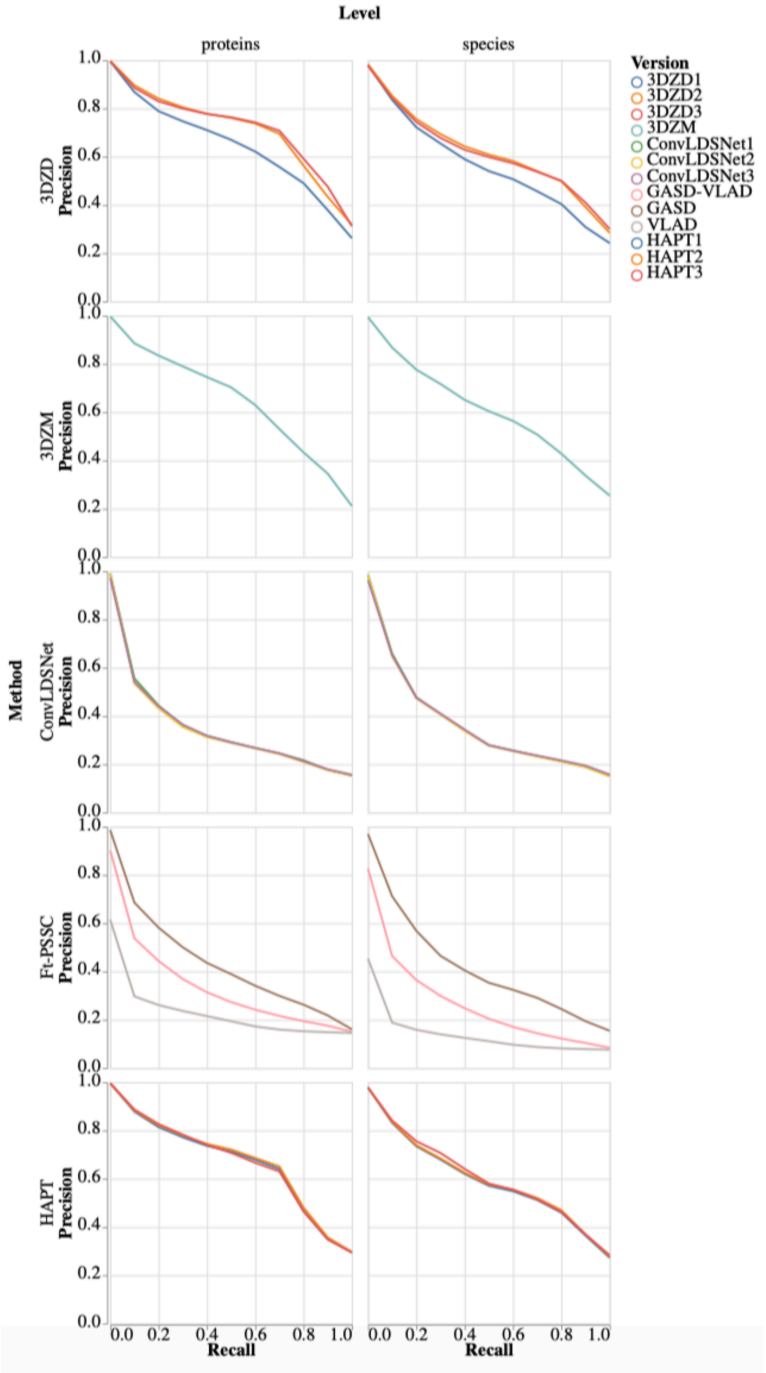


Figure 4.2: Precision-Recall curves for the *proteins* (left) and *species* (right) level. Each row shows the precision-recall curve for one method. Adapted from [135].

Table 4.4: Nearest-neighbor (NN), First Tier (FT), Second Tier (ST), Mean Average Precision (MAP) average values computed at the *species* level

Method	Approach	NN	FT	ST	MAP
3DZD	3DZD1	0.931	0.528	0.641	0.556
3DZD	3DZD2	0.964	0.589	0.705	0.610
3DZD	3DZD3	0.951	0.577	0.716	0.605
3DZM	3DZM	0.989	0.577	0.675	0.604
ConvLDSNet	ConvLDSNet1	0.975	0.332	0.421	0.355
ConvLDSNet	ConvLDSNet2	0.975	0.330	0.423	0.353
ConvLDSNet	ConvLDSNet3	0.952	0.333	0.422	0.355
Ft-PSSC	GASD-VLAD	0.688	0.266	0.385	0.245
Ft-PSSC	GASD	0.955	0.382	0.467	0.405
Ft-PSSC	VLAD	0.266	0.144	0.244	0.122
HAPT	HAPT1	0.947	0.555	0.705	0.578
HAPT	HAPT2	0.951	0.549	0.693	0.584
HAPT	HAPT3	0.944	0.563	0.709	0.588

This table is adapted from [135].

Table 4.5: Nearest-neighbor (NN), First Tier (FT), Second Tier (ST), Mean Average Precision (MAP) average values computed at the *proteins* level

Method	Approach	NN	FT	ST	MAP
3DZD	3DZD1	0.988	0.579	0.729	0.638
3DZD	3DZD2	0.993	0.658	0.789	0.712
3DZD	3DZD3	0.995	0.664	0.802	0.720
3DZM	3DZM	0.994	0.583	0.706	0.649
ConvLDSNet	ConvLDSNet1	0.984	0.303	0.458	0.329
ConvLDSNet	ConvLDSNet2	0.984	0.296	0.457	0.324
ConvLDSNet	ConvLDSNet3	0.961	0.301	0.458	0.328
Ft-PSSC	GASD-VLAD	0.797	0.315	0.481	0.315
Ft-PSSC	GASD	0.977	0.372	0.506	0.417
Ft-PSSC	VLAD	0.390	0.226	0.380	0.206
HAPT	HAPT1	0.988	0.616	0.734	0.659
HAPT	HAPT2	0.988	0.624	0.738	0.666
HAPT	HAPT3	0.991	0.613	0.732	0.658

This table is adapted from [135].

4.3.3 Performance in classification in cryo-electron tomograms in SHREC 2019

Our approach (named as "2.5D-Resnet") does not achieve good localization and classification of tomograms compared to top performing methods from four other groups. In localization, our method shows F1 score = 0.5126, while the best performing method 3D-Unet has achieved F1 score = 0.9169 (Table 4.6). Similar results are observed in classification (Table 4.7). The organizers have also grouped the protein complexes into four groups (Table 4.8) by their sizes and calculated the F1 score for each group. Our method shows better F1 score for proteins with medium and large sizes compared to proteins that are tiny and small. This trend is also observed in other methods. For tiny and small proteins, the F1 score is always much worse compared to medium and large proteins.

Table 4.6: Results of localization evaluation

Submission	RR	TP	FP	FN	MH	RO	AD	Recall	Precision	Miss rate	F1 Score
DoG-3D-CNN	1813	1690	110	850	13	1	2.4519	0.6653	0.9923	0.3346	0.7966
3D-Unet	2887	2163	709	377	15	24	3.5063	0.8515	0.9931	0.1484	0.9169
2.5D-Resnet	4524	1507	1185	1033	876	1	3.9866	0.5933	0.4513	0.4066	0.5126
3D-TM	2429	814	356	1726	425	313	2.5608	0.3204	0.3926	0.6795	0.3529
3D-HN-localization	2127	455	867	2085	311	48	5.9316	0.1791	0.3611	0.8208	0.2394
3D-Unet-CNN-8	2500	1367	372	1173	480	13	4.1660	0.5381	0.6423	0.4618	0.5856
3D-Unet-CNN-12	2500	1438	555	1102	352	12	4.4083	0.5661	0.7393	0.4338	0.6412
2.5D-SSD-3D-CNN	1977	710	196	1830	485	7	4.6453	0.2795	0.3986	0.7204	0.3286

This table is adapted from [136]. RR: results reported; TP: true positive, unique particles found; FP: false positive, reported non-existant particles; FN: false negative, unique particles not found; MH: multiple hits: unique particles that had more than one result; RO: results otuside of volume; AD: average Euclidean distance from predicted particle center; Recall: percentage of total results correctly localized; Precision: percentage of results which are relevant; Miss rate: percentage of results which yield negative results; F1 Score: harmonic average of the precision and recall. The best results in each column are highlighted.

Table 4.7: Results of classification evaluation for all classes

Submission	1bxn	1qvr	1s3x	1u6g	2cg9	3cf3	3d2f	3gl1	3h84	3qm1	4b4t	4d8q
DoG-3D-CNN	0.866	0.619	0.047	0.446	0.343	0.657	0.358	0.225	0.25	0.160	0.222	0.911
3D-Unet	0.904	0.800	0.154	0.522	0.330	0.784	0.584	0.318	0.332	0.193	0.907	0.951
2.5D-Resnet	0.087	0.405	0.119	0.263	0.018	0.566	0.366	0.039	0.293	0.037	0.489	0.359
3D-TM	0.684	0.020	0.005	0.024	0.008	0.194	0.008	0.019	0.032	0.018	0.211	0.890
3D-Unet-CNN-8	0.702	0.559	0	0.234	0.268	0.501	0.209	0.029	0.008	0	0.684	0.711
3D-Unet-CNN-12	0.663	0.577	0	0.243	0.273	0.477	0.209	0.038	0.008	0	0.671	0.728
2.5D-SSD-3D-CNN	0.312	0.343	0	0.054	0	0.166	0.040	0.010	0	0	0.379	0.234

This table is adapted from [136].

Table 4.8: Grouping proteins included in the dataset by their size

Size	Proteins
Tiny	1s3x, 3qml, 3gl1
Small	3d2f, 1u6g, 2cg9, 3h84
Medium	1qvr, 1bxn, 3cf3
Large	4b4t, 4d8q

This table is adapted from [136].

Table 4.9: F1 scores of each submission for size classes defined in Table 4.8

Submission	Tiny	Small	Medium	Large
DoG-3D-CNN	0.144	0.300	0.714	0.566
3D-Unet	0.222	0.400	0.830	0.929
2.5D-Resnet	0.065	0.200	0.352	0.424
3D-TM	0.014	0.000	0.299	0.550
3D-Unet-CNN-8	0.009	0.200	0.587	0.697
3D-Unet-CNN-12	0.012	0.200	0.572	0.699
2.5D-SSD-3D-CNN	0.003	0.000	0.274	0.306

This table is adapted from [136].

4.4 Discussion

Protein structures are very different from the traditional objects commonly used in 3D shape retrievals in computer vision. The differences between protein structures are much smaller. Proteins may also undergo conformational changes to perform their activities. Therefore, it is helpful to develop specific methods for quantifying the shape similarity of protein structures.

Our group has participated twice in the Protein Shape Retrieval track in SHREC. In SHREC 2017, we have used a combination of 3DZD and MM-align to generate the rank list for each query protein. Similarities between two proteins are quantified by the Euclidean distance of their 3DZD vectors, size ratio, and TM-score. With this approach, our lab has achieved the second-best performance among six methods from four groups. In this evaluation, results from 3DZM were considered as the ground truth, which maximizes the correlation coefficient between 3D Zernike moments of two protein structures. As 3DZD (the other approach proposed by the organizers' group) also maximizes the CC, it demonstrates results similar to the ones from 3DZM. But it is questionable whether we should consider the results from 3DZM as the ground truth or not. It is possible that proteins declare to be similar in 3DZM do not share similar structures or functions. Given that we have the atomic structure of all proteins in the dataset, TM-score can be a good approach to quantify the structure similarity in a biologically meaningful way. In SHREC 2019, conformational changes of protein structures were taken into consideration. The organizers have selected protein domains that are solved by NMR and have multiple conformers. We have submitted three 3DZD based approaches in this round: the first approach (3DZD1) simply uses Euclidean distance of 3DZD vectors to quantify domain similarity; the second and third approaches (3DZD2 and 3DZD3) are deep learning-based models trained on *species* and *proteins* level, respectively. As 3DZD2 and 3DZD3 consider more complicated ways to combine the components in the 3DZD vector, they always show better retrieval performance compared to 3DZD1, especially at medium and high recall values. This result suggests that usage of deep learning-based models can be helpful in improving the rank results from 3DZD. One possible direction to explore in future rounds of SHREC is to use a pair of protein structures as the direct input to the neural network and output the similarity score of those two proteins. Given the complicated architecture of neural network, it may be able to capture more detailed shape information in the protein and thus achieve better performance compared to 3DZD.

Protein structures solved by X-ray crystallography and NMR can provide knowledge of cellular life at the molecular level. Cryo-electron tomography has the potential to bridge the knowledge gap between the molecular level and the cellular life to offer a more complete picture. As tomograms have a low signal-to-noise ratio and the amount of data is very big, it is essential to develop automated methods for localization and classification of molecules in the tomogram. Besides the protein shape retrieval track, our lab has also participated in the classification of cryo-electron tomograms track in SHREC 2019. To localize and classify proteins in the tomogram, we have developed a deep learning model which takes the 2D image slices as the input and outputs 13 probability scores. There are still some gaps between the performance from our method and the top performing method in this track. One potential way to further improve the performance is to take in 3D data directly, instead of using 2D slices. Other researchers have shown that, at the same architecture, 3D convolutions perform better than 2D convolutions for 3D data [142]. Protein sizes are also directly correlated with the method performance. The classification task is easier for medium and large proteins compared to tiny and small proteins.

CHAPTER 5. PROTEIN 3D STRUCTURE AND ELECTRON MICROSCOPY MAP RETRIEVAL USING 3D-SURFER2.0 AND EM-SURFER[3]

In Chapter 4, I presented the performance of our 3D Zernike descriptors-based approaches in protein shape retrieval track in SHREC 2017 and SHREC 2019. Compared to other submitted methods, our 3DZD based approaches can retrieve more correct structures at the top. 3DZD is also integrated into two web-based tools (3D-SURFER and EM-SURFER) developed by our lab to provide real-time comparison and analysis of protein structures. In this chapter, I will describe the algorithms and usage of each tool.

5.1 Background

Proteins perform a vast array of functions and participate in essentially every cellular process. The tertiary structure of proteins provides physical platform for carrying out functions. The structure information of proteins thus forms the basis for understanding principles of life and developing new strategies to regulate biological pathways and other processes. As protein structure is directly related to their molecular function, similarity in structure level is more preserved than similarity in sequence level [143]. Therefore, structure-based protein comparison is capable of revealing remote relationships that are hard to detect from sequences.

With the exponential growth of solved protein structures in the Protein Data Bank (PDB) [30] and Electron Microscopy Data Bank (EMDB) [62], it is crucial to develop search tools that can compare protein structures and help understand the relationship between them. Conventional approach for protein structure comparison is aligning atoms or residues of proteins. This approach is time-consuming as sampling of different orientations in the three-dimensional (3D) space is needed, making it inappropriate for searching against a whole structure database in real-time. 3D Zernike descriptors (3DZD) [100] is proven to be suitable for efficient structure comparisons in previous works [4, 5, 7, 99]. It represents a 3D object in a compact vector and in a rotation-invariant fashion, enabling fast search against structure databases.

[3]Portions of this chapter have been previously published [66]

We have developed 3D-SURFER [4, 5] and EM-SURER [7], which are web-based platforms for high-throughput protein structure comparison and analysis. Structures in each server are automatically synchronized with PDB or EMDB weekly. Currently, 3D-SURFER holds over 800,000 entries (including chain, domain, and complex structures) and EM-SURFER holds over 7,000 maps at various map resolutions. Both of them utilize 3DZD to extract global surface information from proteins and quantify shape similarity by calculating Euclidean distance between a pair of 3DZDs. In 3D-SURFER, the VisGrid [144] and LIGSITE^{csc} [145] algorithms are employed to characterize local geometric features of a query protein, including pocket, cavity, protrusion, and flat regions.

5.2 Methods

5.2.1 3DZD calculation in 3D-SURFER

3DZD are based on mathematical series expansion of a given 3D function. It has been applied in various comparisons of biomolecular data [100] including protein-protein docking [146-148], ligand binding pocket comparison [149, 150] and ligand molecule search [151]. It is rotation invariant, that is, prior structure alignment is not required for calculation of 3DZD.

In 3D-SURFER, the calculation of 3DZD starts by construction of surface triangulation using MSROLL [106] and MSMS program [105]. The constructed triangle mesh is then mapped to a 3D grid. Voxels that overlap with protein surface are assigned value 1 and 0 otherwise. This discrete representation is used as the input function *f(x)* for 3DZD calculation. The 3D function *f(x)* is expanded into a series in terms of Zernike-Canterakis basis [123] defined as follows:

$$\Omega_{nl}^{m} = \frac{3}{4\pi} \int_{|x|\leq 1} f(x)\, \bar{Z}_{nl}^{m}(x) dx \tag{5.1}$$

where

$$Z_{nl}^{m}(r, \theta, \Phi) = R_{nl}(r) Y_{l}^{m}(\theta, \Phi) \tag{5.2}$$

The ranges of parameters *m* and *l* depend on order *n*: $-l \leq m \leq l$, $0 \leq l \leq n$, and (*n*-*l*) is even. Previous study has shown that order n = 20 offers a sufficiently accurate representation [123]. Therefore, we set n = 20 which generates 121 invariants. Here $Y_{l}^{m}(\theta, \Phi)$ are spherical harmonics [152] and R_{nl}(r) are the radical functions defined by Canterakis. The rotation-invariant 3D Zernike descriptors are calculated as norms of Ω_{nl}^{m}:

$$F_{nl} = \sqrt{\sum_{m=-l}^{m=l} (\Omega_{nl}^{m})^2} \quad (5.3)$$

The similarity between two sets of 3DZDs is quantified by their Euclidean distance d_E:

$$d_E = \sqrt{\sum_{i=1}^{121} (X_i - Y_i)^2} \quad (5.4)$$

where X_i and Y_i represent the ith invariant for each protein.

5.2.2 Calculation of RMSD

Besides the shape similarity measured by Euclidean distance of 3DZD vectors, 3D-SURFER also calculates the root mean squared deviation (RMSD) of Cα atoms between a query protein and a retrieved structure in the result page. 3D-SURFER uses the combinatorial extension (CE) method [9] to compare and align structures. This algorithm breaks each protein into a set of fragments and then tries to assemble the fragments into the longest continuous path of aligned fragment pairs (AFPs). When adding the next AFP to the alignment path, it considers only the best AFP which extends the path and satisfies the similarity criteria. It uses a z-score to evaluate the statistical significance of the longest alignment path, which reflects the probability of finding an alignment path of the same length with the same or smaller number of gaps and distance from a random comparison of structures using a non-redundant set [153].

5.2.3 Local surface geometry analysis

In 3D-SURFER, the VisGrid and LIGSITE^{csc} algorithms are integrated for characterizing the geometry of local surface regions in a query structure. VisGrid uses the visibility criterion to identify geometrical features of the query protein surface, including cavity, protrusion and flat regions. VisGrid projects the query protein onto a 3D grid with grid size of 0.9 Å [144]. Each atom is represented as a sphere with radius equal to the van der Waals radius plus the radius of a water molecule. A voxel is marked as filled by the protein if it is within the sphere of any protein atom. It then calculates the visibility for each voxel, which is the fraction of visible directions from the target voxel. A cavity is recognized as a set of grouped voxels with low visibility. A protrusion is the pocket region of the negative image of the protein.

LIGSITE^{csc} is an algorithm for automatic identification of pockets on protein surface using the Connolly surface and the degree of conservation. Similar to VisGrid, LIGSITE^{csc} also projects

the query protein onto a 3D grid, but grid size is set to 1 Å [145]. The grid points are labelled as protein, surface, or solvent according to its distance to protein atoms and the Connolly surface. LIGSITE^{csc} scans the x, y, z directions and four cubic diagonals to identify a sequence of grid points that starts and ends with surface grid points and has solvent grid points in between. This event is called a surface-solvent-surface event. A solvent grid is marked as pocket if it has more than six surface-solvent-surface events. All pocket grid points are finally clustered by their spatial proximity. The top three clusters are retained and re-ranked according to the degree of conservation of their surface residues.

5.2.4 3DZD calculation in EM-SURFER

In 3DZD calculation in 3D-SURFER, it firstly performs the surface triangulation and maps the voxels onto a 3D grid. As an EM map is represented as a 3D grid already, this pre-processing step is skipped for EM maps. Voxels are marked as 0 or 1 depending on the contour level specified. A voxel is marked as 1 if its electron density value is equal to or larger than contour level, and 0 otherwise.

For each structure in EMDB, the author has suggested a recommended contour level. Besides this author-recommended density level, a voxelization at one standard deviation of electron density, and two additional voxelizations at higher density levels, 1/3 and 2/3 of the highest density, were computed. The purpose of the additional map descriptions with one lower and two higher densities is to capture shapes at different contour levels of the molecules. Each contour level yields its own vector of 121 3DZD invariants. In total, EM-SURFER provides five options for contour shape representation: “EMDB contour”, “EMDB contour + 1/3 core”, “EMDB contour + 2/3 core”, “EMDB contour + 1/3 + 2/3 core”, and “EMDB contour + 1 std dev”. The first option (“EMDB contour”) is the recommended contour level suggested by the author of the query structure in EMDB. It is the default setting if not specified. The second (“EMDB contour + 1/3 core”) and third option (“EMDB contour + 2/3 core”) combine the 3DZD generated using the first option (121 invariants), and the 3DZD generated using 1/3*(max density) or 2/3 *(max density) as contour level (121 invariants). Those are concatenations of two sets of 3DZD (242 invariants) and able to represent regions closer to the core of the molecule. The fourth option (“EMDB contour + 1/3 core + 2/3 core”) is a combination of three sets of 3DZD (363 invariants), which captures information at different depths of the structure. The last option (“EMDB contour

+ 1 std") is a concatenation of the first option and 3DZD generated from isosurface at one standard deviation (242 invariants).

5.3 Search protein 3D structures using 3D-SURFER

5.3.1 Overview of search procedure

3D-SURFER can be freely accessed at http://kiharalab.org/3d-surfer/. The input required to run the online 3D-SURFER server is either the structure ID in PDB or a PDB format atom coordinate file. Top retrieved structures and their information, as well as local analysis of a query protein, are presented in a result web page. Besides using the search box to enter a specific structure identification (ID) code from PDB, users can also upload their own structure file. The file to upload should be in the PDB format with atom coordinates. If analysis of a structure domain is desired, the amino acid range of the domain needs to be specified in the domain range box. Whole structure is utilized to run the search if no domain range is specified.

Figure 5.1 illustrates the search page of 3D-SURFER. To enter a query protein, users can either type the PDB ID of the protein or upload a PDB-format file. The structure ID box accepts three categories of inputs: chain, domain, and complex. When entering structure ID in the box, structures in all three categories having the same ID as entered will appear in a drop-down menu, which can be scrolled down and selected.

After entering the structure ID, next step is to select a surface representation method. All atom representations utilize coordinates of all atoms in the structure to build the global surface. The main chain atom representation only includes Cα, C, and N atoms in the main chain. The choice should be made according to the purpose of the search. In general, if the purpose is to find structures with the same fold classification as CATH [26] or SCOP [154], the main chain atom representation performs best as long as the query molecule has a globular shape; the all-atom representation performs best if a query has a long tail or unstructured loop region.

Users can also select the structure database to search. Similar to selecting input structure categories, there are chain, domain, and complex template databases, and also a database that contains all three of them. Depending on the purpose of the search, users can apply the CATH filter and the length filter in the search process. To filter out similar structures, users can specify up to which CATH hierarchy level they want to use in the search. By default, no CATH filter is

applied to the search. If the length filter is enabled, only proteins whose sizes are between 0.57 and 1.75 times the size of the query protein will be retrieved in the results page. The length filter is on by default. The length filter is useful if users want to retrieve proteins with a similar size to the query protein. Since 3DZD considers shape similarity but ignores size, proteins of a similar shape but different size can be retrieved with the length filter off.

The search page of 3D-SURFER allows user to submit one protein structure as the query. To submit a batch of queries, users can go to the benchmark page in 3D-SURFER website. Users can either type a list of structure IDs or upload a structure ID list file. Most options are the same as those introduced before. One difference is that users can pick how many top results they want to obtain in the output file using the scroll-down window in step 5. The result for each query is listed in a separate file. Results for all queries submitted can also be downloaded as a single compressed zip file. Besides the batch search function, benchmark page also provides users the option to either download pre-calculated 3DZD for existing protein structures or calculate 3DZD for their own proteins of interest. To calculate 3DZD for new protein structures, users can upload the coordinate files of their proteins as a zipped file and provide their email address. 3D-SURFER then processes the uploaded structures in real-time and send the calculated 3DZD results back to the user's email address.

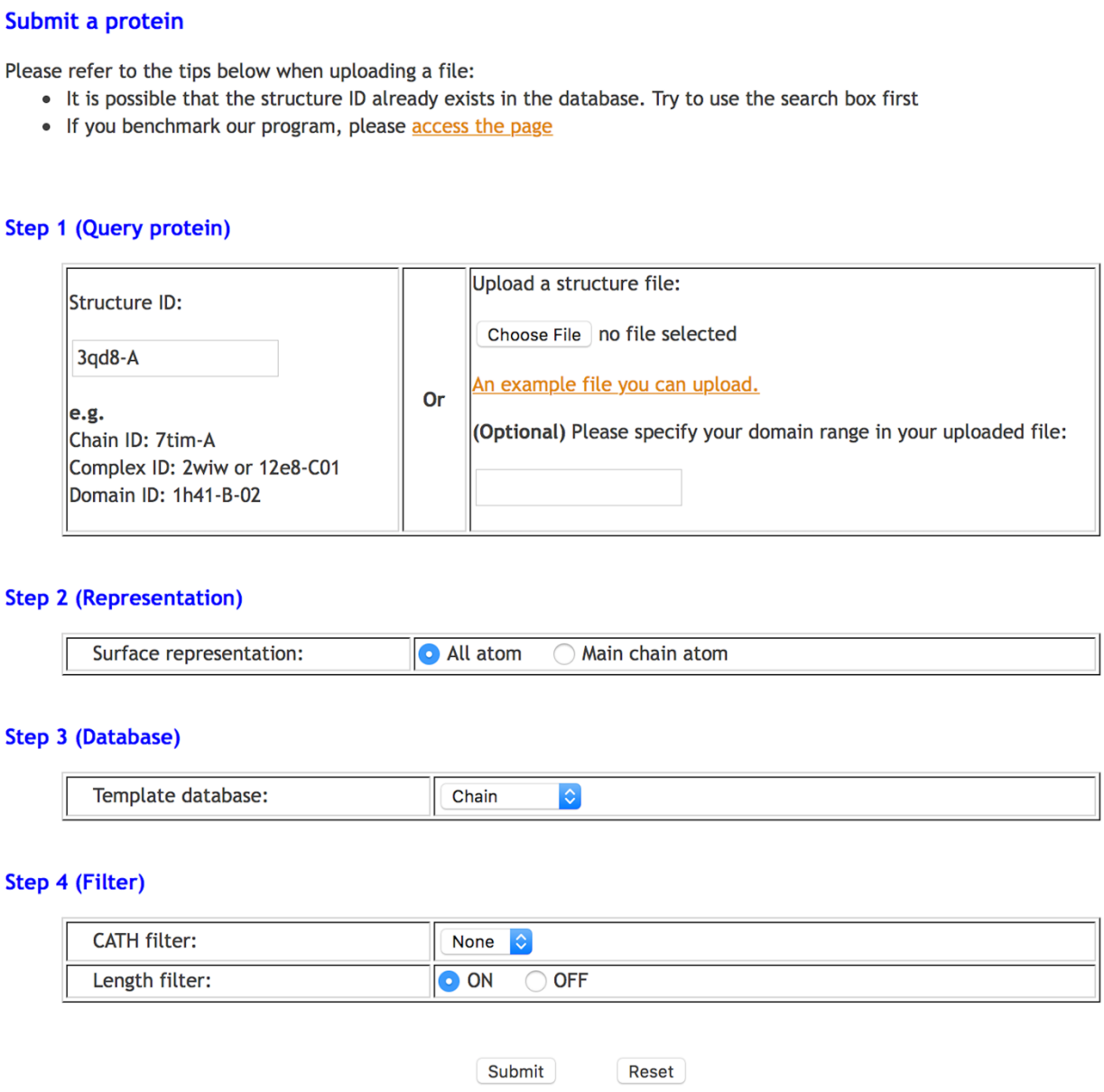


Figure 5.1: Screenshot of the job submission page in 3D-SURFER.

5.3.2 Comparison and analysis results from 3D-SURFER

The results page in 3D-SURFER shows the top 25 structures that share similar global surface shape to the query protein. We use the Euclidean distance between the 3D Zernike descriptors (3DZD) of a pair of proteins to quantify their similarity. Empirically speaking, the surface similarity between two proteins is significant if the distance is below 10. Besides surface comparison results, users can also analyze geometric features of the query protein and run structure alignment between the query protein and a specific retrieved structure. Below, we explain and discuss the search results using 3qd8-A as an example, which is chain A in ferritin BrfB from *Mycobacterium tuberculosis* [155].

At the top of the results page are shown the query ID, filters enabled in the search, the length of the query protein, and its CATH ID if available (Figure 5.2). The visualization of the query protein is generated by the JSmol applet at the top left panel. The representation of the protein can be changed using the JSmol menu by clicking the right button of the mouse. To analyze geometric features of the query protein, click the Cavity, Protrusion, or Flat button to color residues that correspond to specified geometry on the protein surface calculated by VisGrid. Red means the largest cavity or protrusion, green means the second largest, and blue means the third largest. Flat regions are colored in yellow. The VisGrid algorithm identifies local geometric features of protein surfaces using the visibility criterion. Visibility is defined as the fraction of open directions from a target position on the protein surface. Thus, a protrusion is defined as a region that has high visibility while a cavity is a region with low visibility. The surface area and volume for each cavity, protrusion, or flat region are calculated for the convex hull formed by all residues in that region. Convex hull is the smallest polygon that contains all the residues in that region. It is computed with the Qhull program [156].

Identification of surface pockets by LIGSITE^{csc} is invoked by clicking the Pocket button. If identified, the first, second, and third largest pocket residues will be colored using the same color scheme used for results from VisGrid. The surface area and volume of each pocket is also calculated with Qhull.

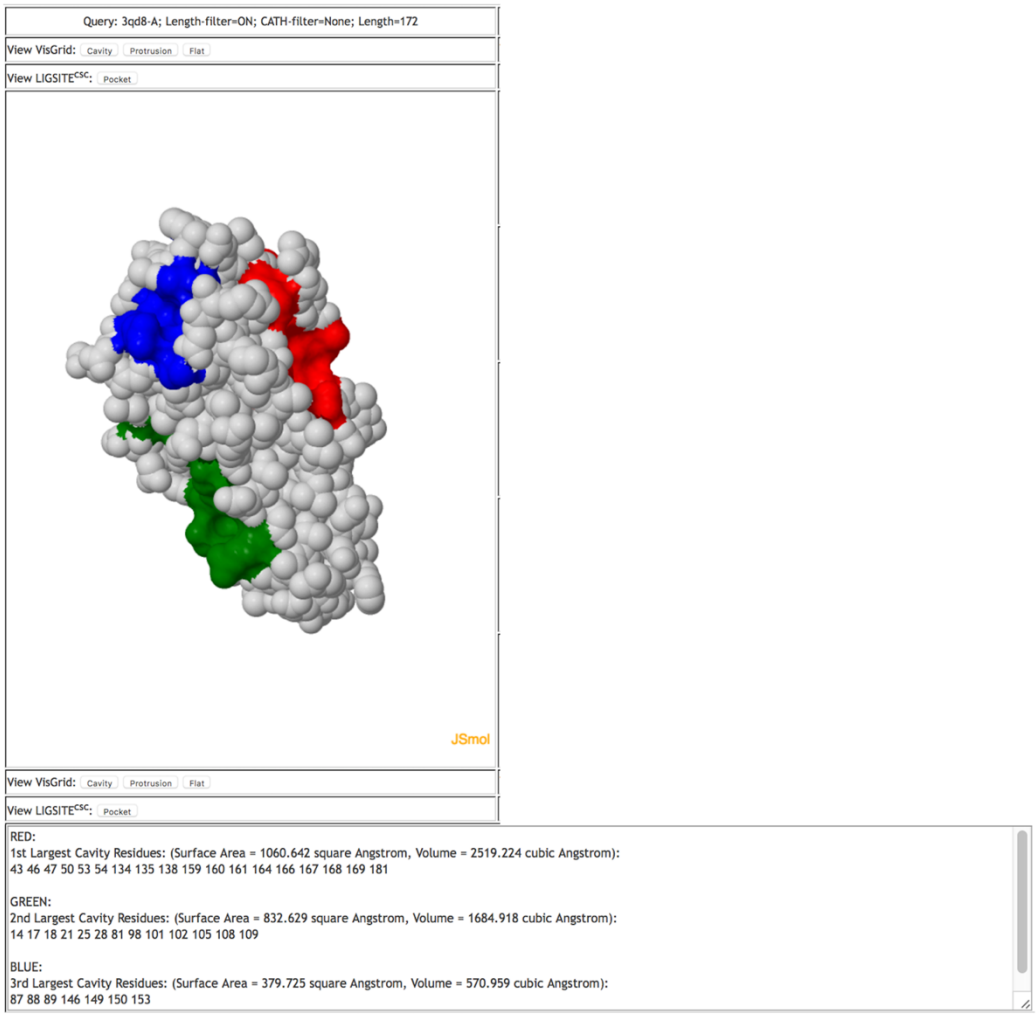


Figure 5.2: Geometric analysis of a query protein. The top part shows the query ID, filters used in the search, length of the query protein, and its CATH annotation if available. The query structure is visualized using the Jsmol applet. Users can click the Cavity, Protrusion, and Flat button to identify those regions using VisGrid. The first, second, and third largest cavities in this example are colored in red, green, and blue, respectively. The bottom panel lists residues in each cavity as well as its surface area and the volume.

Retrieved structures are sorted according to the Euclidean distance between their 3DZD and the 3DZD of the query, which is shown as "EucD" in each panel (Figure 5.3). In this example, the top 19 retrieved structures are ferritin homologs of BrfB with a Euclidean distance of 2.261 or less. By moving the mouse over the image of a protein, it will rotate 360°along the x and y axes to give a through representation of the protein. Each retrieved protein is annotated by its structure ID, length, and CATH ID if available. The structure ID is linked to the corresponding entry in the PDB Web site.

Users can also run structure alignment between query and a retrieved protein by checking the "Rmsd" box below that structure. The alignment is performed using the Combinatorial Extension (CE) program [9]. The RMSD value and the coverage (the number of aligned amino acids divided by the length of the query entry) will be displayed and a new Rmsd button will appear. By clicking the Rmsd button, the structure alignment is displayed using JSmol applet on the left panel. To visualize detailed alignment results, users can click on the RMSD result ["0.31A (0.93)" in this example] and analyze the alignment file displayed in a new pop-up window.

Results for the top 25, 50, 100, 250, 500, and 1000 retrieved structures are available by choosing from the drop-down menu next to "Top results in text format." After clicking Show, it will display structure ID, Euclidean distance, CATH ID, and length for proteins within the cutoff in a new window. At the bottom of the results page, a line chart and exact numbers of the 3DZD (121 invariants) for the query protein (Figure 5.4) are displayed.

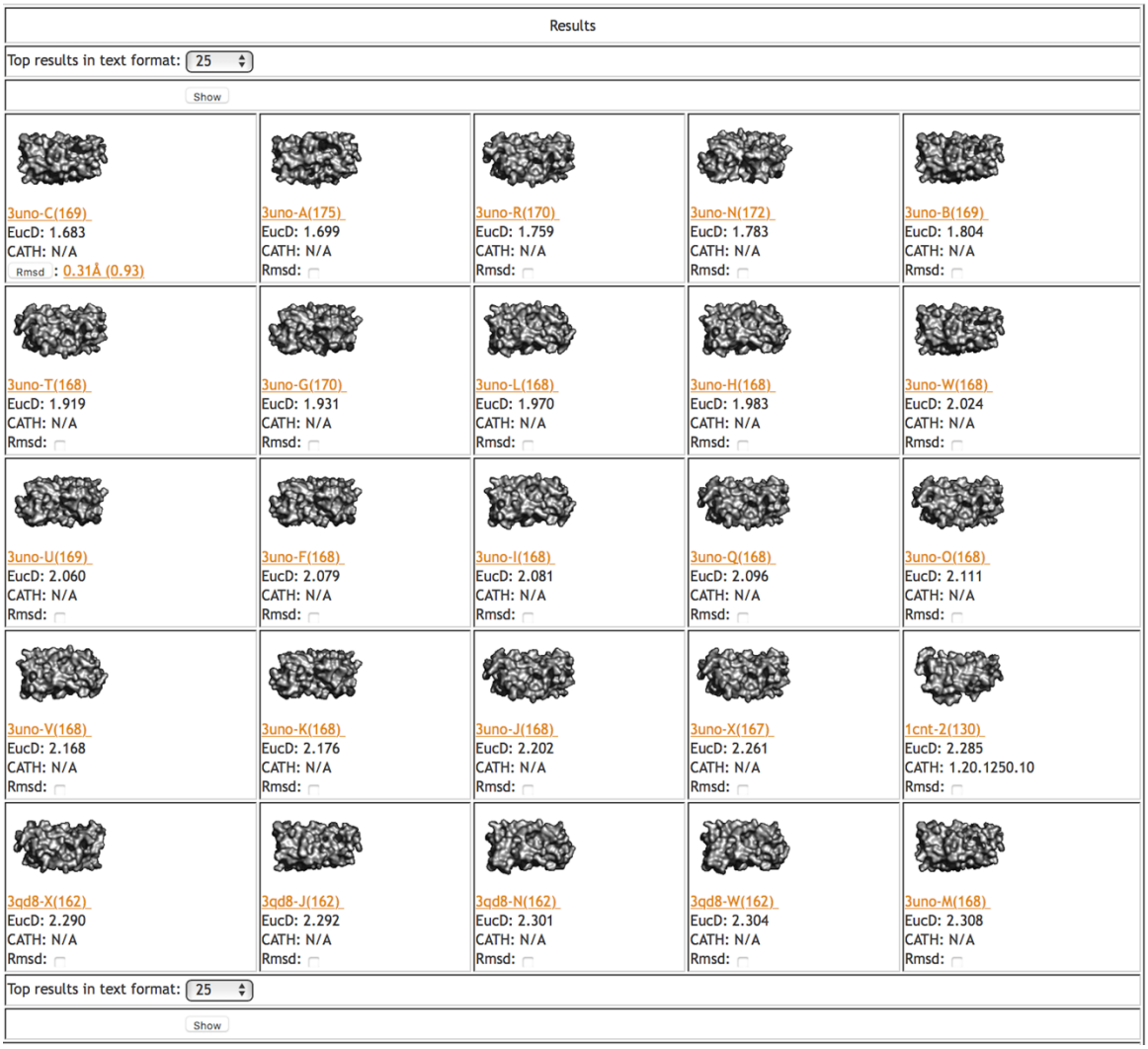


Figure 5.3: Illustration of the top 25 retrieved structures in 3D-SURFER. Each hit is displayed with its structure ID, length, Euclidean distance to the query, and CATH classification if available. To calculate root mean squared deviation (RMSD) between the query and a specific hit, users can click on the checkbox following "Rmsd." In this example, the RMSD between 3qd8-A and 3uno-C is 0.31 A, and coverage is 93%. A list of the top 20, 50, 100, 250, 500, and 1000 retrieved structures can be displayed by specifying at the drop-down menu at top and clicking the Show button.

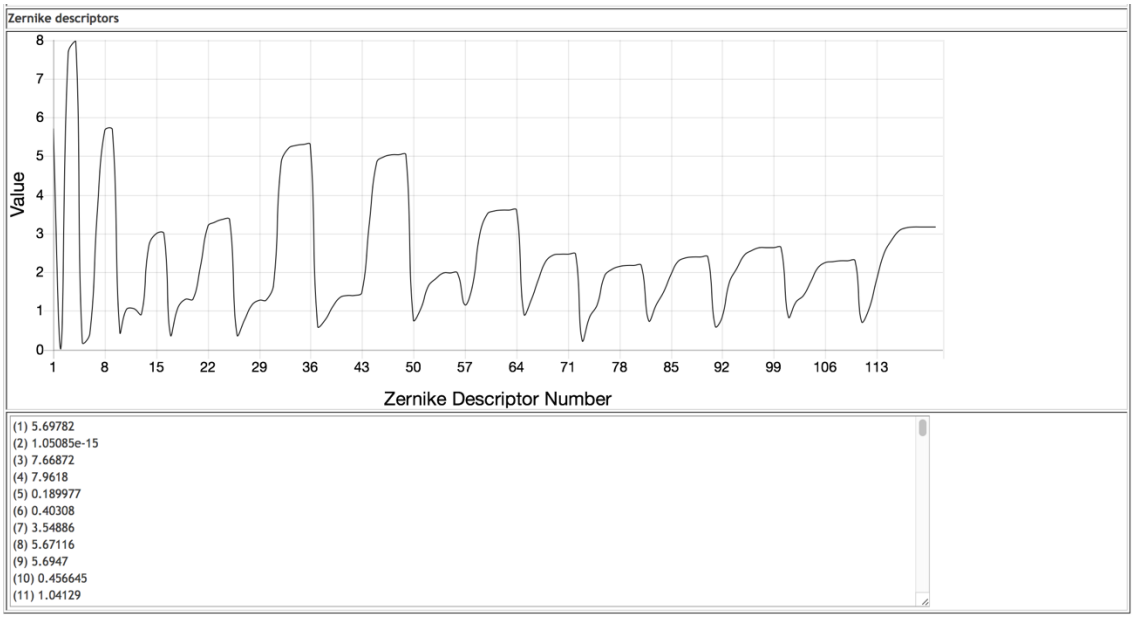


Figure 5.4: Graphic and text representation of 3DZD for a query protein. In this example, it displays 3DZD for 3qd8-A.

5.3.3 Examples of results retrieved by 3D-SURFER

Here we show two examples of search results from 3D-SURFER. As 3DZD is capable of retrieving proteins with similar global surface, it can identify functionally related proteins with low sequence identity and insignificant structure similarity. 1a31-A is the structure of DNA topoisomerase I in human [157].

Using 1a31-A as query in 3D-SURFER, the top three most similar structures (PDB ID: 1ej9-A, 1a35-A, 2b9s-A) are DNA topoisomerase I in human and Leishmania donovani with Euclidean distances less than 3 (Figure 5.5A). The default search setting was used in this example. The structure retrieved at rank 4 is DNA polymerase lambda (3hwt-A) in human. Similar to DNA topoisomerase I, DNA polymerase lambda shares a characteristic central pore that binds to DNA double strands (Figure 5.5B, 5.5C). The sequence identity between 1a31-A and 3hwt-A is low (21.6%), and the RMSD between the two structures is 9.6 Å. However, 3DZD is able to identify their overall surface similarity with a 3.41 Euclidean distance.

The second example is a search for a single chain protein query against the protein complex database. All other parameters were set to default in this search. 2ixf-A is the ATPase domain of TAP1, a subunit of ATP-binding cassette (ABC) transporter TAP in *Rattus norvegicus* ([158]; Figure 5.6). The top three retrieved structures (Figure 5.6A) are all dimeric complexes with the ABC-ATPase fold, which have a Euclidean distance less than 3.2 to the query. The first retrieved complex is the SMCcd-SMCcd homodimer from *Pyrococcus furiosus* (1xew; [159]; Figure 5.6C). An SMC protein has a catalytic ATP binding cassette (ABC) domain with the ATPase activity. It has a similar global surface to 2ixf-A as reflected in its Euclidean distance, but different secondary structure arrangements, which give an RMSD of 6.24 Å.

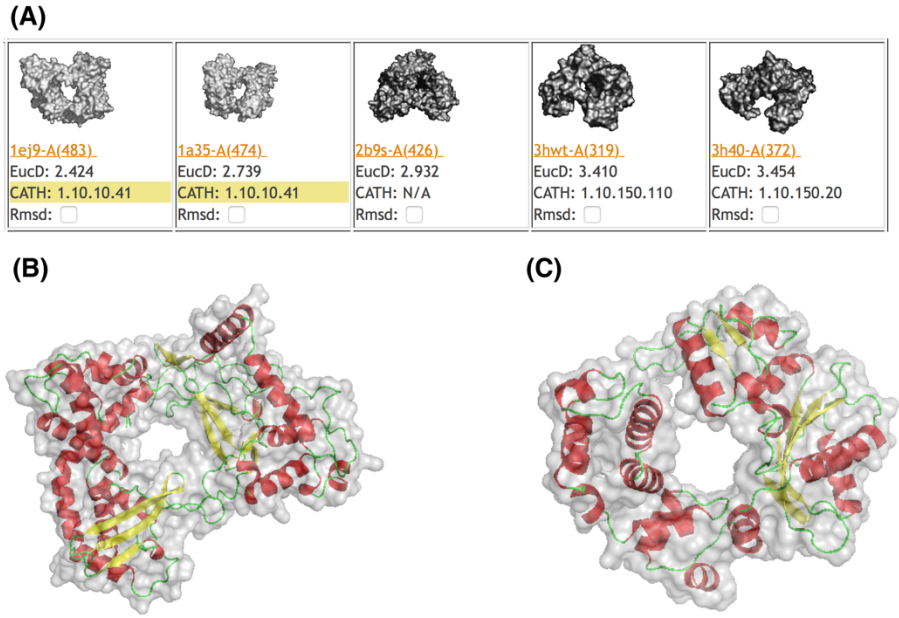


Figure 5.5: An example of a protein pair with similar global surface but different folds. (A) Part of search results for 1a31-A on 3D-SURFER. The top 5 hits are shown. (B) 1a31-A, DNA topoisomerase I from human. (C) 3hwt-A, DNA polymerase lambda from human.

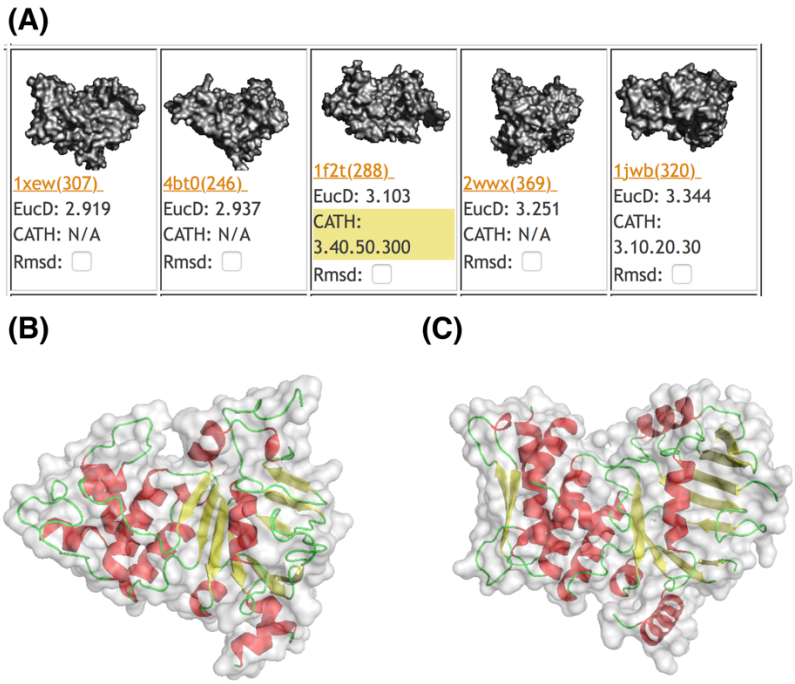


Figure 5.6: An example of search results for ATPase domain in TAP1 (PDB ID: 2ixf-A) against the complex database. (A) Part of search results for 2ixf-A by 3D-SURFER. The top 5 hits are shown. (B) Structure of query protein 2ixf-A, ATPase domain in TAP1 from *Rattus norvegicus*. (C), Structure of the top hit, 1xew, SMCcd-SMCcd homodimer from *Pyrococcus furiosus*.

5.4 Search electron microscopy maps using EM-SURFER

5.4.1 Overview of search procedure

EM-SURFER can be freely accessed at http://kiharalab.org/em-surfer/. To start a search in EM-SURFER, the only input required is either the four-digit EMDB entry ID or the electron microscopy map for the query structure. Retrieved similar structures are presented in the form of a Web page.

Figure 5.7 illustrates the search page of EM-SURFER. The first step is to specify the contour level that is used to represent the 3D shape of the query map. There are five options provided: EMDB contour, EMDB contour + 1/3 core, EMDB contour + 2/3 core, EMDB contour + 1/3 core + 2/3 core, and EMDB contour + 1 std. Second step is to provide the query information. Users can provide the four-digit EMDB entry ID or upload an EM map. When the upload option is used, the input map should be in the MAP or MRC format. Users should also specify the contour level they want to use. Next step is to choose the volume and the resolution filters. As size information of a protein is not reflected in its 3DZD descriptor, users can enable the volume filter to retrieve entries with similar volume (0.8 to 1.2 times the volume of the query). If users want to retrieve only structures within a certain resolution range, they can specify the resolution range in the resolution filter. If only a maximum or a minimum value is entered, it will be the only resolution restriction imposed. No resolution filter is applied by default.

Similar to 3D-SURFER, if users want to submit a batch of entries, they can utilize the batch mode of EM-SURFER by clicking the Benchmark button at the top panel. Users can either type in all structure IDs or upload a list file. There is an option to specify the number of top results shown in the final output. By choosing from 10, 20, or 30 in the drop-down menu, users will get a corresponding number of retrieved structures for each query. The output for each query submitted can be downloaded separately or as a single compressed zip file. The benchmark page also allows users to download pre-calculated 3DZDs for existing EM maps.

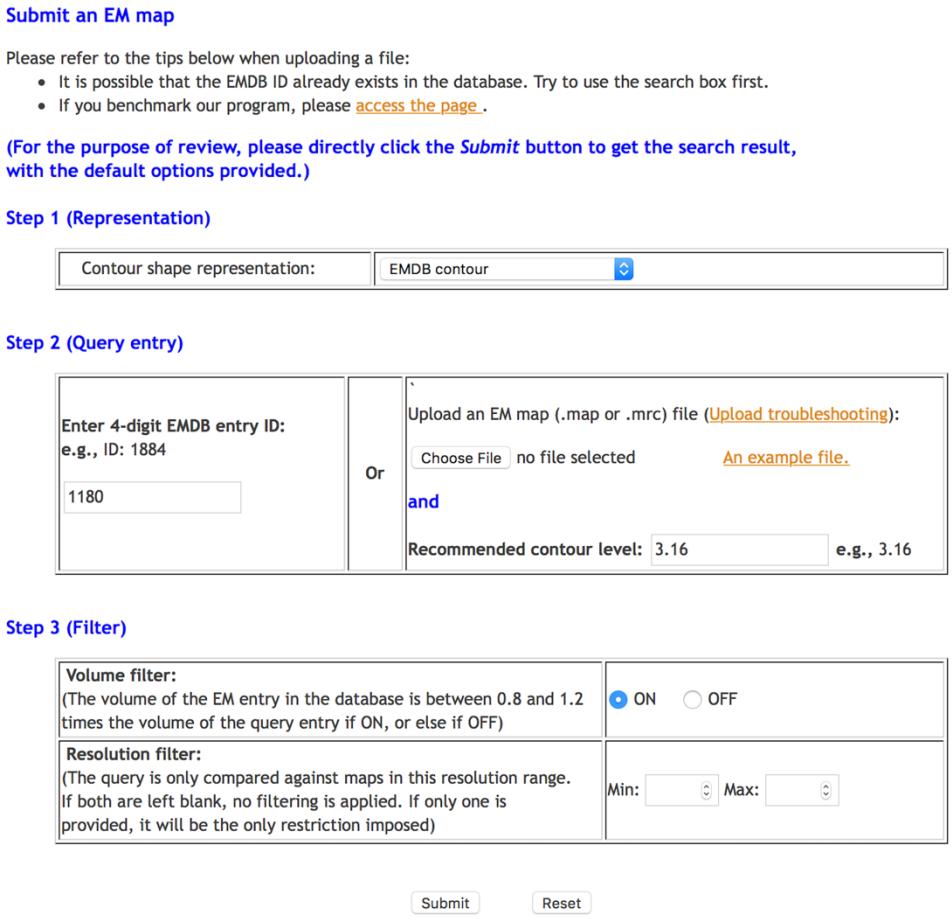


Figure 5.7: Screenshot of the job submission page in EM-SURFER.

5.4.2 Comparison and analysis results from EM-SURFER

A search result of EM-SURFER is displayed in a similar layout as in 3D-SURFER. The results page displays the top 20 structures that share a similar global isosurface shape to the input map. Similar to 3D-SURFER, global surface similarity between two maps is quantified by the Euclidean distance of their 3DZDs. The smaller the distance, the more similar the two EM maps are. Empirically, two biomolecules in EM maps are biologically related if the distance is smaller than 8.0. Below, we explain and discuss search results of the EM-SURFER server using EMD-1180 as a query, which is a GroEL-ATP-GroES complex [160], as shown in Figure 5.8.

The top panel in the results page displays the EMDB ID of the query, its description, and a figure showing its overall structure, which is taken from EMDB. Graphic and text forms of the 3DZD of the query protein are placed next to the structure image. To download text results for the top 50 structures, users can click the "Download text results here" button located in the first row. In the text results, retrieved structures are ranked by Euclidean distances of their 3DZD to the query. Resolution information is also provided, if available, in the map information in EMDB.

The search results panel shows the top 20 structures with the most similar global isosurface shape to the query (Figure 5.8). In this example, the top 13 retrieved EM maps are all GroELs, which indicate that the search is successful in identifying biologically relevant maps to the query from EMDB. Retrieved structures are ranked by the Euclidean distance of their 3DZDs to that of the query (shown after "EucD:"). Smaller Euclidean distance indicates that the surface of two structures is more similar. Each hit is displayed with its EMDB ID, a short description, the ratio of volume to query, and its resolution. The EMDB ID of the structure is linked to its corresponding entry in the EMDB Web site to allow users to obtain more detailed structure information. By clicking on the image of a retrieved structure, users can start a new search from the clicked entry. Default filter settings are used in this new search.

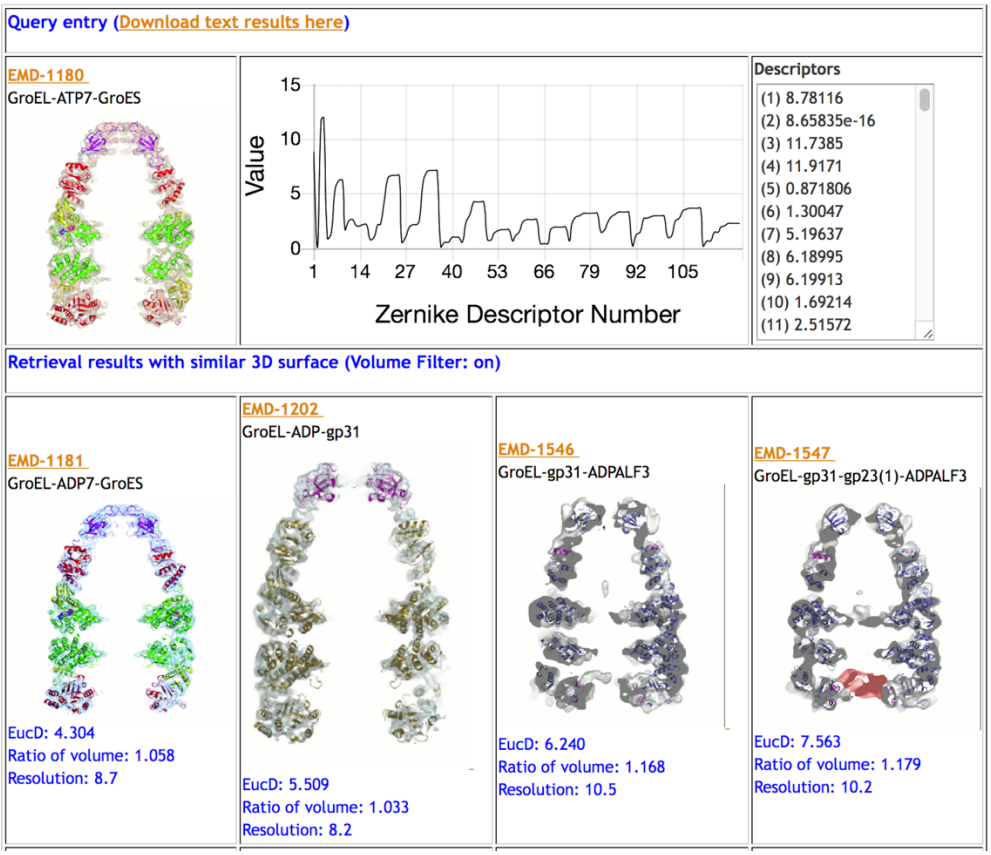


Figure 5.8: An example of the EM-SURFER results page for query EMD-1180. The top left panel displays an image of the query structure, or the filename if a user's EM map is uploaded. The graph and the text depiction of the 3DZD for the query are shown next to the map image. The search results panel shows top 20 hits with their EMDB ID, a short description, structure image, and detailed values of their Euclidean distance, volume ratio, and resolution. Here we only show part of the results page (top 4 hits) for EMD-1180.

5.4.3 Examples of results retrieved by EM-SURFER

Here we show another search result by EM-SURFER, using EMD-1226 as the query (Figure 5.9A). For this search, the author-recommended contour level was used and the volume filter was on. The query is yeast heat shock protein Hsp26 in compact form [161]. Six different three-dimensional structures of wild-type Hsp26 and modified forms were solved in the original paper (EMD-1221, 1226 to 1230). All of them are retrieved at the top in the search results. There are two distinct forms of Hsp26, one in a compact and another in an expanded form. The internal organization of those two forms is different, but their external diameter is only 4% different. As both forms share similar global surface shape, using EMD-1226 as the query, all other Hsp26 proteins are ranked top 1 to top 5 in the EM-SURFER results. For comparison, in Figure 5.9B we show the search results from the Omokage server, another tool for searching PDB and EMDB [6]. It uses a combination of incremental distance rank (iDR) profile and the principal component analysis (PCA) profile to characterize shape similarity. Although five other Hsp26 proteins are retrieved among top nine hits on the Omokage server, they are separated by some other proteins, like bacteriophage EL chaperonin. Thus, in this case, EM-SURFER had a better performance than Omokage.

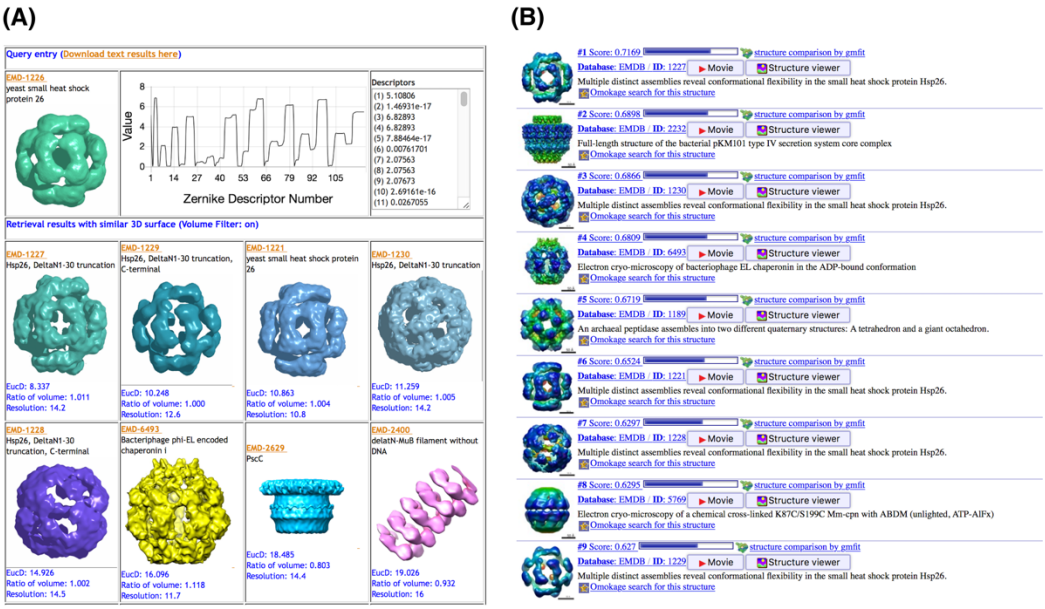


Figure 5.9: Search results for a yeast heat shock protein Hsp26, EMD-1226. (A) Search results in EM-SURFER. The EMD-1227, 1229, 1221, 1230, and 1228 are EM maps for Hsp26. (B) Search results in the Omokage 1230, 1221, 1228, and 1229 are EM maps for Hsp26.

5.5 Discussion

With the rapid growth in the number of solved protein structures stored in the Protein Data Bank (PDB) and the Electron Microscopy Data Bank (EMDB), it is essential to develop tools to perform real-time structure similarity searches against the entire structure database. Since conventional structure alignment methods need to sample different orientations of proteins in the three-dimensional space, they are time consuming and unsuitable for rapid, real-time database searches. To this end, we have developed 3D-SURFER and EM-SURFER, which utilize 3D Zernike descriptors (3DZD) to conduct high-throughput protein structure comparison, visualization, and analysis. Taking an atomic structure or an electron microscopy map of a protein or a protein complex as input, the 3DZD of a query protein is computed and compared with the 3DZD of all other proteins in PDB or EMDB. In addition, local geometrical characteristics of a query protein can be analyzed using VisGrid and LIGSITE^{csc} in 3D-SURFER.

Although both 3D-SURFER and EM-SURFER have several checks and repair mechanisms to ensure correct processing of structure ID and uploaded files, there can still be cases where users have troubles obtaining the search results. Thus, here we also provide some troubleshooting tips for potential problems. When a query ID is not recognized in the structure ID box in 3D-SURFER, it is possible that the structure entered is obsolete in PDB or the structure sequence is too short. PDB entries can be made obsolete following an author's request to PDB when better experimental data has been collected or a better interpretation of existing data has been produced. Obsolete entries reported in PDB are removed from the 3D-SURFER database, making those structure IDs unrecognizable in the search. In this case, users can go to the PDB Web page, identify a superseding entry for that obsolete structure, and input its successor into the search. Another possibility is that the input structure contains less than 10 residues. Those short structures are also removed in 3D-SURFER, as their structure information is limited. When an error occurs in the uploading process in EM-SURFER, it may come from a network connectivity issue, incorrect format in map file, etc.

CHAPTER 6. DISCUSSION AND SUMMARY

6.1 Remaining challenges

We have represented the protein global surface shape using 3DZD and analyzed the protein shape universe in Chapter 2. 3DZD treats proteins as rigid bodies and ignores the protein structure flexibility. Proteins are flexible molecules. Although the native structure of a protein is stabilized by physical interactions of atoms, the structure still admits flexible motions. Motions include those of side-chains, and some parts of main-chains, especially regions that do not form the secondary structures, which are often called loop regions. In many cases, the flexibility of proteins plays an important or essential role in the biological functions of the proteins. When we built the single-chain and complex dataset, we applied the sequence similarity cutoff of 25% and thus only one conformation for each protein was retained in the non-redundant dataset. To better reflect how the proteins exist in nature, it would be helpful to take different conformations of the same protein into consideration in the analysis of protein structure universe.

3D-SURFER website was originally developed in 2009 to offer the surface comparison and analysis of protein atomic structures [5]. With recent updates of operating system, hardware, and software, it becomes more and more time-consuming and difficult to maintain existing scripts in 3D-SURFER website. We recently added the support for mmCIF format, which became the standard PDB archive format in 2014 and imposed no limitations for the number of atoms, residues or chains in a single PDB entry. As previous scripts only accept PDB format coordinate files, it took a long time to rewrite the scripts for every data processing step and benchmark the calculation results to enable the switch to mmCIF. As 3D-SURFER integrates many other programs for atomic structure processing and analysis, it also requires periodic check to make sure those programs are working properly. RMSD value is calculated from the structure-based alignment obtained from CE program [9]. Jsmol is used to visualize the protein structure and geometric features. VisGrid [144] and LIGSITE^{csc} [145] characterizes the local geometric features of a query protein. Pymol generates the images and animations for all structures in 3D-SURFER database. Maintenance of 3D-SURFER website also requires special attention to all those programs, which are developed by different authors at different times and provide different functions. One potential solution is to

package up all programs, libraries, and other dependencies in a container, which can significantly save the time for periodic check after system and software updates.

6.2 Future work

We have applied principal component analysis to project 3DZD from 121-dimension to three-dimensional space in Chapter 2. There are several other directions we can further explore for the projection space. Top three PCs explained 52.64% and 47.76% of the total variation in the single-chain and complex datasets, respectively. We can further look into the factor loadings in each of top three PCs to understand the contribution from each component in 3DZD vector. One possible direction is to replace 3DZD to x, y, and z coordinates in this projection space and analyze the retrieval performance of protein atomic structures. The second possible direction is to use lower order in 3DZD calculation. Currently, the order n is set to 20 which generates 121 invariants in the 3DZD vector. As eccentricity is used to quantify the protein shape and it does not pay much attention to small changes in protein surface, a coarser surface shape representation can be utilized to extract the shape information and construct the projection space. We can then compare the projection space built with original 3DZD to the space built with lower order 3DZD and analyze the differences in protein distributions. Chapter 2 analyzes protein atomic structures in PDB, but similar analysis can also be conducted for EM maps in EMDB to present an overview of the shape universe of EM maps.

To quantify the similarity of protein structures, Chapter 4 and Chapter 5 have calculated the Euclidean distance of 3DZD vectors. As shown in Chapter 4, if we feed 3DZD as the input to a neural network and obtain a similarity score for a pair of protein structures, the retrieval performance is always better than using Euclidean distance directly. This suggests that neural network is a useful tool to integrate into our protein shape comparison approach. Previously, Furuya and Ohbuchi have combined low-level local geometrical features with Deep Neural Network to learn part-in-whole relation of 3D shapes [162]. Gainza *et al.* have developed a framework named MaSIF, which is based on a geometric deep learning method to capture fingerprints for different applications [163]. It decomposes a surface into overlapping patches and calculates geometry and chemistry features for each point within a patch. MaSIF learns to embed the features into a numerical vector descriptor depending on the application of interest. We can

also develop a new method with the integration of neural network to further improve the retrieval performance.

Both 3D-SURFER and EM-SURFER utilize 3DZD to describe the shape information of protein structures. Similar to what is offered in Omokage search, we can develop a new search service to enable the search of both atomic structures and EM maps given a query structure. But from some previous explorations, direct usage of Euclidean distance of 3DZD vectors does not provide a good performance. Atomic structures in PDB have much higher resolution compared to most EM maps in EMDB. EM maps are also much noisier compared to atomic structures. As a result, we have observed a big Euclidean distance even if the atomic structure and the EM map is from the same protein. One potential solution to convert atomic structures to EM maps at lower resolutions and then calculate 3DZD for the simulated map instead of the original atomic structure. We can explore multiple resolution level and contour level combinations to determine the optimal parameters to use.

6.3 Outlook

3D shape matching is an important and active research area in computer vision. The remarkable advances in the deep learning architectures on 2D data have led to significant improvements in classification [164], segmentation [165, 166], and detection and localization. Similar approaches have been applied into 3D shape matching, which may also help improve the comparison of protein 3D structures. Analysis of protein structure also has strong implications for protein design. We can dive further into the protein shape universe and reveal protein shapes that are difficult to construct with proteins. We may also design proteins with certain properties based on their shape and structure similarity to existing proteins.

REFERENCES

1. Berg JM, Tymoczko JL, Stryer L. Biochemistry. 5th ed. New York: W.H. Freeman; 2002. xxxviii, 974, 76 p. p.
2. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular biology of the cell. 4th ed. New York: Garland Science; 2002. xxxiv, 1548 pages p.
3. Orengo CA, Thornton JM. Protein families and their evolution-a structural perspective. Annu Rev Biochem. 2005;74:867-900.
4. Xiong Y, Esquivel-Rodriguez J, Sael L, Kihara D. 3D-SURFER 2.0: web platform for real-time search and characterization of protein surfaces. Methods Mol Biol. 2014;1137:105-17.
5. La D, Esquivel-Rodriguez J, Venkatraman V, Li B, Sael L, Ueng S, et al. 3D-SURFER: software for high-throughput protein surface comparison and analysis. Bioinformatics. 2009;25(21):2843-4.
6. Suzuki H, Kawabata T, Nakamura H. Omokage search: shape similarity search service for biomolecular structures in both the PDB and EMDB. Bioinformatics. 2016;32(4):619-20.
7. Esquivel-Rodriguez J, Xiong Y, Han X, Guang S, Christoffer C, Kihara D. Navigating 3D electron microscopy maps with EM-SURFER. BMC bioinformatics. 2015;16:181.
8. Mukherjee S, Zhang Y. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. Nucleic Acids Res. 2009;37(11):e83.
9. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 1998;11(9):739-47.
10. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10.
11. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7):2302-9.
12. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443-53.
13. Hawkins T, Chitale M, Luban S, Kihara D. PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. Proteins-Structure Function and Bioinformatics. 2009;74(3):566-82.
14. Jain A, Kihara D. Phylo-PFP: improved automated protein function prediction using phylogenetic distance of distantly related sequences. Bioinformatics. 2019;35(5):753-9.
15. Lan L, Djuric N, Guo Y, Vucetic S. MS-kNN: protein function prediction by integrating multiple data sources. BMC Bioinformatics. 2013;14 Suppl 3:S8.
16. Wang Z, Zhao CG, Wang YH, Sun Z, Wang N. PANDA: Protein function prediction using domain architecture and affinity propagation. Sci Rep-Uk. 2018;8.
17. Rentzsch R, Orengo CA. Protein function prediction using domain families. Bmc Bioinformatics. 2013;14.
18. Piovesan D, Giollo M, Leonardi E, Ferrari C, Tosatto SCE. INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity. Nucleic Acids Research. 2015;43(W1):W134-W40.
19. Griffiths AJF. Modern genetic analysis.

20. Blake CC, Koenig DF, Mair GA, North AC, Phillips DC, Sarma VR. Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution. Nature. 1965;206(4986):757-61.
21. Kartha G, Bello J, Harker D. Tertiary structure of ribonuclease. Nature. 1967;213(5079):862-5.
22. Rossmann MG, Moras D, Olsen KW. Chemical and biological evolution of nucleotide-binding protein. Nature. 1974;250(463):194-9.
23. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 2004;32(Database issue):D226-9.
24. 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):536-40.
25. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2014;42(Database issue):D304-9.
26. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, et al. CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 2015;43(Database issue):D376-81.
27. Holm L, Sander C. The FSSP database of structurally aligned protein fold families. Nucleic Acids Res. 1994;22(17):3600-9.
28. Hadley C, Jones DT. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure. 1999;7(9):1099-112.
29. Campbell ID. Timeline: the march of structural biology. Nat Rev Mol Cell Biol. 2002;3(5):377-81.
30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235-42.
31. Magnusson U, Salopek-Sondi B, Luck LA, Mowbray SL. X-ray structures of the leucine-binding protein illustrate conformational changes and the basis of ligand specificity. J Biol Chem. 2004;279(10):8747-52.
32. Bragg WL, Thomson JJ. The diffraction of short electromagnetic waves by a crystal. P Camb Philos Soc. 1914;17:43-57.
33. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature. 1958;181(4610):662-6.
34. Brink C, Hodgkin DC, Lindsey J, Pickworth J, Robertson JR, White JG. X-ray crystallographic evidence on the structure of vitamin B12. Nature. 1954;174(4443):1169-71.
35. Harding M, Hodgkin D, Cole S, Kennedy A, Oconnor A, Rimmer B, et al. X-Ray Studies on the Structure of Insulin. Acta Crystallogr. 1960;13(12):1056-.
36. Deisenhofer J, Epp O, Miki K, Huber R, Michel H. Structure of the Protein Subunits in the Photosynthetic Reaction Center of Rhodopseudomonas-Viridis at 3a Resolution. Nature. 1985;318(6047):618-24.
37. Allen JP, Feher G, Yeates TO, Komiya H, Rees DC. Structure of the Reaction Center from Rhodobacter-Sphaeroides R-26 - the Cofactors .1. P Natl Acad Sci USA. 1987;84(16):5730-4.

38. Bloch F, Hansen WW, Packard M. The Nuclear Induction Experiment. Phys Rev. 1946;70(7-8):474-85.
39. Pound RV, Purcell EM. Measurement of Magnetic Resonance Absorption by Nuclear Moments in a Solid. Phys Rev. 1946;69(11-1):681-.
40. Proctor WG, Yu FC. The Dependence of a Nuclear Magnetic Resonance Frequency Upon Chemical Compound. Phys Rev. 1950;77(5):717-.
41. Saunders M, Wishnia A, Kirkwood JG. The Nuclear Magnetic Resonance Spectrum of Ribonuclease. J Am Chem Soc. 1957;79(12):3289-90.
42. Ernst RR, Anderson WA. Application of Fourier Transform Spectroscopy to Magnetic Resonance. Rev Sci Instrum. 1966;37(1):93-+.
43. Aue WP, Bartholdi E, Ernst RR. 2-Dimensional Spectroscopy - Application to Nuclear Magnetic-Resonance. J Chem Phys. 1976;64(5):2229-46.
44. Hacker C, Cai X, Kegler C, Zhao L, Weickhmann AK, Wurm JP, et al. Structure-based redesign of docking domain interactions modulates the product spectrum of a rhabdopeptide-synthesizing NRPS. Nat Commun. 2018;9(1):4366.
45. Zhao J, Beyrakhova K, Liu Y, Alvarez CP, Bueler SA, Xu L, et al. Molecular basis for the binding and modulation of V-ATPase by a bacterial effector protein. PLoS Pathog. 2017;13(6):e1006394.
46. Adrian M, Dubochet J, Lepault J, McDowall AW. Cryo-electron microscopy of viruses. Nature. 1984;308(5954):32-6.
47. Dubochet J, Adrian M, Chang JJ, Homo JC, Lepault J, McDowall AW, et al. Cryo-electron microscopy of vitrified specimens. Q Rev Biophys. 1988;21(2):129-228.
48. Kuhlbrandt W. Biochemistry. The resolution revolution. Science. 2014;343(6178):1443-4.
49. Murata K, Wolf M. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. Biochim Biophys Acta Gen Subj. 2018;1862(2):324-34.
50. Vinothkumar KR, Zhu JP, Hirst J. Architecture of mammalian respiratory complex I. Nature. 2014;515(7525):80-+.
51. Bai XC, Yan CY, Yang GH, Lu PL, Ma D, Sun LF, et al. An atomic structure of human gamma-secretase. Nature. 2015;525(7568):212-+.
52. Grigorieff N, Harrison SC. Near-atomic resolution reconstructions of icosahedral viruses from electron cryo-microscopy. Curr Opin Struc Biol. 2011;21(2):265-73.
53. Henderson R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. Q Rev Biophys. 1995;28(2):171-93.
54. Milo R. What is the total number of protein molecules per cell volume? A call to rethink some published values. Bioessays. 2013;35(12):1050-5.
55. Ettinger A, Wittmann T. Fluorescence live cell imaging. Methods Cell Biol. 2014;123:77-94.
56. Chen YX, Hrabe T, Pfeffer S, Pauly O, Mateus D, Navab N, et al. Detection and Identification of Macromolecular Complexes in Cryo-Electron Tomograms Using Support Vector Machines. 2012 9th Ieee International Symposium on Biomedical Imaging (Isbi). 2012:1373-6.
57. Che CQ, Lin RG, Zeng XR, Elmaaroufi K, Galeotti J, Xu M. Improved deep learning-based macromolecules structure classification from electron cryo-tomograms. Mach Vision Appl. 2018;29(8):1227-36.

58. Baldwin PR, Tan YZ, Eng ET, Rice WJ, Noble AJ, Negro CJ, et al. Big data in cryoEM: automated collection, processing and accessibility of EM data. Curr Opin Microbiol. 2018;43:1-8.
59. Wang F, Gong H, Liu G, Li M, Yan C, Xia T, et al. DeepPicker: A deep learning approach for fully automated particle picking in cryo-EM. J Struct Biol. 2016;195(3):325-36.
60. Chen M, Dai W, Sun SY, Jonasch D, He CY, Schmid MF, et al. Convolutional neural networks for automated annotation of cellular cryo-electron tomograms. Nat Methods. 2017;14(10):983-5.
61. Eltsov M, Dube N, Yu Z, Pasakarnis L, Haselmann-Weiss U, Brunner D, et al. Quantitative analysis of cytoskeletal reorganization during epithelial tissue sealing by large-volume electron tomography. Nat Cell Biol. 2015;17(5):605-14.
62. Lawson CL, Patwardhan A, Baker ML, Hryc C, Garcia ES, Hudson BP, et al. EMDataBank unified data resource for 3DEM. Nucleic Acids Res. 2016;44(D1):D396-403.
63. Hasegawa H, Holm L. Advances and pitfalls of protein structural alignment. Curr Opin Struct Biol. 2009;19(3):341-8.
64. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol. 1993;233(1):123-38.
65. Ye Y, Godzik A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics. 2003;19 Suppl 2:ii246-55.
66. Han X, Wei Q, Kihara D. Protein 3D Structure and Electron Microscopy Map Retrieval Using 3D-SURFER2.0 and EM-SURFER. Curr Protoc. 2017;60:3 14 1-3 5.
67. Zhang J, Baciu G, Zheng D, Liang C, Li G, Hu J. IDSS: a novel representation for woven fabrics. IEEE Trans Vis Comput Graph. 2013;19(3):420-32.
68. Liu YS, Li Q, Zheng GQ, Ramani K, Benjamin W. Using diffusion distances for flexible molecular shape comparison. Bmc Bioinformatics. 2010;11.
69. Wang HW, Chu CH, Wang WC, Pai TW. A local average distance descriptor for flexible protein structure comparison. Bmc Bioinformatics. 2014;15.
70. Joseph AP, Lagerstedt I, Patwardhan A, Topf M, Winn M. Improved metrics for comparing structures of macromolecular assemblies determined by 3D electron-microscopy. J Struct Biol. 2017;199(1):12-26.
71. Shatsky M, Hall RJ, Brenner SE, Glaeser RM. A method for the alignment of heterogeneous macromolecules from electron microscopy. Journal of Structural Biology. 2009;166(1):67-78.
72. Vasishtan D, Topf M. Scoring functions for cryoEM density fitting. Journal of Structural Biology. 2011;174(2):333-43.
73. Chen ZZ, Husz ZL, Wallace I, Wallace AM. Video object tracking based on a chamfer distance transform. Ieee Image Proc. 2007:1485-+.
74. Ceulemans H, Russell RB. Fast fitting of atomic structures to low-resolution electron density maps by surface overlap maximization. Journal of Molecular Biology. 2004;338(4):783-93.
75. Osadchy M, Kolodny R. Maps of protein structure space reveal a fundamental relationship between protein structure and function. Proc Natl Acad Sci U S A. 2011;108(30):12301-6.

76. Hou J, Jun SR, Zhang C, Kim SH. Global mapping of the protein structure space and application in structure-based inference of protein function. Proc Natl Acad Sci U S A. 2005;102(10):3651-6.
77. Hou J, Sims GE, Zhang C, Kim SH. A global representation of the protein fold space. Proc Natl Acad Sci U S A. 2003;100(5):2386-90.
78. King NP, Sheffler W, Sawaya MR, Vollmar BS, Sumida JP, Andre I, et al. Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science. 2012;336(6085):1171-4.
79. Lin YR, Koga N, Tatsumi-Koga R, Liu G, Clouser AF, Montelione GT, et al. Control over overall shape and size in de novo designed proteins. Proc Natl Acad Sci U S A. 2015;112(40):E5478-85.
80. Bhardwaj G, Mulligan VK, Bahl CD, Gilmore JM, Harvey PJ, Cheneval O, et al. Accurate de novo design of hyperstable constrained peptides. Nature. 2016;538(7625):329-35.
81. Han X, Sit A, Christoffer C, Chen S, Kihara D. A global map of the protein shape universe. PLoS Comput Biol. 2019;15(4):e1006969.
82. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, et al. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science. 2011;332(6031):816-21.
83. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, et al. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic acids research. 2017;45(D1):D289-D95.
84. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. JMolBiol. 1995;247(4):536-40.
85. Schaeffer RD, Liao Y, Cheng H, Grishin NV. ECOD: new developments in the evolutionary classification of domains. Nucleic acids research. 2017;45(D1):D296-D302.
86. Chothia C. Proteins. One thousand families for the molecular biologist. Nature. 1992;357(6379):543-4.
87. Liu X, Fan K, Wang W. The number of protein folds and their distribution over families in nature. Proteins. 2004;54(3):491-9.
88. Magner A, Szpankowski W, Kihara D. On the origin of protein superfamilies and superfolds. Scientific reports. 2015;5:8166.
89. Abeln S, Deane CM. Fold usage on genomes and protein fold evolution. Proteins. 2005;60(4):690-700.
90. Qian J, Luscombe NM, Gerstein M. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. Journal of molecular biology. 2001;313(4):673-81.
91. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J. 1986;5(4):823-6.
92. Finkelstein AV, Ptitsyn OB. Why do globular proteins fit the limited set of folding patterns? Progress in biophysics and molecular biology. 1987;50(3):171-90.
93. Efimov AV. Structural trees for protein superfamilies. Proteins. 1997;28(2):241-60.
94. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science. 1991;253(5016):164-70.

95. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. Journal of molecular biology. 1997;268(1):209-25.
96. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. Journal of molecular biology. 1993;233(1):123.
97. Lucic V, Rigort A, Baumeister W. Cryo-electron tomography: the challenge of doing structural biology in situ. J Cell Biol. 2013;202(3):407-19.
98. Kuhlbrandt W. Cryo-EM enters a new era. eLife. 2014;3:e03678.
99. Sael L, Li B, La D, Fang Y, Ramani K, Rustamov R, et al. Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins. 2008;72(4):1259-73.
100. Kihara D, Sael L, Chikhi R, Esquivel-Rodriguez J. Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking. Curr Protein Pept Sci. 2011;12(6):520-30.
101. Han X, Wei Q, Kihara D. Protein 3D Structure and Electron Microscopy Map Retrieval Using 3D-SURFER2.0 and EM-SURFER. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2017;60:3.14.1-3..5.
102. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics. 2003;19(12):1589-91.
103. The UniProt C. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):D158-D69.
104. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol. 2007;372(3):774-97.
105. Sanner MF, Olson AJ, Spehner JC. Reduced surface: an efficient way to compute molecular surfaces. Biopolymers. 1996;38(3):305-20.
106. Connolly ML. The molecular surface package. Journal of molecular graphics. 1993;11(2):139-41.
107. Canterakis N. 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition. Proc11th Scandinavian Conference on Image Analysis. 1999:85.
108. Sael L, Kihara D. Improved protein surface comparison and application to low-resolution protein structure data. BMC bioinformatics. 2010;11 Suppl 11:S2.
109. Sael L, La D, Li B, Rustamov R, Kihara D. Rapid comparison of properties on protein surface. Proteins. 2008;73(1):1-10.
110. Ejlali N, Faghihi MR, Sadeghi M. Bayesian comparison of protein structures using partial Procrustes distance. Statistical applications in genetics and molecular biology. 2017;16(4):243-57.
111. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7):2302.
112. Chen CR, Makhatadze GI. ProteinVolume: calculating molecular van der Waals and void volumes in proteins. BMC bioinformatics. 2015;16:101.
113. Oliphant TE. Python for Scientific Computing. Computing in Science & Engineering. 2007;9:10-20.
114. Li B, Turuvekere S, Agrawal M, La D, Ramani K, Kihara D. Characterization of local geometry of protein surfaces with the visibility criterion. Proteins. 2008;71(2):670-83.
115. Xu D, Zhang Y. Generating triangulated macromolecular surfaces by Euclidean Distance Transform. PLoS One. 2009;4(12):e8140.

116. Ladner JE, Pan M, Hurwitz J, Kelman Z. Crystal structures of two active proliferating cell nuclear antigens (PCNAs) encoded by Thermococcus kodakaraensis. Proc Natl Acad Sci U S A. 2011;108(7):2711-6.
117. Skolnick J, Arakaki AK, Lee SY, Brylinski M. The continuity of protein structure space is an intrinsic property of proteins. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(37):15690-5.
118. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG. SCOP2 prototype: a new approach to protein structure mining. Nucleic acids research. 2014;42(Database issue):D310-4.
119. Gerstein M. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. Journal of molecular biology. 1997;274(4):562.
120. Ando T, Yu I, Feig M, Sugita Y. Thermodynamics of Macromolecular Association in Heterogeneous Crowding Environments: Theoretical and Simulation Studies with a Simplified Model. The journal of physical chemistry B. 2016;120(46):11856-65.
121. Nogales E. The development of cryo-EM into a mainstream structural biology technique. Nat Methods. 2016;13(1):24-7.
122. Bai XC, McMullan G, Scheres SH. How cryo-EM is revolutionizing structural biology. Trends Biochem Sci. 2015;40(1):49-57.
123. Novotni M KR. 3D Zernike descriptors for content based shape retrieval. ACM symposium on solid and physical modeling proceedings of the 8th ACM symposium on Solid modeling and applications. 2003:216-25.
124. Kawabata T. Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a gaussian mixture model. Biophys J. 2008;95(10):4643-58.
125. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605-12.
126. Frigo M, G. JS. The Design and Implementation of FFTW3: Proceedings of the IEEE; 2005.
127. Terashi G, Kihara D. De novo main-chain modeling for EM maps using MAINMAST. Nat Commun. 2018;9(1):1618.
128. Terashi G, Kihara D. De novo main-chain modeling with MAINMAST in 2015/2016 EM Model Challenge. J Struct Biol. 2018;204(2):351-9.
129. Joseph AP, Malhotra S, Burnley T, Wood C, Clare DK, Winn M, et al. Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. Methods. 2016;100:42-9.
130. Whicher JR, Dutta S, Hansen DA, Hale WA, Chemler JA, Dosey AM, et al. Structural rearrangements of a polyketide synthase module during its catalytic cycle. Nature. 2014;510(7506):560-4.
131. Carroni M, Kummer E, Oguchi Y, Wendler P, Clare DK, Sinning I, et al. Head-to-tail interactions of the coiled-coil domains regulate ClpB activity and cooperation with Hsp70 in protein disaggregation. Elife. 2014;3:e02481.
132. Yan Z, Yin M, Xu D, Zhu Y, Li X. Structural insights into the secretin translocation channel in the type II secretion system. Nat Struct Mol Biol. 2017;24(2):177-83.

133. Monroe L, Terashi G, Kihara D. Variability of Protein Structure Models from Electron Microscopy. Structure. 2017;25(4):592-602 e2.
134. Na Song DC, Charles W. Christoffer, Xusi Han, Daisuke Kihara, Guillaume Levieux, Matthieu Montes, Hong Qin, Pranjal Sahu, Tenki Terashi, Haiguang Liu. SHREC 2017 – Classification of Protein Shapes. Eurographics Workshop on 3D Object Retrieval: Eurographics Workshop on 3D Object Retrieval; 2017.
135. Florent Langenfeld AA, Halim Benhabiles, Petros Daras, Andrea Giachetti, Xusi Han, Karim Hammoudi, Daisuke Kihara, Tuan M. Lai, Mahmoud Melkemi, Stelios K. Mylonas, Genki Terashi, Yufan Wang, Feryal Windal, and Matthieu Montes. SHREC'19 Protein Shape Retrieval Contest. Eurographics Workshop on 3D Object Retrieval: Eurographics Workshop on 3D Object Retrieval; 2019.
136. Ilja Gubins GvdS, Remco C. Veltkamp, Friedrich Förster, Xuefeng Du, Xiangrui Zeng, Zhenxi Zhu, Lufan Chang, Min Xu, Emmanuel Moebel, Tuan M. Lai, Xusi Han, Genki Terashi, Daisuke Kihara, Benjamin A. Himes, Xiaohua Wan, Jingrong Zhang, Shan Gao, Yu Hao, Zhilong Lv, Xiaohua Wan, Zhidong Yang, Zijun Ding, Xuefeng Cui, Fa Zhang. SHREC'19 Track: Classification in Cryo-Electron Tomograms. Eurographics Workshop on 3D Object Retrieval: Eurographics Workshop on 3D Object Retrieval; 2019.
137. Xu M, Singla J, Tocheva EI, Chang YW, Stevens RC, Jensen GJ, et al. De Novo Structural Pattern Mining in Cellular Electron Cryotomograms. Structure. 2019;27(4):679-+.
138. Goodsell DS, Dutta S, Zardecki C, Voigt M, Berman HM, Burley SK. The RCSB PDB "Molecule of the Month": Inspiring a Molecular View of Biology. PLoS Biol. 2015;13(5):e1002140.
139. Trapani S, Navaza J. Calculation of spherical harmonics and Wigner d functions by FFT. Applications to fast rotational matching in molecular replacement and implementation into AMoRe. Acta Crystallogr A. 2006;62(Pt 4):262-9.
140. Xu D, Li H, Zhang Y. Protein depth calculation and the use for improving accuracy of protein fold recognition. J Comput Biol. 2013;20(10):805-16.
141. He KM, Zhang XY, Ren SQ, Sun J. Deep Residual Learning for Image Recognition. Proc Cvpr Ieee. 2016:770-8.
142. Deniz CM, Xiang S, Hallyburton RS, Welbeck A, Babb JS, Honig S, et al. Segmentation of the Proximal Femur from MR Images using Deep Convolutional Neural Networks. Sci Rep. 2018;8(1):16485.
143. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol. 2000;297(1):233-49.
144. Li B, Turuvekere S, Agrawal M, La D, Ramani K, Kihara D. Characterization of local geometry of protein surfaces with the visibility criterion. Proteins. 2008;71(2):670-83.
145. Huang B, Schroeder M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol. 2006;6:19.
146. Esquivel-Rodriguez J, Yang YD, Kihara D. Multi-LZerD: multiple protein docking for asymmetric complexes. Proteins. 2012;80(7):1818-33.
147. Li B KD. Protein docking prediction using predicted protein-protein interface. BMC Bioinformatics. 2012;13(7).
148. Venkatraman V, Yang YD, Sael L, Kihara D. Protein-protein docking using region-based 3D Zernike descriptors. BMC Bioinformatics. 2009;10:407.

149. Sael L, Kihara D. Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. Proteins. 2012;80(4):1177-95.
150. Chikhi R, Sael L, Kihara D. Real-time ligand binding pocket database search using local surface descriptors. Proteins. 2010;78(9):2007-28.
151. Venkatraman V, Chakravarthy PR, Kihara D. Application of 3D Zernike descriptors to shape-based ligand similarity searching. J Cheminform. 2009;1:19.
152. Dym H, McKean HP. Fourier series and integrals. New York,: Academic Press; 1972. x, 295 p. p.
153. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Sci. 1992;1(3):409-17.
154. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, et al. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 2008;36(Database issue):D419-25.
155. Khare G, Gupta V, Nangpal P, Gupta RK, Sauter NK, Tyagi AK. Ferritin structure from Mycobacterium tuberculosis: comparative study with homologues identifies extended C-terminus involved in ferroxidase activity. PLoS One. 2011;6(4):e18570.
156. Barber CB DD, Huhdanpaa H. The Quickhull algorithm for convex hulls. ACM T Math Software. 1996;22(4):469–83.
157. Redinbo MR, Stewart L, Kuhn P, Champoux JJ, Hol WG. Crystal structures of human topoisomerase I in covalent and noncovalent complexes with DNA. Science. 1998;279(5356):1504-13.
158. Procko E, Ferrin-O'Connell I, Ng SL, Gaudet R. Distinct structural and functional properties of the ATPase sites in an asymmetric ABC transporter. Mol Cell. 2006;24(1):51-62.
159. Lammens A, Schele A, Hopfner KP. Structural biochemistry of ATP-driven dimerization and DNA-stimulated activation of SMC ATPases. Curr Biol. 2004;14(19):1778-82.
160. Ranson NA, Clare DK, Farr GW, Houldershaw D, Horwich AL, Saibil HR. Allosteric signaling of ATP hydrolysis in GroEL-GroES complexes. Nat Struct Mol Biol. 2006;13(2):147-52.
161. White HE, Orlova EV, Chen S, Wang L, Ignatiou A, Gowen B, et al. Multiple distinct assemblies reveal conformational flexibility in the small heat shock protein Hsp26. Structure. 2006;14(7):1197-204.
162. Furuya T, Ohbuchi R. Learning part-in-whole relation of 3D shapes for part-based 3D model retrieval. Computer Vision and Image Understanding. 2018;166:102-14.
163. Gainza P, Sverrisson F, Monti F, Rodola E, Boscaini D, Bronstein MM, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods. 2019.
164. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Commun Acm. 2017;60(6):84-90.
165. Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation. 2015 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr). 2015:3431-40.
166. Noh H, Hong S, Han B. Learning Deconvolution Network for Semantic Segmentation. Ieee I Conf Comp Vis. 2015:1520-8.

VITA

Xusi Han was born and raised in Puyang, Henan, China. She attended No. 1 High school, where biology was one of her favorite subjects. She obtained her Bachelor of Science degree in Biology from Minzu University of China at Beijing in 2013. During her undergraduate study, she did independent research under the supervision of Professor Fei Gao and Professor Yijun Zhou. Xusi started her graduate study at Purdue University in August 2013 in the PULSe program. During her PhD career, she received rewards including Outstanding Proposal Award from ACLS summer school and Graduate School Summer Research Grant from the Department of Biological Sciences. Meanwhile, she also obtained Master of Science degree in Applied Statistics at Purdue University.