Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By SAEL LEE

Entitled

HIGH THROUGHPUT SCREENING OF GLOBAL AND LOCAL PROTEIN SURFACES

For the degree of DOCTOR OF PHILOSOPHY

Is approved by the final examining committee:

DAISUKE KIHARA (Chair)

LUO SI

MICHAEL R GRIBSKOV

ROBERT D SKEEL

MIKHAIL ATALLAH

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): DAISUKE KIHARA

Approved by: ADITYA MATHUR / WILLIAM J GORMAN (Head of the Graduate Program) JULY 26, 2010 (Date)

Graduate School Form 20
(Revised 1/10)

PURDUE UNIVERSITY
GRADUATE SCHOOL

Research Integrity and Copyright Disclaimer

Title of Thesis/Dissertation:
HIGH THROUGHPUT SCREENING OF GLOBAL AND LOCAL PROTEIN SURFACES

For the degree of DOCTOR OF PHILOSOPHY

I certify that in the preparation of this thesis, I have observed the provisions of *Purdue University Teaching, Research, and Outreach Policy on Research Misconduct (VIII.3.1)*, October 1, 2008.*

Further, I certify that this work is free of plagiarism and all materials appearing in this thesis/dissertation have been properly quoted and attributed.

I certify that all copyrighted material incorporated into this thesis/dissertation is in compliance with the United States' copyright law and that I have received written permission from the copyright owners for my use of their work, which is beyond the scope of the law. I agree to indemnify and save harmless Purdue University from any and all claims that may be asserted or that may arise from any copyright violation.

SAEL LEE
Printed Name and Signature of Candidate

07/16/2010
Date (month/day/year)

*Located at http://www.purdue.edu/policies/pages/teach_res_outreach/viii_3_1.html

HIGH THROUGHPUT SCREENING OF GLOBAL AND LOCAL PROTEIN SURFACES

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Sael Lee

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2010

Purdue University

West Lafayette, Indiana

ACKNOWLEDGMENTS

I greatly thank my advisor, Daisuke Kihara, who has helped me in every step of my graduate studies. He has brought me into the intriguing field of structural bioinformatics and I have learned many important things about research under his guidance. I would like to thank my advisory committee, Michael Gribskov, Robert Skeel, Lou Si, and Mikhail Atallah, for their interest and generous advise. I would also like to thank my lab mates for helpful comments and their friendship. I would especially like to thank David La for his encouragement. I thank my family for their love and trust. Most of all, I thank the Lord for always being there for me.

TABLE OF CONTENTS

Page

LIST OF TABLES iv

LIST OF FIGURES v

ABSTRACT vi

1 INTRODUCTION 1

2 REVIEW OF THE PROTEIN SURFACE REPRESENTATION AND COMPARISON 7
- 2.1 Evaluation criteria 7
- 2.2 The surface representation 9
  - 2.2.1 General object representation 9
  - 2.2.2 Protein surface definition 9
- 2.3 General object analysis methods 10
  - 2.3.1 Global shape analysis 10
  - 2.3.2 Local shape analysis 15
- 2.4 Protein surface analysis methods 15
  - 2.4.1 Graph-based methods 16
  - 2.4.2 Geometric hashing 17
  - 2.4.3 Methods using series expansion of 3D function 18
  - 2.4.4 Other methods 19
- 2.5 3D Zernike descriptors (3DZDs) 21
  - 2.5.1 Characteristics of 3DZDs 21
  - 2.5.2 Derivation of 3DZDs 22
  - 2.5.3 Computing distances of 3DZDs 23
  - 2.5.4 3DZD and spherical harmonic descriptor 24
- 2.6 Chapter summary 26

3 PROTEIN TERTIARY STRUCTURE RETRIEVAL BASED ON GLOBAL SURFACE SHAPE SIMILARITY 27
- 3.1 Materials and methods 28
  - 3.1.1 Benchmark dataset 28
  - 3.1.2 Input generation 29
  - 3.1.3 DALI protein structure comparion program 30
  - 3.1.4 Benchmark procedure 30
- 3.2 Results 31
  - 3.2.1 Examples of 3DZD 31

Page

3.2.2 Rotation invariance test . . . 32
3.2.3 Comparison of the CE and the SCOP classifications on the benchmark dataset . . . 34
3.2.4 Retrieval performance of 3DZD and DALI against CE classification . . . 36
3.2.5 Examples of individual 3DZD retrieval cases . . . 38
3.2.6 Resolution of 3DZDs . . . 41
3.2.7 Computation time . . . 43
3.2.8 3D-Surfer: A web application . . . 45
3.3 Chapter summary . . . 45

4 IMPROVED GLOBAL PROTEIN STRUCTURE SEARCH . . . 47
4.1 Materials and methods . . . 48
4.1.1 Benchmark dataset . . . 48
4.1.2 Computing four types of protein surfaces . . . 49
4.1.3 Evaluation method . . . 49
4.1.4 Combining two descriptors: AASurf and CACNO . . . 51
4.2 Results . . . 52
4.2.1 Retrieval performance . . . 52
4.2.2 Evaluation of AASurf and CACNO . . . 53
4.2.3 Improving retrieval performance . . . 56
4.2.4 Towards application in comparing EM density maps . . . 58
4.3 Chapter summary . . . 61

5 RAPID GLOBAL COMPARISON OF PROPERTIES ON PROTEIN SURFACE . . . 63
5.1 Materials and methods . . . 64
5.1.1 Surface representation . . . 64
5.1.2 Extracting 3DZD for physicochemial properties . . . 65
5.2 Results . . . 65
5.2.1 Clustering of color patterns on spheres . . . 65
5.2.2 Globin family proteins . . . 66
5.2.3 Thermophilic and mesophilic proteins . . . 72
5.2.4 Computation time . . . 75
5.3 Chapter summary . . . 75

6 LOCAL COMPARISON OF SEGMENTED PROTEIN SURFACES . . . 76
6.1 Materials and methods . . . 78
6.1.1 Datasets . . . 79
6.1.2 Protein surface property calculation . . . 80
6.1.3 Ligand binding pocket extraction by ray casting . . . 82
6.1.4 Local surface patch extraction . . . 83
6.1.5 Pocket comparison and ligand binding prediction . . . 84

Page

6.1.6 Pocket retrieval performance evaluation and ligand type prediction score . . . 88
6.2 Results . . . 89
6.2.1 Binding sites of TIM barrel proteins . . . 89
6.2.2 Reconstruction of local surface shape . . . 89
6.2.3 Ligand local property variance . . . 92
6.2.4 Retrieval performance difference of global 3DZD and local 3DZD methods . . . 92
6.2.5 Ligand type retrieval and prediction accuracy of individual pocket types . . . 96
6.2.6 Performance of local 3DZD method on flexible ligands . . . 98
6.2.7 Retrieving chemical moiety using local 3DZD method . . . 101
6.2.8 Computation time . . . 103
6.3 Chapter summary . . . 103

7 TOWARDS ANNOTATING PROTEIN SURFACES . . . 105
7.1 Materials and methods . . . 106
7.1.1 Data set . . . 107
7.1.2 Segmenting protein surface to local patches . . . 107
7.1.3 SOM and ESOM . . . 108
7.2 Results . . . 110
7.2.1 ESOM maps . . . 110
7.2.2 Local patch types of ligand binding surfaces . . . 112
7.2.3 Computation time . . . 116
7.3 Chapter summary . . . 117

8 CONCLUSION . . . 118

LIST OF REFERENCES . . . 121

VITA . . . 134

LIST OF TABLES

Table		Page
3.1	Comparison between the CE based benchmark dataset and the SCOP database	35
3.2	Summary of the structure retrieval using the three distance measures	37
3.3	Sructural comparison performed by CE and 3DZD between the pairs of protein	40
3.4	Pairs of proteins that have similar surface shape defined by 3DZD	42
3.5	Execution time	45
4.1	Precision-recall AUC improvement using weighted distance	57
4.2	Precision-recall AUC for database retrieval of EM isosurfaces	61
5.1	Range of 3DZD distances using representative protein set and globins	67
6.1	The ligand pocket benchmark dataset	81
6.2	ROC AUCs using different parameters	95
6.3	Variance of ROC AUCs in ligand type using different combination of protein properties	97
6.4	Summary of the binding ligand prediction on Kahraman dataset	99
6.5	Database search speed of pocket comparison methods	103

LIST OF FIGURES

Figure		Page
2.1	Reconstruction of 3D Zernike moments	22
3.1	3DZDs of two example proteins.	32
3.2	Variance of 3DZDs upon rotation of proteins.	33
3.3	The sensitivity-specificity plot of the benchmark dataset using the three distance definitions of 3DZD	38
3.4	Examples of protein pairs that have the same main chain orientation but with different surface shapes and pairs that have similar surface shapes but with different main chain orientation	39
3.5	Examples of protein pairs whose surface shapes are judged to be similar by 3DZD	41
3.6	Resolution of 3DZDs	44
4.1	Four protein surface representations	50
4.2	Precision-recall curves on the CE and the SCOP classification	53
4.3	Effect of the sphericity and the tail-like structure to the precision-recall AUCs of AASurf and CACNO representations.	54
4.4	Differences in the database retrieval performance of the AASurf and the backbone representations	55
4.5	Examples of retrieval performances with the AASurf and CACNO combination	58
4.6	Comparisons of surfaces constructed from EM maps and reconstructed molecular surfaces from 3D Zernike moments	59
5.1	Analysis of coloring patterns on spheres using complete linkage clustering.	66
5.2	Distribution of the 3DZD distances using the representative protein set and the globin family	68
5.3	Histograms of CC distances between the electrostatic potentials of the representative proteins using different orders	70
5.4	Electrostatic potential distances of globins	71

Figure Page

5.5 Surface electrostatic potential comparisons of thermophilic proteins and their mesophilic homologues 73

5.6 Complete linkage clustering of surface electrostatic potentials for GDHs and TBPs using HL2 measure 74

6.1 Flow chart of local 3DZD method 78

6.2 Nine ligand structures 79

6.3 Binding site electrostatic potential of 19 TIM barrel structures 90

6.4 Local surface shape reconstruction from the 3DZDs 91

6.5 Average 3DZD distance values map to closest ligand atoms 93

6.6 ROC curves of pocket shape retrieval using global 3DZD and local 3DZD methods 94

6.7 Examples of pocket matches of the flexible ligands FAD and NAD . . . 100

6.8 Adenosine region matching of AMP, ATP, FAD, and NAD with the local 3DZD method. 101

6.9 Flavin binding region matching with local 3DZD method 102

7.1 Flowchart of the protein surface classification method. 106

7.2 ESOM maps of local surface patches. 111

7.3 Affinity propagation clustering of ESOM neurons. 113

7.4 Histograms of the ESOM neuron clusters for binding sites of the six ligand molecules 114

7.5 Distribution of the ligand binding surface patches on the ESOM map. . 115

ABSTRACT

Lee, Sael. Ph.D., Purdue University, August 2010. High throughput screening of global and local protein surfaces. Major Professor: Daisuke Kihara.

The comparative study of protein tertiary structures provides rich information for investigating the function and evolution of proteins, which are the constructors and maintainers of all living things. Traditionally, comparative studies have focused on using genetic sequences to study the function and the evolution of proteins. However, with the accumulation of structural information, more direct approaches for gaining information through protein structure comparisons are also possible. These require methods that are able to quickly find proteins of similar structure over large datasets. However, most of the existing structure comparison methods use full structure alignment methods that are not quick enough for database searching.

In this dissertation, methods for rapid comparison of protein tertiary structures are proposed. For this purpose, among the protein tertiary structure representations, the protein surface is used. The shape, physicochemical properties and other properties of the protein surface determine the recognition process of other proteins, ligands, DNA, or other molecules with which it interacts. To enable quick searching, the 3D Zernike descriptor, one of the object abstraction methods used in the graphics field, is used to efficiently represent and compare protein surfaces. The most attractive aspect of the 3D Zernike descriptor is rotational invariance. Rotation invariance means that structural alignment is not necessary, which speeds up the searching process.

This dissertation makes three major contributions: 1) the development of a rapid method for protein function prediction using the global 3D shape, electrostatic potential, and hydrophobicity of the protein surface; 2) the development of local protein surface comparison methods and their application to ligand binding prediction; 3) the

characterization and classificiation of local protein surface patches. The first method assumes that the protein structure relationships are preserved by evolution even when proteins share little sequence similarity. The latter two approaches assume that there are structural similarities among proteins of similar function even when structures are globally different. Local methods directly search for geometrical and/or physico-chemical properties of significant sites; it is possible to predict molecular functions of proteins that lack global sequence/structural homology to proteins of known function.

1 INTRODUCTION

Bioinformatics is the large-scale computational analysis of biological data using traditional methods in computer science and statistics to provide new biological insights. Molecular biological data include DNA sequences, amino acid sequences, protein structures, gene expressions, protein interactions, etc. Some of the informatics methods used are string comparison methods, pattern extraction and analysis methods such as machine learning and data clustering, computer vision methods such as 3D searching, simulation methods that utilize various numerical analyses, and network analysis methods. There are several specific research areas in bioinformatics that can be characterized by the data and/or methods used.

The branch of bioinformatics that relates mainly to the structure of biomolecules is called structural bioinformatics. Problems in structural bioinformatics include protein structure prediction, protein folding, protein-ligand interaction, protein-protein docking, establishing protein structure-function relationships, and protein design. Proteins are the basic constructors and maintainers of all living organisms, as they structure the cell and catalyze biochemical reactions. They are also the main players in cell signaling, immune system responses, and many other critical biological processes. The structures and physicochemical properties of proteins determine their functional roles. This dissertation proposes novel comparative protein analysis approaches that utilize protein surfaces encoded with 3D Zernike descriptors. A brief history of bioinformatics/structural bioinformatics is outlined to provide a better understanding of where the current study fits in the area of bioinformatics.

The field of bioinformatics has been surging in the past decade and is becoming even more important as biological information increases. According to Hagen [1], although bioinformatics seem to be a new field, the emergence of computational work in biology dates back to the early 1960's with the expansion of the idea that pro-

teins carry information, along with the new availability of computers to the academy. Two events impacted the understanding of biology that proteins carry information as amino acid sequences: 1) the sequencing of α-chain insulin by Frederick Sanger in 1951, which was the first protein to be sequenced; and 2) the solving of the first two protein structures by Max Perutz (hemoglobin) and Sir John Kendrew (myoglobin) in 1958 [2,3]. These two events influenced the growth of areas in molecular sequencing, structural biology, and evolution of genes and proteins. Some of the early efforts in these areas are pointed out in the reviews by Hagen [1] and Ouzounis and Valencia [4].

Pioneering studies in the 1960's include development of computer programs to aid protein sequence determinations [5, 6] by Margaret Dayhoff; the start of Atlas of Protein Sequence and Structure by Dayhoff in 1965 [7]; the first attempts to use the sequence information on the study of evolution in 1965 [8]; the first construction of phylogenetics trees in 1967 [9]; the first construction of computational molecular model in 1966 [10].

In the 1970's to 1990's, pioneering algorithms for sequence analysis were introduced and there were many centralized efforts to obtain and maintain sequence data. Additionally, structural biology steadily developed as computers became essential tools for solving the structure of proteins. Some of the events in the period include the publication of the Needleman-Wunsch algorithm for global sequence alignment in 1970 [11]; the establishment of the Protein Data Bank, originally containing seven structures in 1971; invention of rapid DNA sequencing methods by Sanger [12], and Maxam and Gilbert [13] in the mid 1970's; the first gene sequencing of an organism, Bacteriophage FX174, in the 1970's; the publication of the Smith-Waterman algorithm for local sequence alignment [14]; the publishing of the fast heuristic database searching algorithm for amino acids, FASTP, by Lipman and Pearson in 1985 [15]; the announcement of the Human Genome Initiative by the Department of Energy (DOE) in 1986, which later developed into the Human Genome Project; the establishment of the National Center for Biotechnology Information (NCBI) in 1988 to provide access to biomedical and genomic information; and the Human Genome Project, which was

formally initiated by the DOE and National Institutes of Health (NIH) in 1990 with the goal of identifing all the genes in human DNA and determining the sequences of the billions of base pairs that make up human DNA. The first complete genome was sequenced for *Haemophilus influenza*, in 1995 [16] and the human genome was completed in 2003.

With the increasing availability of sequence information, researchers hope to gain understanding of how genes encode protein function, and thus better understand disease mechanisms. However, understanding protein function based only on sequence has its limits. A fuller understanding of protein function, e.g. the protein-ligand interaction, relies heavily on protein structures. Moreover, protein structures are generally more conserved than sequence, which can also be infered from the consensus that although the sequence variance can be thought as infinite there are only approximately 10,000 protein families of structure, i.e. proteins with similar structure [17]. Thus, interest spread to the accumulation of protein structures to fill the gap between the knowledge of sequence and structure. As a result, in the year 2000 the Protein Structure Initiative (PSI) started with the 10-year goal to 1) develop methods to increase the efficiency of structure determination, 2) construct and automate a protein production and structure determination pipeline, and 3) determine protein structures for all non-homologous proteins (less than 30% sequence identity) [18]. As of June 2010, a total of 61,318 proteins have been deposited in the wwPDB [19], which is a database formed with the mission of maintaining an freely available structural database of macromolecular structural data, stored in the protein data bank (PDB) format, for the global community [19].

The importance and the impact of computational studies in biology is ever increasing, with the accumulation of large scale biological data from various sources, including genome sequences, protein-protein interaction, and protein structures. This data contains key information for understanding the behavior of molecules in the biological systems essential for sustaining life. It is expected that bioinformatics will play a significant role in analyzing such data, as computational techniques are indis-

pensable in the analyses. A particularly important and interesting problem is protein structure analysis for function prediction, as structural genomics projects have been solving an increasing number of protein structures of unknown function. Indeed, as of June 2010, there are over 3100 proteins of unknown function in the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Structure Databank (PDB) [20]. These structures remain as proteins of unknown function because no experiments have been done to characterize their functions. Ongoing efforts to improve function prediction include development of sequence-based methods which are more sensitive and accurate than conventional methods [21, 22].

Alternatively, one can use tertiary structure information for identifying similarity to proteins with known function that are stored in the PDB. Potential advantages of using structural information are twofold: 1) evolutionarily distant proteins to a given query protein could be identified when the global structure is more conserved than the primary sequence [23, 24]; 2) local physical features of proteins can be directly compared for identifying functional sites where interaction of ligand molecules or other proteins takes place [25]. Assuming that proteins of similar structure have similar function [24, 26], there have been various attempts to predict a function of a protein by evaluating the functions of proteins that have similar structures to the protein [27]. Commonly used representations of protein structures include the backbone Cα positions [28, 29], distance maps [30], and molecular surfaces [31]. Computational methods for comparing protein structures depend on the representations in use, such as the dynamic programming method for backbone comparisons [24]. Different representations capture different aspects of protein structure so they may not necessarily agree in the degree of structural similarity they can identify [32].

In our study, the protein surface is considered in the comparison of proteins. Using surface information and not the backbone enables us to compare proteins that are evolutionarily distant. Furthermore, the 3D shape and physicochemical properties on protein surfaces carry essential information for understanding protein function. For example, the enzyme catalytic reaction is realized by a set of atoms in the active

site [33]. Also residues at an protein-protein docking interface surface establish physical contact in protein-protein interactions [34, 35]. Surface comparison is useful for function prediction because global and local geometrical and physicochemical properties of protein structures of known function can be directly compared to those with unknown functions.

The goal of this study is to develop a high throughput protein surface analysis method that is able to globally and locally examine the similarity of protein surface properties, such as geometric shape and physicochemistry. The focus of this work is the application and development of computational methods to overcome three main challenges: 1) testing an efficient protein surface encoding method to reduce the time and space complexity for comparing proteins; 2) local matching of segmented protein surfaces, where distantly related proteins that may share a function can be detected by local similarities in the surfaces; 3) developing and testing a scoring function that will accurately incorporate the heterogeneous protein surface information, such as geometric shape, electrostatic potential, and hydrophobicity.

To address these challenges, the 3D Zernike descriptor [36] is used to efficiently represente protein surface properties. We have tested other 3D shape descriptors but have found 3D Zernike descriptor to be most efficient in terms of time, space and accuracy. The 3D Zernike descriptor is based on a series expansion of a given 3D function which has several valuable characteristics. First, it allows fast retrieval of protein structures due to compactness and rotation invariance character of the descriptor. This is a major advantage since current major structure databases, including PDB, CATH [37], and SCOP [38] only allow keyword searching and browsing of a pre-computed classification. The DALI server [30], VAST search at NCBI [39], eF-seek database [40], and ProFunc [41] allow users to search the database with a query structure, but a search often requires hours to finish. Ideally, a routinely used protein structure comparison requires comparisons against a large number of structures and should be done instantly, just as a BLAST [42] sequence search. Secondly, since the 3D Zernike descriptor is a 3D function, any scalar characteristic of a protein surface

can be naturally incorporated into the descriptor by considering an appropriate 3D function. Moreover, the resolution of the description of protein structures can be simply altered by changing the order of the 3D Zernike descriptor considered.

The contribution of our work is threefolds: 1) development of a rapid method for protein function prediction using the global 3D shape, electrostatic potential, and hydrophobicity of the protein surface; 2) development of local protein surface comparison methods and their application to ligand binding prediction; and 3) characterization and classification of local protein surface patches so that they can be used in the annotation of protein surfaces. Thus, our work effectively encodes the complex nature of the protein surface and may help start a cascade of structural bioinformatics work related to surfaces, including the protein-protein docking, ligand docking, and protein database searching. There are also further applications in the protein function prediction and EM image analysis.

The dissertation is organizated as follows. A review of computational techniques for representing and comparing protein 3D surface shapes is provided. The 3D Zernike descriptor, which is used throughout the dissertation, is also described. Then three chapters introduce a method of global protein surface comparison using 3D Zernike descriptors. First, high-throughput protein surface shape comparison is performed. Then the retrieval performance of all-atom surface shape and main-chain surface shape is compared. Next, a quantitative method for comparing protein surface physicochemical properties is introduced. The last two chapters focus on the analysis of local regions of protein surfaces, starting with the introduction of the method for local structure based function prediction, followed by a proposal for a protein surface annotation based on characterizing protein surface using classified local surface patches.

2 REVIEW OF THE PROTEIN SURFACE REPRESENTATION AND COMPARISON: NEW APPROACHES IN STRUCTURAL PROTEOMICS [1]

In this chapter, a review of computational techniques for representing and comparing protein 3D surface shapes is provided. First, an overview of evaluation criteria for 3D shape analysis methods is provided. Then discussion of how a protein surface is defined is performed. Next, 3D object analysis methods developed in computational geometry and graphics field are reviewed. In the last section of this chapter, 3D Zernike descriptors, which have recently been applied to 3D object comparison [36], are discussed. Applications of the 3D Zernike descriptor (3DZD) to the protein surface, both globally and locally, will be provided in the subsequent chapters.

2.1 Evaluation criteria

Tangelder and Veltkamp suggest six criteria for evaluating the characteristics of general object shape and protein surface analysis methods [44]. These criteria will be refered to when describing existing methods in the following sections.

1. Invariance to Euclidean transformation. The three Euclidean transformations, i.e. rotation, scaling, and translation, do not change the shape of the original object. Some methods use representations that are invariant to Euclidean transformation. Euclidean transformation invariant representations are convenient in comparing objects, because their representations can be directly compared without preprocessing. When non-invariant representations are employed, either alignment of compared objects is needed, which is time consuming, or a normalization process of object positions is needed prior to analysis.

[1] This chapter reuses published work in [43].

2. Need for pose normalization. Pose normalization is the rotation, translation, and scaling of 3D object to a standard position for comparison. Normalizing in terms of translation and scaling can easily be done, for example, by moving the object in such a way that its center of gravity is located at the origin and then enlarged or shrinked the object so that it fits compactly into a unit sphere. However, normalization against rotation is often not robust. Principal Component Analysis (PCA) is the most widely practiced method for pose normalization. PCA often fails to provide a unique robust solution when an object has symmetrical mass distribution; i.e. PCA generates many equal eigenvalues, which suggests more than one possible positioning of the object. This is especially problematic in handling protein shapes, because they are more or less spherical.

3. Ability for partial matching. Partial matching aims to find similar local regions of two objects. It is especially important in protein matching considering that the function of a protein is attributed to local surface regions such as active sites and protein docking interfaces.

4. Capability of changing resolution. Depending on the context of the object analysis, being able to adjust the level of details of an object description can be useful. Higher resolutions provide detailed information and is more sensitive to small changes while lower resolutions are more focused in the overall shape and are less sensitive to small changes in shape.

5. Tolerance to small amounts of noise and/or changes in shape. This property is also important for protein shape analysis. Proteins are flexible in nature and structures are solved experimentally at different resolutions depending on the experimental conditions and the methods used. Moreover, if a predicted structure is used for analysis, some errors are unavoidable. Thus it is important to account for these changes by allowing resolution changes or by making the method tolerant to small differences in shapes which can be caused by noise and/or by flexible nature of proteins.

6. Ability to incorporate additional properties. Non-shape properties such as electrostatic potential, hydrophobicity, and residue conservation are important fac-

tors in determining and analyzing the function of proteins. Thus, being able to use additional properties can extend the applicability of protein surface analysis.

2.2 The surface representation

2.2.1 General object representation

There are three types of object representations, volume-, boundary-based and point cloud methods. Well known volume-based representations are voxels and octrees. In voxels, the volume of an object decomposed on a 3D grid while in octrees the object space is hierarchically subdivided. Widely used boundary based representations include polygon meshes. A polygonal mesh comprises nodes and edges that form faces that are connected to completely cover the surface of an object. In point clouds, a set of (x, y, z) points are used to represent a object.

Generally, volume-based and point cloud representations are used to for experimental data, such as computed tomography scans. On the other hand, boundary-based representations are used for computer designed objects, such as ones used in computer games. Volume-based representation requires a larger space but can provide information about the interior of an object while boundary-based representation is efficient in drawing an object on the computer screen.

2.2.2 Protein surface definition

The physical surface of a protein is the set of atomic surfaces that are defined by the van der Waals radius of each atom in the protein. Thus an intuitive way of defining a protein surface is to compute the union of boundaries of spheres with the van der Waals radii corresponding to the protein atoms (the van der Waals surface). Inflated (i.e. enlarged) van der Waals radii are often used for defining the surface. However, direct use of the van der Waals sphere of atoms usually leaves unoccupied spaces between atoms, making small clefts and cavities on the surface. Those small cavities,

which water molecules and ions cannot enter, are negligible but cause unnecessary noise for many applications of protein surface representations. A common way to obtain a smoother surface is to roll a probe sphere (usually of the size of a water molecule) over the van der Waals surface and to trace the center of the sphere (solvent accessible surface) or to trace the inward-facing surface of the probe sphere (solvent excluded surface or Connolly surface [45]). Other protein surface definitions include α-surface [46]. The algorithm for construction of an α-surface connects points to construct triangle meshes, whose resolution is controlled by a parameter, α. The solvent accessible surface and the Connolly surface are also usually represented by triangle meshes. Other representations, such as point cloud and voxels are also used.

There is not strong preferences in the representation of proteins and the choice of representation primarily depends context and methods used in the studies.

2.3 General object analysis methods

This section provides a list of well known shape analysis methods in computer science and other engineering fields that have been or have the potential of being applied to proteins. Roughly, methods can be classified as global and local shape analysis methods.

2.3.1 Global shape analysis

Global shape descriptors represent the overall shape of objects. They can be classified into three categories, feature and feature distribution-based methods, 3D coordinate centered methods, and view-based methods.

Feature/feature distribution-based methods

Methods in this category describe an object by one or a set of features of the object, such as the volume and the area of surface. These methods are some of the earlier

methods developed, and are still actively applied individually or as a component of more complex methods. An advantage of these methods is that non-shape properties of objects can be easily combined with shape-based features. On the other hand, the disadvantage is that they are obviously less descriptive since an object shape is represented by smaller number of features. Multiple features are represented as a feature vector.

Elad et al. [47] use statistical moments, which describe the distribution of the position of vertices on a polygon mesh of an object, i.e. the center of gravity, variance, and skewness etc. as features of the object. The extracted moments are used as a feature vector and compared using a weighted Euclidean distance. In their method, the objects are pose normalized by normalizing against the first two moments (center of gravity and variance) of the surface points prior to extracting the moments. Zhang and Chen also utilize the statistical moment in addition to some other features [48]. They propose an efficient method for computing and comparing the global features including the volume, the surface area, the volume-surface ratio, the statistical moments, and the coefficients of the Fourier transform of 3D objects.

Rather than representing an object by a single value, feature distribution-based methods use a histogram of global shape features as a descriptor. An example is the shape distribution, which uses a histogram of the Euclidean distance of randomly chosen two points on the surface of an object [49]. The solid angle histogram places a sphere at each representing points of an object and computes the fraction of volume of the sphere occupied by the object [50].

3D coordinate centered methods

These methods directly represent the 3D shapes of objects in space. There are two main approaches in this category. The first approach is a mathematical transformation of a 3D object, which approximate the object by fittng the mathematical

function to the object. Another approach is to compute how an object occupies the 3D space by its volume when the object is represented in voxels.

Mathematical transformations have been widely studied in 2D image processing. The 3D version of those transformations, such as Fourier, Hough, Radon, and wavelet transformations, have been applied for 3D objects. These methods are not invariant to Euclidean transformations and need pose normalization prior to extraction of descriptors. More recently, spherical harmonics have been widely explored in context of 3D shape analysis. Spherical harmonics are functions with two parameters, a polar angle θ, and a colatitude angle φ: $Y_l^m(\theta, \varphi)$. Since spherical harmonics form an orthonormal set of functions, a 3D function (thus an 3D object) can be expanded as a series of spherical harmonics with a different degree l and an order m on the unit sphere. Limitations in direct applications of spherical harmonics include its non-invariance to Euclidean transformations and also that they can correctly capture only star like shapes, i.e. shapes that have no re-entrant surfaces.

Duncan and Olson [51] introduced a direct application of spherical harmonics on representing proteins surface which provided an efficient representation of protein surface. However, because the spherical harmonic functions form basis set on defined on unit sphere, only protein that are topologically similar to spheres could be used.

Funkhouser's group introduced a spherical harmonics-based shape descriptor that is rotation invariant and can also be applied to non-star like shapes [52]. The method first segments an object into concentric spheres and then computes spherical harmonics for each of the spheres. Since rotating a spherical function does not change its L^2 norm, combining the L^2 norm computed for each group of harmonics of the same parameters (l and m) yields a rotation invariant (thus invariant to Euclidean transformations) descriptor. Non-star like shapes are better handled by the segmentation to concentric spheres. Application of spherical harmonics in partial matching has also been made by the same group as an extension of the spherical harmonic-based method [53].

The above method considers the radial information (the distance from the center of an object) by the segmentation of an object into concentric spheres. In contrast, 3D Zernike descriptors uses Zernike-Canterakis basis $Z_{nl}^{m}(\mathbf{x})$, which are 3D functions that can be defined over Cartesian coordinates $\mathbf{x} = (x, y, z)$ [54] with a parameter l called the order and two parameters m and n that depends on l. Thus 3D Zernike descriptors are a convenient wasy to handle a 3D object described in points or voxels in Cartesian coordinates. Rotation invariance was obtained later by Novotni and Klein [36] by combining the L^2 norms of same parameters (l and m) similar to the approach of Kazhdan et al. [52]. This thesis describes application of 3D Zernike descriptors for protein surface comparison, which will be discussed in sections 2.5.1 and 2.5.4. Also the details on the derivation of the 3D Zernike descriptor is provided in 2.5.2.

The other special functions introduced in 3D object analysis include spherical wavelets and Krawtchouk polynomials. Mathematically, the spherical wavelet descriptor [55] has two advantages over spherical harmonics. First, the level of detail of the description can be locally controlled. Second, the sampling of points is more uniform. The weighted 3D Krawtchouk descriptor [56] uses polynomials of discrete variables, and thus eliminates the need for a spatial discretization process. Hence no numerical approximation is involved in handling a voxelized object data. A drawback of Weighted 3D Krawtchouk descriptor is again the need for pose normalization.

When objects are represented by voxels, two objects can be compared by computing the difference of distribution of the occupied voxels (volumetric difference methods). Occupied voxels of an object can be represented by a tree data structure, e.g., an octree. Thereafter efficient comparison is done based on the tree representation. Volumetric difference methods are generally slower than other global methods and still need pose normalization.

An interesting idea of representing an object is to compute 'energies' or the cost needed to morph the object to a sphere. Then comparison is done by computing the difference in the morphing energies. A method by Leifman et al. [57] calculates

sphere projection energy as $E = \int_{dist} \vec{F} \cdot \, d\,\vec{r}$, where $dist$ is the distance between the sphere and the object surface, and $\vec{F}$ is the applied force which is assumed constant . In a method proposed by Yu et al. [58], a feature map is used to record a local energy needed to morph an object. The object is first normalized and fit into a unit sphere. Then local energy at each point is computed by two terms. The first term is the distance from the object surface to the bounding sphere. The second term is the number of object surfaces penetrated when a ray is shot from the sphere center. This method additionally uses the Fourier transform of the feature map, which shows better tolerance to noise that may have been introduced in the pose normalization process.

View-based methods

View-based methods describe a 3D object as a set of projected 2D images of the 3D object from different viewing angles. Each image contains characteristics of the object from that angle, however, relative spatial information between images from different view points is not captured.

The most well known view-based method are the Light Field Descriptors [59]. In computing the Light Field Descriptor of a 3D object, the object is first scaled and placed into a bounding sphere. Then a light field, which consists of 20 uniformly distributed silhouettes of the object from 10 rotational positions on the bounding sphere, of the object is created. Subsequently, a combination of 2D Zernike moments and Fourier transforms are used as a 2D descriptor for each silhouette. To compare descriptors of two objects, silhouettes of the two objects are compared exhaustively to find matches.

Ohbuchi et al. proposed another view based technique, which captures the depth of an object from each angle in addition to the 2D silhouettes [60]. Then for each image a Fourier transform based descriptor is generated.

2.3.2 Local shape analysis

Local shape analysis aims to capture geometrical features of a local region around a given point on a surface. A curve of a local surface is described using Gaussian curvature, mean curvature, and/or the shape index. Among them, the shape index has been used for protein surface analysis. The shape index [61] is a single-value ranging from $[-1, 1]$, which measures the slope of a local surface using principal curvatures. The spin image is another popular method used to describe a local shape [62, 63]. A spin image is a 2D histogram of distances that are measured from central vertices to their neighboring vertices. Two distances characterize the spatial relationship between a central vertex and a nearby vertex: the radial distance, α, which is defined as the perpendicular distance between a nearby vertex and the surface normal vector of the central vertex; and the axial distance, β, a signed perpendicular distance between the nearby vertex and the tangent plane of the surface normal computed from the central vertex. By definition, a spin image does not change upon rotating around the norm of a central vertex. The spin image is calculated for each vertex on a surface mesh of an object.

Using surface curvature information captured at each vertex, a larger surface region can be described by connecting the vertices as a graph. A graph captures the relative spatial information of vertices and enables partial matching of two local surfaces. In general, however, partial graph matching can often be slow for comparing large graphs. Methods for global shape analysis, such as spherical harmonic-based methods, can also be used for describing a local shape around a vertex.

2.4 Protein surface analysis methods

In this section, we discuss existing methods for protein surface representation and comparison. Identification of similar global and local protein surfaces has application to structure-based function prediction. Protein surface representation has also been studied in context of protein-protein docking and protein-small ligand docking,

in which complementarity of two surfaces is taken into account. Three major categories of protein surface analysis methods, i.e., graph-based, geometric hashing, and methods using series expansion of 3D function, are described.

2.4.1 Graph-based methods

Graph theoretical approaches are frequently applied to protein surface comparison especially for mesh-based surface representation, which can be considered as a graph. In a graph representation of a protein surface, the vertices contain geometrical and physicochemical features of a local region around each vertex; and the edges describe the positional relationship of the vertices they connect. The advantage of a graph representation is that partial matching of two protein surfaces can be done using existing graph theoretical algorithms.

The method proposed by Pickering et al. [64] first generates a Connolly surface of the region of interest. Then, for each vertex point, shape information, such as shape index and radius of curvature are calculated, as well as biological features such as types of residues. The matching process involves solving the maximum common subgraph isomorphism problem of the two graphs representing the protein surfaces.

Kinoshita et al. [31] developed a database of protein surfaces of functional sites, named eF-site, and a method to search for the similar local surface sites in a query protein run on the database. The triangular meshes of a Connolly surface constitute a graph of the protein. Each vertex is described by a electrostatic potential and a curvature of the protein surface region it represents. To find similar local regions of two proteins, a clique detection algorithm on an association graph is used. An association graph of two graphs is formed first by creating nodes from pairs of vertices, which two vertices are selected each from the two proteins, that have similar features. The edges of the association graph are created connecting nodes that have similar spatial distances between the pairs of original vertices belonging to the nodes. Next, the largest clique in the association graph, i.e. the largest fully connected subgraph,

is selected with a clique detection algorithm. The selected clique is considered as the most similar part between the two protein surfaces.

SURFCOMP also uses a clique detection algorithm on an association graph [65]. In SURFCOMP, surface critical points are considered as vertices. Each vertex is labeled as one of the three curvature classes: convex, concave, or saddle point. Rather than using all the vertices in a Connolly surface, using critical points reduces the number of vertices to be considered, making the method more efficient. The graphs are further simplified by several filters that compare surrounding shapes, local arrangement of the critical points, and physicochemical properties.

Baldacci et al. [66] further reduce the number of graph nodes by considering surface patches. A patch in a protein surface is a local circular region where the included residues have homogeneous geometrical and physicochemical properties. The properties considered are geometrical curvature, electrostatic potential, and hydrophobicity. Each patch contains at least ten amino acids and typically a protein surface is represented by less than ten patches. Neighboring patches are connected by edges, representing a protein by a spatial graph. They used the spatial graphs for classifying proteins by similarity of patterns of patches.

2.4.2 Geometric hashing

Wolfson and Nussinov [67] apply the geometric hashing technique, which was originally developed for computer vision applications. The method first extracts sparse critical points defined at the centers of mesh faces abstracted as convex, concave, or saddle of the protein surface [68]. The geometric hashing is composed of two stages, a hashing stage and a recognition (matching) stage. In the hashing stage, transformation-invariant information describing the protein surface shapes to be compared (called models) is extracted and stored in a hash table. Concretely, a protein surface shape represented by critical points is placed relative to every possible admissible reference frame and its position and features are stored in the hash table.

This stage can be executed off-line and the table can be reused once created. In the recognition stage, a query protein surface is placed relative to every possible reference frame and the hash table is accessed to find matching model critical points. Then a vote is registered for each pair of model and target reference frames if their critical points match. Geometric hashing allows partial surface matching, and a query protein can be compared with multiple proteins at the same time once they are hashed in a table. Later, they also applied geometric hashing for protein-small ligand molecule docking, and protein-protein docking [69, 70].

2.4.3 Methods using series expansion of 3D function

Usage of mathematical transformation has become popular in protein surface analyses as well as in 3D object analysis. In methods using mathematical transformation, a protein surface is treated as a 3D function, which is expanded as a series. A major advantage of these methods is the compactness of the description, which allows rapid comparison against a large number of proteins. A series expansion is also suitable for changing resolutions of surface descriptions. Another advantage is that properties on a surface can be naturally incorporated in a surface description.

An early work in this category uses a Fourier series expansion as a shape descriptor [71]. Protein surfaces are superimposed and Fourier coefficients are extracted and compared at various resolutions. Gerstein used the method to compare shape of antigen-combining sites of antibody molecules.

Thornton and her colleagues [72] used spherical harmonics to describe the volume of ligand binding pockets of proteins. A ligand binding pocket in a protein surface is detected using the SURFNET program, which identifies a pocket by inserting spheres of a certain size. Thus a pocket is represented as overlapping spheres, which fill the volume of the pocket. Then the spherical harmonic expansion is applied to the volumetric representation of the pocket and the coefficients are taken as the descriptor. An interesting application of their approach is the direct comparison of pocket shapes

with the corresponding ligand molecules, which is possible because both pockets and ligands are represented as a closed volumes. For comparison, shapes should be pose normalized.

Spherical harmonics also have been applied to protein-protein docking prediction [73]. By using spherical harmonics, a complete search of docking conformations over all six degrees of freedom can be conveniently performed by rotating and translating the initial expansion coefficients.

A 3D Zernike descriptor, which is the descriptor that is used throughout this dissertation, is also a series expansion method. The most favorable features of the 3D Zernike descriptor, aside from its advantages originating from spherical harmonics, are its rotation invariance and applicability to non-star-like shapes. These two advantages are worth further attention and will be described extensively in sections 2.5.1 and 2.5.4.

2.4.4 Other methods

The volumetric difference method, which was originally developed for 3D object representation, has been applied for protein surface comparison. Masek et al. [74] defines molecular 'skins', which are a thin layer of voxels comprising the protein surface. The method compares shapes by computing the similarity of the maximum overlap between a pair of protein skins. Another volumetric difference method utilizes a genetic algorithm to find the optimal superimposition of protein surfaces or fragments of proteins [75]. The spin image representation also has been applied to identify structurally equivalent surface regions in two proteins [76].

Shentu et al. [77] proposed a local surface structure characterization method named context shape, which considers visible directions from critical points on a protein surface. The visible direction of a point on a surface is the direction from the point to the region outside the protein. A context shape of a critical point essentially describes the visible directions from the point to a surrounding sphere of a

given radius, which is not blocked by voxels occupied by the protein volume. The context shape is represented as a binary string of size around 1,000 with 1 to mark blocked and 0 to mark visible directions. They used this method to evaluate shape complementarities of two protein surfaces in protein-protein docking prediction.

Pawlowski and Godzik proposed a method which aimed to compare physicochemical features of protein surfaces, such as electrostatic potential and hydrophobicity [77]. Those features are mapped on a surrounding sphere of a protein by casting rays from the center of gravity of the protein to the sphere and mapping the feature of the surface that is intersected by each ray to the sphere surface that the ray ends at. The comparison of feature mapped spheres are done after the spheres are superimposed. As obvious from its design, this method does not compare shapes of proteins but only physicochemical properties, thus it can only analyze proteins of similar structure (e.g., proteins of the same family). Nevertheless this method is interesting as it can quantify the difference of properties on the surface.

There are several methods that combine surface shape information with residue or sequence information. As sequence motifs (e.g., the PROSITE database) or spatial arrangements of catalytic residues [78], are traditionally used in function prediction in protein bioinformatics, these methods can take advantage of accumulated knowledge of sequence-function relationships of proteins.

The SURFACE database stores a library of functionally important residues found at pocket regions of proteins [79]. In selecting functional sites of proteins, pockets in protein surfaces are identified by SURFNET and the residues which reside in the pockets are compared to functional motif databases including PROSITE. Two local sites are compared in terms of the root mean square deviation of positions of the superimposed amino acids.

A method developed by Binkowski et al. [80] utilizes the local sequence information of binding pockets and surface shape to predict the function of proteins. The local sequence of a pocket region is extracted by concatenating short sequences which

comprise the pocket. To compare the extracted local sequence, local sequence alignment by dynamic programming algorithm is performed.

2.5 3D Zernike descriptors (3DZDs)

The protein surface analysis methods in the following chapters are based on application of 3D Zernike descriptors (3DZD). A 3DZD is categorized as a projection based method using special kernel functions that is obtained through normalization of 3D Zernike moments. Canterakis first introduced 3D Zernike moments that combine a radial function with spherical harmonics to describe objects in a 3D Cartesian coordinate system [54, 81]. Novotni and Klein later applied 3D Zernike moments to construct rotation invariant descriptor of 3D objects [36, 82].

2.5.1 Characteristics of 3DZDs

3DZDs have several significant advantages with regards to the comparison of protein surfaces. First, 3DZDs represent proteins compactly allowing fast retrieval for real-time database search. Second, 3DZDs are rotation invariant, that is, protein structures need not be pose normalized for comparison. Related works, such as spherical harmonics for binding pocket and ligand comparisons by Thornton's group [83], need pose normalization because the methods are not rotation invariant. Pose normalization is problematic, especially in comparison of protein shapes, because proteins are almost globular and the principle axes are not robustly determined. Third, resolutions of descriptions of protein structures can be easily adjusted by changing the order of 3DZDs. Figure 2.1 illustrates five different resolutions visualized by protein surfaces reconstructed from five different orders of 3DZD: 5, 10, 15, 20, and 25. When a lower order is used, a pear-like global surface shape of protein 1ew0A is highlighted, a while more detailed description of local geometry shows up as the order becomes higher. Moreover, physicochemical properties of the protein surface, such as electrostatic potential and hydrophobicity, can be incorporated into the description.

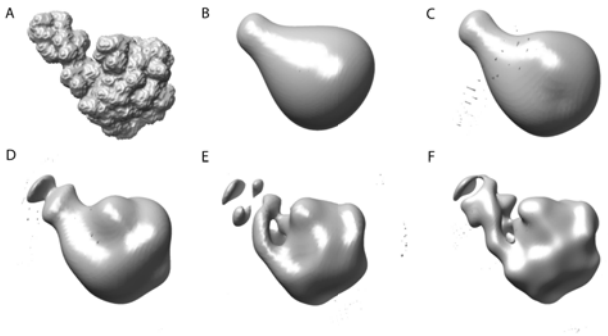

Figure 2.1. Reconstruction of 3D Zernike moments. **A**, surface abstraction of 1ew0A, which is used as the input. **B** through **F** are reconstructed figures of 3D Zernike moments using different order from 5 up to 25 increasing with an interval of 5.

2.5.2 Derivation of 3DZDs

Surface function $f(\mathbf{x})$ defined on 3D coordinate $\mathbf{x} = (x, y, z)$ is defined to be 1 (or a real number) if it is on the surface, and 0 if it is not on the surface of a protein. To obtain 3DZDs, a given 3D function $f(\mathbf{x})$ is expanded into a series in terms of the Zernike-Canterakis basis defined by the collection of functions

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

where $-l < m < l, 0 \leq l \leq n$, and $(n - l)$ is even non-negative integer. Index n is called the order of the descriptor, which determines the resolution of the descriptor; and indices m and l are determined by the value of n [54, 82]. Here $Y_{l}^{m}(\vartheta, \phi)$ are spherical harmonics and $R_{nl}(r)$ are radial functions defined by Canterakis constructed so that $Z_{nl}^{m}(r, \vartheta, \phi)$ become polynomials $Z_{nl}^{m}(\mathbf{x})$ when written in terms of Cartesian coordinates [54]. 3D Zernike moments of the input $f(\mathbf{x})$ are defined as the coefficients of the expansion of this orthogonal basis:

$$\Omega_{nl}^{m} = \frac{3}{4\pi} \int_{|\mathbf{x}| \leq 1} f(\mathbf{x}) \overline{Z}_{nl}^{m}(\mathbf{x}) d\mathbf{x}, \tag{2.2}$$

where the $\overline{Z}_{nl}^{m}(\mathbf{x})$ is the complex conjugates of $Z_{nl}^{m}(\mathbf{x})$. The rotational invariant 3DZD, F_{nl}, is obtained as the norms of vectors Ω_{nl}^{m}:

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

The length of the descriptor is dependent on the order n and is computed as $(\frac{n}{2}+1)^2$ when n is even and $\frac{(n+1)(n+3)}{4}$ when n is odd. $n = 20$, which yields a total of 121 invariants (coefficients), is shown to provide sufficient accuracy in the previous work of shape comparison [82]. For further details readers are referred to papers by Canterakis [54] and Novotni and Klein [82].

Final normalization of 3DZDs is an optional step. Two normalizations are newly introduced in the thesis. The first normalization method normalizes each invariants in a 3DZD by the sum of the total numbers of the descriptor. This normalization is found to reduce the rotation error which will be shown in section 3.2.1. The second normalization divides each number in 3DZD by the norm of the descriptor vector making a unit vector descriptor. This normalization is found to reduce the dependency of 3DZD on the number of voxels used to represent a protein, i.e., further reduce the dependency on protein size.

2.5.3 Computing distances of 3DZDs

There are three distances used to compare 3DZDs. Manhatten distance (MHT) is defined as

$$MHT = \sum_{i=1}^{on} (Z_{Ai} - Z_{Bi})^2, \tag{2.4}$$

where on is length of order n 3DZD vector. Manhatten distance is also known as L^1 norm.

The Euclidean distance (EUC) is defined as

$$EUC = \sqrt{\sum_{i=1}^{on} (Z_{Ai} - Z_{Bi})^2}, \tag{2.5}$$

where Z_{Ai} and Z_{Bi} are the ith number of 3DZD of protein A and B, respectively. The Euclidean distance is also know as L^2 norm.

The correlation coefficient based distance (CC), is defined as

$$CC \;=\; 1 \;-\; r(Z_A, Z_B), \tag{2.6}$$

where r is the correlation coefficient of two 3DZDs. CC is 0 when two 3DZDs correlate perfectly. All or one of the distances is used in the following chapters.

2.5.4 3DZD and spherical harmonic descriptor

In this section, discussion is made on 3DZDs mainly in comparison with the spherical harmonics descriptors (SHD) [52, 84], which are a popular spherical harmonics-based projection techniques used for general 3D object comparison. Projection-based techniques have been used extensively in two-dimensional (2D) image analysis and pattern recognition [85–88]. In particular, 2D Zernike moments have proved exceptionally useful for the analysis of 2D shapes arising in many areas ranging from face recognition, [89] cell part recognition [90], and optical scattering pattern recognition for identifying bacterial colonies [91]. Yeh et al. [92] applied 2D Zernike moments to protein 3D structure retrieval by characterizing a structure with a set of 2D projections from 100 different directions. Finally, Canterakis was able to extend 2D Zernike polynomials and moments to 3D, introducing 3D Zernike-Canterakis polynomials [54]. Later, rotationally invariant descriptors based on Zernike-Canterakis moments were explored for 3D shape retrieval by Novotni and Klein, [82] who reported improved precision-recall curves at a lower storage cost when compared with SHD. For SHD no radial distance parameter is used; rather, the 3D space is sampled into concentric spherical shells around the center of mass. Then, a volume of a target object within each concentric sphere of a radius r centering at the center of mass of the object, $f_r(\vartheta, \varphi)$, is expanded in the series of spherical harmonics, $Y_l^m(\vartheta, \varphi)$

$$f_r(\vartheta,\varphi) \;=\; \sum_l f_r^l(\vartheta,\varphi) \;=\; \sum_l \sum_{m=-l}^{l} c_{r,l}^m Y_{r,l}^m(\vartheta,\varphi) \tag{2.7}$$

where $c_{r,l}^m$ is the expansion coefficients that can be obtained through multiplying the above equation by the spherical harmonics and integrating over the defined angles. Spherical harmonics differ under different orientations, (ϑ,φ). However, since the L^2 norm of the function is rotation invariant, a rotation invariant signature for $f_r(\vartheta,\varphi)$ is constructed as the collection of L^2 norms of $f_r^l(\vartheta,\varphi)$ at each l, that is, $\{\|f_r^0\|, |f_r^1\|, \dots\}$. Finally, collecting the signatures for each radius, r, will give the SHD of a protein structure.

The 3DZDs genuinely belong to the 3D realm, while SHD are essentially a combination of 2D descriptors. Indeed, note that SHD measures similarity of objects by comparing them shell-wise. There are quite a few practical implications of this fact: (1) SHD does not capture object coherence in the radial direction, thereby incorporating less object characteristic information [36]. For example, since the descriptors for each shell are calculated separately, the shells can be rotated independently by random angles without changing the resulting descriptors. (2) The orthonormality of the Zernike-Canterakis basis results in less information redundancy. One should note that in SHD, descriptors coming from adjacent shells are highly correlated, making them redundant to some extent. Indeed, using 154 3DZDs (max order 21) yields better retrieval results than using 928 SHDs (32 shells, 29 descriptors per shell) as tested on the Princeton Shape Benchmark, which is a database of general 3D objects such as airplanes and chairs [36, 82, 84]. (3) SHDs require polar sampling, which was pointed out to be problematic for the robustness of rotation invariancy [55]. Securing robustness of SHD requires a distance field-based voxelization procedure where voxels are assigned continuous values between 0 and 1. On the other hand, the Zernike-Canterakis basis consists entirely of polynomials in Cartesian coordinates, thus avoiding polar sampling, and making possible to treat all voxels in the model on equal footing. In addition, 3DZDs show optimal performance when simple binary

voxelization is used [36, 82]. Because a sizeable amount of the computational time is consumed by the voxelization process, this simplicity results in faster response times for user-search engine transactions. (4) One can naturally add other protein surface properties within the 3D Zernike framework. For example, to add electrostatics, it is enough to calculate 3DZDs of $f(\mathbf{x})$ set equal to the electrostatic potential value on the surface, and zero otherwise. This is not as straightforward with the SHD, because of the aforementioned robustness problem.

2.6 Chapter summary

In this chapter, we reviewed methods for 3D object shape analysis in the context of protein shape analysis.

Protein surface analysis is especially difficult because most of proteins have a more or less sphere-like shape. Therefore a descriptor needs to differentiate relatively small differences. For the same reason, typical pose normalization methods, such as PCA, do not give a unique solution. Non-shape features should also be considered, such as physicochemical properties and residue conservation, as they are important in understanding protein function. In addition, protein comparisons should be fast enough for a real-time database search.

To meet these requirements, we chose 3D Zernike descriptor (3DZD), which will be extensively studied in the following chapters, for protein surface comparison . Conventionally, proteins have been long analyzed and classified in terms of their sequences and main-chain conformations. However, there are cases where these methods are not capable of detecting similarities and dissimilarities in a biologically meaningful way. In such cases, the 3DZD-based surface analysis can often do a better job than these methods, as it captures global and local protein surface shapes, which are directly responsible for biological function, and it also is able to quantify similarity of physicochemical properties.

3 PROTEIN TERTIARY STRUCTURE RETRIEVAL BASED ON GLOBAL SURFACE SHAPE SIMILARITY [1]

Characterization and identification of similar tertiary structure of proteins provides rich information for investigating the function and evolution. The function of a protein of unknown function can be predicted by searching for proteins of similar structures that has already been characterized with a function. As the number of known structures increase, the speed as well as the accuracy of the searching process becomes important. A crucial drawback of conventional protein structure comparison methods, which compare structures by their main chain orientation or the spatial arrangement of secondary structure, is that a database search is too slow to be done in real-time.

In this chapter, structure comparison is performed using a global surface shape represented by 3D Zernike descriptors (3DZDs), which have been discussed in the previous chapter. With this simplified representation, a search against a few thousand structures takes less than a minute. To investigate the agreement between the surface representation defined by 3DZD and conventional main-chain based representation, a benchmark was performed against a protein classification generated by the combinatorial extension algorithm (CE) [29]. Despite the difference in representation, 3DZD retrieved proteins of the same conformation defined by CE in 89.6% of the cases within the top five closest structures.

This chapter is organized as follows: first, differences between 3DZD and the other projection-based methods are extensively discussed. Next, we report the results of our benchmark on the performance in protein structure search using a large dataset of 2432 proteins. The overall results show a good agreement with the structural

[1] This chapter is a reuse of published work [93]. The main contributors of each section of the work, if it is not L. Sael, is mentioned in the footnotes of the corresponding sections or subsections.

classification made by CE program [29], which compares main chain orientations of proteins, despite the difference in view of the protein shapes by the two methods. We also compared 3DZD with another standard protein structure comparison method, DALI [30]. Finally, differences between CE and 3D Zernike are discussed, emphasizing the advantage of 3DZD. The effect of shape comparison at different resolutions is also discussed.

3.1 Materials and methods

3.1.1 Benchmark dataset

The benchmark dataset of protein structures consists of 2432 protein structures classified into 185 fold groups. These are a subset of structures extracted from the structure comparison results of the CE program [29] (`ftp://ftp.sdsc.edu/pub/sdsc/biology/CE/db/ata_3_8.txt`). Note again that the structure representations of CE and 3DZD are fundamentally different: the former considers a protein structure as the spatial position of main-chain residues and the latter represents a protein structure as a surface shape. The purpose of this benchmark study is to investigate the extent of similarity between the two methods. If we observe a significant agreement between the two methods, then 3DZD can not only be used for comparing similar surface shapes but also can be an effective tool for fast searching of protein structures with a similar main-chain orientation (i.e., a conventional sense of protein structure similarity, which also implies evolutionary relationship). On the other hand, it is also expected that interesting cases can be found, such as cases where two proteins share a similar surface shape but different main-chain orientation.

The CE is one of the standard programs for protein main-chain comparison that classifies proteins solely by geometrical aspect of proteins without consideration of evolutionary relationship as, for example, the SCOP database [94] does. Given two protein structures, CE first identifies eight residue-long fragments with similar conformation in the two proteins by comparing corresponding distances of pairs of residues

within each fragment. Then, identified fragment pairs from the two proteins are combined to find larger structurally similar regions by comparing corresponding interfragment distances. Dynamic Programming (DP) is used for the calculation, thus fragment pairs are combined in a sequential order from the N-terminus to the C-terminus.

A description of the procedure that is used to select the benchmark proteins follows. The original CE database consisted of 50,246 protein structures classified into 7,386 fold groups. Each fold group consists of a "representing" protein, a set of "represented" proteins that satisfy several similarity criteria against the representing protein, and a set of "similar" proteins has weak similarities to the "representing" protein. First, starting from the CE database, separate fold groups are merged if the structure of their "representing" proteins is sufficiently similar, i.e. having a Z-score of 3.8 or higher by CE. The Z-score of 3.8 is recommended by the authors of the database to filter out random similarities. Next, the set of "similar" proteins are eliminated from a fold group. Then, "represented" proteins are eliminated from a fold group if the size is more than 12.5% different in length from the "representing" protein, or if the quality of the structure is not appropriate: structures that lack coordinates of more than 10 residues, or structures that do not have side-chain coordinates. Small proteins that have less than 100 residues are also eliminated. In addition, structures that have coordinates of hydrogen atoms of more than 3% of residues are filtered out, because they significantly affect surface shape of the protein. Finally, groups that only contain three or fewer "represented" proteins (and "representing" protein) are removed.

3.1.2 Building protein surfaces and generating 3DZD inputs

The first step of computing a 3DZD of a protein is to define the protein surface in 3D space. To begin, atoms in the PDB file that are not part of the protein sequence are removed. Then, the MSROLL program in the Molecular Surface Package version

3.9.333 is used to compute the Connolly surface (triangle mesh) of the protein using default parameters [95]. Next, the triangle mesh is placed in a 3D cubic grid of N^3 (N=200 used), compactly fitting the protein to the grid. Each voxel (a cube defined by the grid) is assigned either 1 or 0; 1 for a surface voxel is located closer than 1.7 grid interval from any triangle defining the protein surface, and 0 otherwise. The resulting protein surface has thickness of approximately 3 grid intervals. The inside of a protein is kept empty so that 3DZD focuses on capturing the surface shape of a protein.

3.1.3 DALI protein structure comparion program

In addition, we also run the DALI algorithm [30] against the CE based benchmark dataset. DALI is another widely used protein structure comparison algorithm that was established in 1993. DALI compares two protein structures by comparing the 2D residue-residue (Cα-Cα) distance maps of the two proteins. First, DALI identifies similar structural fragments of a fixed size (generally 6 residues long) between the two proteins by picking up pairs of similar subdistance maps from the two distance maps. This step captures local regions of the two proteins that have a similar residue contact pattern. Next, the algorithm combines identified pairs of similar subdistance maps to find all sub-structures that are similar between the two proteins. We used the stand-alone program of the DALI algorithm, DaliLite [96], which is available for download at `http://www.ebi.ac.uk/DaliLite/`.

3.1.4 Benchmark procedure [2]

For each query protein, the proteins in the dataset are ranked by their descriptor distances to the query protein. For a given distance threshold value, the sensitivity and the specificity are averaged within a group, then again averaged among all groups

[2]Benchmark procedure was implemented by B. Li.

to give a final value in the plots (see Fig. 3.3). The sensitivity and the specificity are defined as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.1}$$

$$Specificity = \frac{TP}{TP + F} \tag{3.2}$$

where TP, true positive, is the number of proteins in a fold group of the query protein retrieved with a distance closer than the threshold; FN, false negative, the number of proteins in the fold group of the query protein not retrieved with a distance larger than the threshold; FP, false positive, is the number of proteins that are not included in the fold group of the query protein but incorrectly retrieved in the search. Thus, the denominator in Eq. 3.1 is the total number of all members of the fold group. The denominator in Eq. 3.2 is the total number of proteins retrieved above the threshold.

3.2 Results

3.2.1 Examples of 3DZD

Figure 3.1 shows 3DZD examples of two proteins, triosephosphate isomerase (PDB code: 7tim, A chain) and interleukin-4 receptor A-chain (1iarB). Globally, 7timA has a round-shaped surface and 1iarB is an L-shaped structure [Fig. 3.1**A**]. This apparent difference of their global surface shape is reflected by distinctive 3DZDs shown in Figure 3.1**B**. The difference in the overall shape tends to appear in the first couple of orders of the descriptor, resulting in a relatively large Euclidean distance of 38.84, and correlation coefficient-based distance of 0.656.

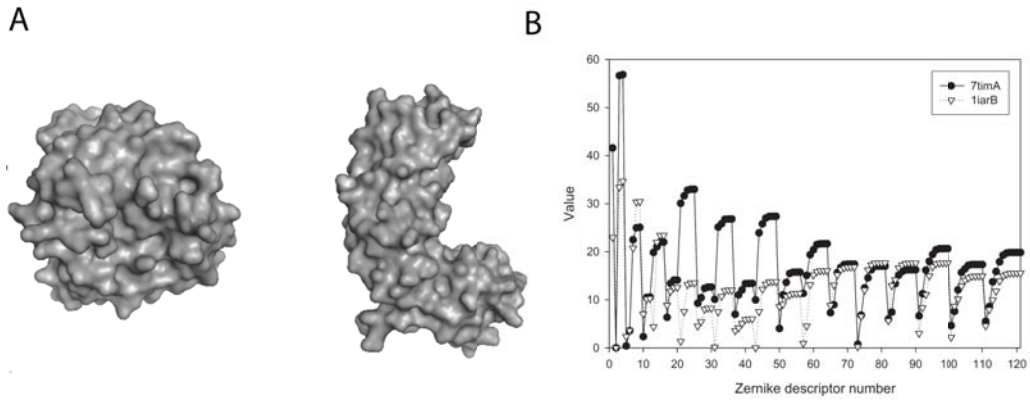

Figure 3.1. 3DZD of two example proteins. **A**, The global surface shape of the two proteins; 7timA (left) and 1iarB (right). **B**, 3DZD of the two proteins.

3.2.2 Rotation invariance test

As described earlier and in the Methods section, 3DZD are rotationally invariant. This is one of the largest advantages of 3DZDs. However, in practice, 3DZDs of rotated protein structures are not identical. A major source of error is introduced when a protein surface is discretized into voxels. We found that in computing all the distance measures, i.e. Manhattan (MHT) [Eq. 2.4], Euclidean (EUC) [Eq. 2.5] and correlation coefficient-based (CC) [Eq. 2.6], normalizing each number in a 3DZD by the sum of the 121 numbers of the descriptor gives the best error reduction among tested methods.

Figure 3.2 shows an example of the variance in 3DZDs upon rotation. Here, each of the two proteins used in Figure 3.1 is rotated and EUC and CC to the surface at original orientation are computed. The two distances are the top two performing functions in our protein shape search benchmark (see the next section). As for EUC [Fig. 3.2**A**], approximately 90% of the rotated structures stay within the distance of 10. In the case of the CC [Fig. 3.2**B**], approximately 90% of the rotated structures of the two proteins have less than a distance of 0.03. From this experiment, we can draw a threshold of the significance of the distance, or in the other words,

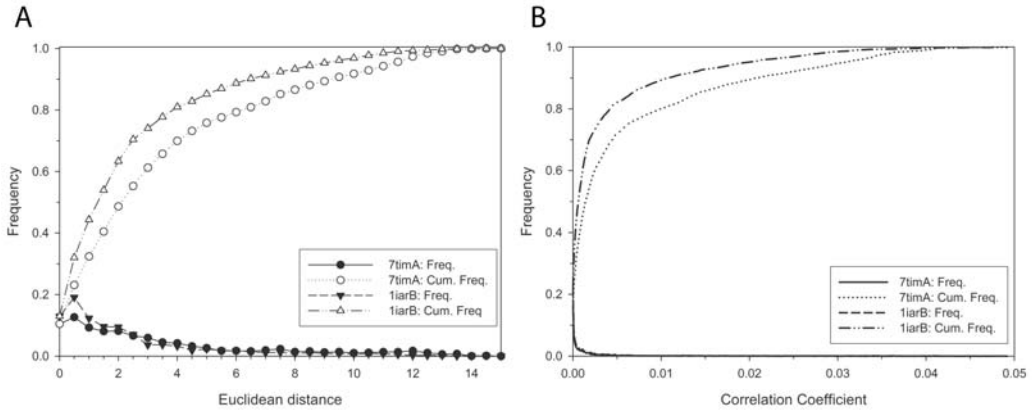

Figure 3.2. Variance of 3DZDs upon rotation of proteins. All the possible rotated positions of two protein structures, 7timA and 1iarB, in three orthogonal directions with a step size of 30 degrees are computed. Histograms of distances between each of the rotated structures and the original one are shown. **A**, The EUC is used. Filled (empty) circles, the frequency (the cumulative frequency) of the EUC of 7timA are plotted. Filled (empty) triangles, the frequency (the cumulative frequency) of EUC of 1iarB are plotted. **B**, The CC is used. The frequency and the cumulative frequency of distances of 7timA and 1iarB are shown by solid line, dotted line, dashed line, dash-dot-dot line, respectively.

determine an "invisible" range of the 3DZD. For example, if two proteins have an EUC of less than 10, these proteins can be considered significantly similar, or more precisely, indistinguishable from cases where the two proteins are identical but placed in a different orientation. To improve the rotation invariance of the 3DZD, different thicknesses of the surface representation were tested, as well as a continuous value assignment between 0 and 1 to surface voxels rather than the binary voxelization, but there were no differences in the performance.

3.2.3 Comparison of the CE and the SCOP classifications on the benchmark dataset

The CE benchmark dataset is compared with the SCOP protein classification database [97] to understand the nature of protein structure classification. A unique feature of the SCOP classification is that the evolutionary relationships of proteins are also taken into account by a manual curation process, together with the protein structure similarities. Thus, SCOP has been serving as an indispensable resource for elucidating relationships of protein structure and function.

In Table 3.1 column **A**, the 185 fold groups in the CE benchmark dataset are compared with the SCOP superfamily classification. The number of CE-fold groups that overlap with a SCOP superfamily by a certain fraction is counted. When a CE-fold group overlaps with several SCOP superfamilies, the SCOP superfamily that gives the largest overlap is counted. The number of SCOP superfamilies that correspond to the CE benchmark dataset is 150, which is smaller than the number of the CE-fold groups. It is found that only 75.1% of the CE-fold groups correspond to one SCOP superfamily. Table 3.1 colunm **B** shows that 82.2% of the CE-fold groups correspond to one SCOP fold. The overlap with SCOP folds looks larger than SCOP superfamilies, because the average size of a SCOP fold is larger, i.e., it is more frequent that multiple CE-fold groups correspond to a SCOP fold. The number of SCOP folds that correspond to the CE benchmark is 117. These results illustrate that even the widely used protein structure comparison method, CE, does not have perfect correspondence

Table 3.1
Comparison between the CE based benchmark dataset and the SCOP database

Overlap[a]	The number of CE groups (%)[b]	
	(A) SCOP superfamily	(B) SCOP fold
0.0∼0.1	1 (0.5)	1 (0.5)
0.1∼0.2	0	0
0.2∼0.3	0	0
0.3∼0.4	1 (0.5)	0
0.4∼0.5	5 (2.7)	1 (0.5)
0.5∼0.6	15 (8.1)	10 (5.4)
0.6∼0.7	5 (2.7)	4 (2.2)
0.7∼0.8	7 (3.8)	8 (4.3)
0.8∼0.9	11 (5.9)	8 (4.3)
0.9∼1.0	2 (1.1)	2 (1.1)
1.0	139 (75.1)	152 (82.2)

[a] The fraction of members of a fold group in the CE-based benchmark dataset that overlap with a superfamily in SCOP. When a CE-fold group corresponds to multiple SCOP superfamilies, a SCOP superfamily that gives the largest overlap with the CE-fold group is used to compute the fraction. [b] The percentage among the 185 CE-fold groups.

with a well-established protein structure classification database, SCOP. A recent work by Sierk and Pearson [98] provides a further benchmark for protein structure comparison methods. Therefore, the aim of the structure retrieval performed in the next section using 3DZD and DALI on the CE benchmark dataset is to understand the similarity and dissimilarity of the three methods, not to evaluating "accuracy" of a particular method.

3.2.4 Retrieval performance of 3DZD and DALI against CE classification

Figure 3.3 shows the sensitivity and specificity plot of the benchmark performance on the dataset of 2432 proteins. Results of the 3DZD with and without prescreening by the length of the proteins are also shown. When the prescreening is used, a protein in the dataset is compared with a query protein only when its length is in the range between 57% to 175% of the query protein length. The three different distance measures, namely, Euclidian (EUC), Manhattan (MHT), and the correlation coefficient-based (CC) are compared. First, regardless of the prescreening, the results are far better than random. Second, among the three distance measures, the performance of MHT is somewhat worse than the other two distance measures, but all three distance measures essentially showed similar performance. Third, it shows that the prescreening is effective in improving the search performance. This is because the scale is normalized so that a structure fits in a unit sphere when computing 3DZD, hence the size information is lost.

Table 3.2 summarizes the search results of 3DZDs with the length-based prescreening. More than 89.0% of proteins retrieved another protein in the same CE-fold group within the top five closest structures. Those successful proteins are not biased to specific types of protein folds, because the successful proteins are distributed among approximately 98% of the fold groups (Top 5, Group 1). When Top 10 hits are considered, 93.1% of the proteins successfully retrieved their CE-fold group member by using the EUC or the CC. The search is successful for at least one protein in almost all the fold groups (99.5% by using the EUC) considering the Top 10 hits. On the other hand, approximately half of the fold groups contain some members that could not retrieve a fold group member within Top 10 (Top 10, Groups All). These are protein structures that are judged to be similar by the main-chain orientation but not by the surface shape. Examples of these cases are given in the next section.

The structure retrieval results by DaliLite are also shown in Table 3.2. Interestingly, only 28.6% of the proteins retrieved another protein in the same CE-fold group

Table 3.2
Summary of the structure retrieval using the three distance measures and DaliLite

	Top 5			Top 10				
	Proteins[a]	Groups 1[b]	Groups All[c]	Proteins	Groups 1	Groups All	Avg. rank[d]	Avg. dist[d]
EUC	2179 (89.6)	182 (98.4)	72 (38.9)	2264 (93.1)	184 (99.5)	91 (49.2)	9.79	8.31
MHT	2165 (89.0)	181 (97.8)	70 (37.8)	2257 (92.8)	183 (98.9)	88 (47.6)	10.07	80.04
CC	2176 (89.5)	183 (98.9)	71 (38.4)	2265 (93.1)	183 (98.9)	92 (49.7)	10.79	0.02
DaliLite[e]	696 (28.6)	85 (3.5)	1 (0.0)	897 (36.9)	104 (4.3)	1 (0.0)	183.07	24.86
Random[f]	508 (20.9)	89 (48.1)	0 (0.0)	806 (33.1)	122 (65.9)	1 (5.4)	87.17	0.14

[a] The number of query proteins that retrieved a correct member in the same group as the first position, within Top 5 or Top 10. In the parentheses, the percentage among all the 2432 proteins in the benchmark set is shown. [b] A group is counted if at least one member in the group successfully retrieved another member in the group as the first position, within Top 5, or Top 10. In the parentheses, the percentage among all the 185 groups in the benchmark set is shown. [c] A group is counted only if all the members in the group successfully retrieved another member in the group as the first position, within Top 5 or Top 10. [d] The average rank/distance of the closest structure judged by the distance metric to the query. The average is first averaged within a group, then averaged across the groups. [e] DaliLite Version 2.4.4 was used. The distance d is defined as d 5 100 2 (the structure similarity Z-score by DaliLite). [f] A random value between 0 and 1 is assigned as the distance between the query to each protein.

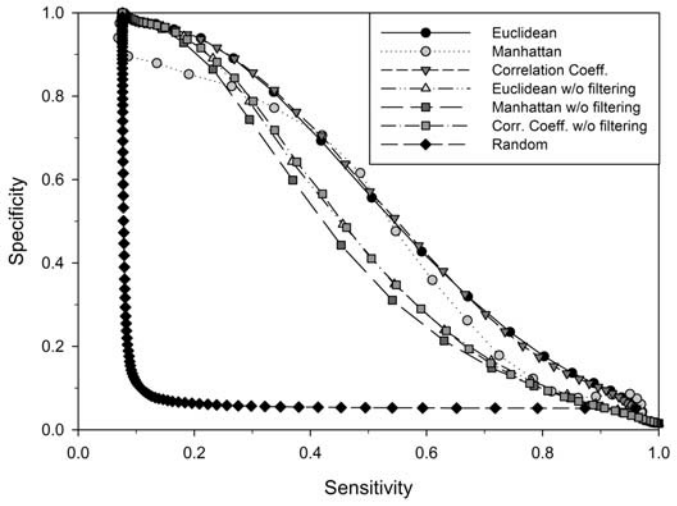

Figure 3.3. The sensitivity and the specificity of the benchmark dataset are plotted using the three distance definitions of 3DZD: the Euclidean, the Manhattan, and the correlation coefficient-based distance with and without prescreening by the sequence length. When the prescreening is used, a protein in the dataset is compared with a query only when its length is in the range of 57%∼175% of that of the query protein.

within Top 5 by using DaliLite. Which shows that there is disagreement between the CE and DALI methods. Although there are disagreements between the structural classifications such that it is controversial to argue that the 3DZD method is a good structure comparison method by showing strong agreement with only one method, the strong agreement between the retrieval results of the 3DZD method on the CE classification shows that 3DZD method is an efficient alternative to CE method.

3.2.5 Examples of individual 3DZD retrieval cases[3]

Figure 3.4 illustrates the difference between CE and 3DZD. Figure 3.4 **(A,B)** shows protein structure pairs that are identified to be significantly similar by CE, but not by 3DZD. They are evolutionarily close and thus classified into the same

[3]Examples of proteins that have disagreement between 3DZD and CE was chosen by D. La.

family in CATH and SCOP. In these two examples, a small portion of the secondary structure elements of the protein is flipped out (figure on the right of **A** and **B**) from the mass of the protein, resulting in the change of the surface shape. In contrast, Figure 3.4 **C** and **D** demonstrate two instances in which 3DZD detects similar global surface shape of proteins with a different overall fold. Figure 3.4**C** is a vivid example of two proteins that have a very similar surface shape but with completely different secondary structure elements, where the left structure is a β class protein and the right structure is an α class protein. Figure 3.4**D** is a protein pair with a different topology (in CATH) forming a very similar surface shape. The results of structural comparison performed by CE and 3DZD between the pairs of proteins in Figure 3.4 are shown in Table 3.3.

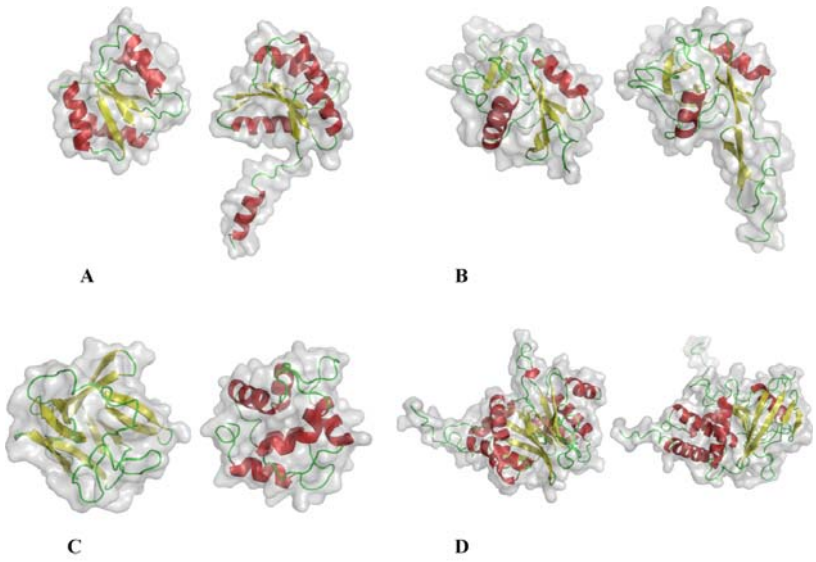

Figure 3.4. Examples of protein pairs that have the same main chain orientation but with different surface shapes **(A, B)** and pairs that have similar surface shapes but with different main chain orientation **(C, D)** are shown. **A** shows 1dz3A (response regulator SPO0A) and 1mb0A (response regulator DIVK). **B** shows 1jznA (galactose-specific C-type lectin) and 1g1qA (P-selectin lectin). **C** shows 1barA (fibroblast growth factor) and 1rro (oncomodulin). **D** shows 1rypB (proteasome subunit) and 1gwz (Tyrosine phosphatase).

Table 3.3
Sructural comparison performed by CE and 3DZD between the pairs of protein in Figure 3.4

	PDB ID		CE			3DZD		
	left	right	RMSD(Å)	Z-score	SeqID(%)	EUC	MHT	CC
A	1dz3A	1mb0A	1.6	5.0	26.9	51.21	438.53	0.620
B	1jznA	1g1qA	2.0	5.9	23.5	52.67	431.16	0.602
C	1barA	1rro	6.7	1.6	3.6	12.66	101.85	0.031
D	1rypB	1gwz	5.0	2.3	9.7	12.73	108.89	0.041

By taking advantage of the 3DZD's ability to find proteins with similar overall protein surface shape, functionally related proteins can be retrieved beyond sequence similarity and significant backbone conformation similarity. Figure 3.5 shows several such examples. Associated Table 3.4 gives detailed data for the proteins in Figure 3.5. Figure 3.5**A** is a pair of DNA topoisomerase I from human and Escherichia coli. The characteristic role of the proteins is to capture DNA double strands. The sequence identity between the two proteins is very low, and the CE only aligns 17.3% of the two proteins. In contrast, 3DZD identifies the overall surface shape with a significant distance (compare the 3DZD with the average distance of the top hit in the benchmark, the right columns in Table 3.2). Figure 3.5**B** shows two DNA binding proteins. Both proteins bind to DNA with the curved U-shaped region. These two proteins have different functions but both have the characteristic surface shape of the DNA binding region that is captured by 3DZD. Figure 3.5**C** is another pair of proteins. These two proteins bind to DNA with their long tail regions. Note that SCOP classifications of these three pairs are also different from each other. Figure 3.5**D** shows a pair of subunits in membrane protein complexes. 2nwl is a subunit of glutamate transporter, which is a pentamer, and 2bbh is a subunit of CorA Mg21 transporter, which is a trimer. In both the cases, two long helices penetrate the

membrane and form the scaffold of the transporters. The last example, Figure 3.5**E** is a pair of transmembrane proteins. In each case [Fig. 3.5**A** and **E**], the sequence identity between the pair of proteins is below 10%, and CE only aligns partial regions of the pair. In contrast, 3DZD captures the overall surface similarity of each pair in that is required to realize their biological function with significantly close distance like the donut shape Figure 3.5**A** and U-shape in Figure 3.5**B** of the DNA binding protein.

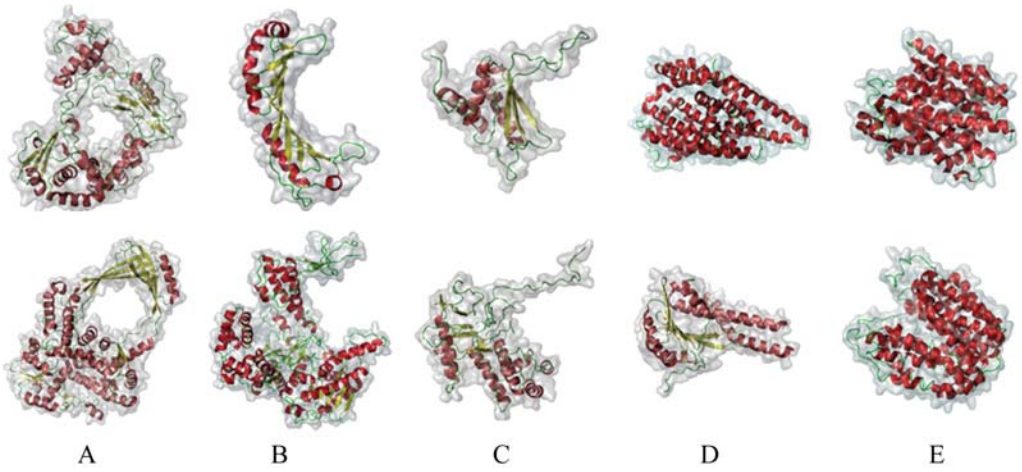

Figure 3.5. Examples of protein pairs whose surface shapes are judged to be similar by 3DZD. **A**, 1a31 and 1cy0 (from top(T) to bottom(B));**B**, 1tbp and 1t7p; **C**, 1b3t and 1adv; **D**, 2nwl and 2bbh; **E**, 2b2i and 2cfp. Detailed data of these protein pairs are shown in Table 3.4.

3.2.6 Resolution of 3DZDs

As described in section 2.5.1, one characteristic of 3DZDs is that the resolution of a shape can be altered by changing the order of the 3DZDs. In Figure 3.6, two different orders [the index n in Equations 2.1 $\sim$ 2.3], 5 and 20, are used to compute similarity (EUC) of the 16 proteins selected from different CE-fold groups. Altering

Table 3.4
Pairs of proteins that have similar surface shape defined by 3DZD

	PDB ID		SCOP classification		Length(aa)		SeqID	CE			3DZD[b]		
	T	B	T	B	T	B	(%)[a]	RMSD	Z-score	Align[c]	EUC	MHT	CC
A	1a31	1cy0	d.163.1.2	e.10.1.1	457	534	5.8	6.3Å	3.9	79(17.3)	5.58	49.9	0.001
B	1tbp	1t7p	d.129.1.1	e.8.1.1	180	662	2.0	4.9Å	4.4	64(35.6)	7.25	58.6	0.08
C	1b3t	1adv	d.58.8.1	g.51.1.1	147	287	9.0	6.7Å	1.6	64(43.5)	7.65	69.4	0.28
D	2nwl	2bbh	N/A[d]	d.328.1.1	422	244	5.7	8.1Å	2.3	88(36.1)	6.04	53.6	0.001
E	2b2i	2cfp	N/A	f.38.1.2	399	417	7.8	4.9Å	4.4	102(25.6)	7.28	58.6	0.08

[a] The sequence identity between the two proteins. [b] The three distances (EUC, MHT, and CC) of the 3DZD. [c] The percentage of the aligned residues relative to the shorter protein among the two. [d] Not included in the current SCOP (ver. 1.73).

the order of descriptors changes the distances of proteins [e.g., the EUC of 1theB to 1o0eA is 28.74 in Fig. 3.6**A**, which is 13.36 in Fig 3.6**B**]. Also, the 3DZD distance of a pair changes, which is obvious from the difference in the length of and topology of the two trees [Fig. 3.6**A,B**]. When order of five is used [Fig. 3.6**A**], emphasis is given to describe overall shapes, such as spherical, cylinder-like, or tadpole-like shapes. With order of 20, clusters made by using the order of five are further decomposed with some topology changes [Fig. 3.6**B**]. To highlight the decomposition of clusters between the two trees, clusters of proteins within the EUC Z-value of 0.35 are shaded by the same color. The Z-value of the EUC using the order of 20 and 5 is computed using the average and the standard deviation of the distribution of distances of protein pairs in the CE benchmark dataset. Reducing resolution also contributes to reduction of the search speed, because the descriptor becomes more compacts. 121 invariants are used in a descriptor when the order is set to 20, and this decreases to 12 when the order is set to 5.

3.2.7 Computation time

The 3DZD allows rapid real-time search on the web, because a protein structure is compactly represented by 121 numbers (order n=20). If a query protein is already transformed into 3DZD, a search to the current benchmark dataset takes less than a second on an Intel Pentium 4 3.0 GHz processor with 2 GB of memory (Table 3.5).

When a protein strucuter is input as the query, the following steps must be performed before the database search: solvent accessible surface triangulation, surface voxelization, and transformation into 3DZD. Taken together with the database search, this entire process takes less than a minute. Because enlarging the database to be searched only affects the execution time of the database search step, a search against the entire PDB (as of August 2007) with 45,000 structures will only take a minute. The search speed can be further increased if the database is prescreened by the length of the query protein. Note that a pairwise structure comparison by CE typically takes

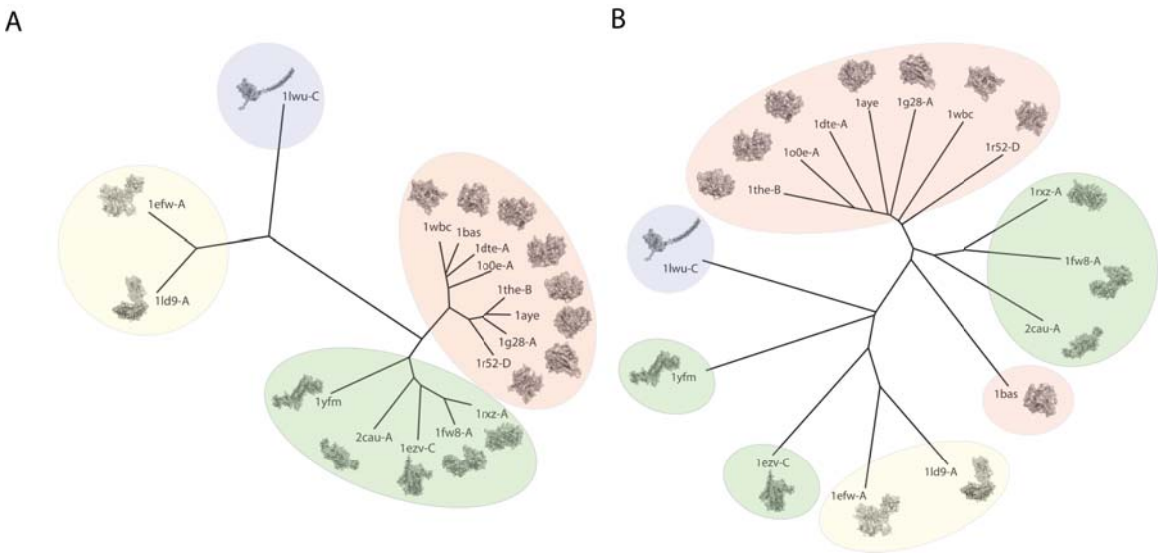

Figure 3.6. Resolution of 3DZDs. In **A** order of five and in **B** the order of 20 are used to construct trees that represent similarity of the surface shape of 16 proteins: 1theB, 1o0eA, 1dteA, 1aye, 1g28A, 1wbc, 1r52D, 1rxzA, 1fw8A, 2cauA, 1bas, 1ld9A, 1efwA, 1ezvC, 1yfm, and 1lwuC. The colors represent proteins in the same cluster in the tree constructed by using the order of five that have distance Z-value within 0.35. The Phylip package [99] Fitch-Margoliash method is used to construct the trees using EUC. The length of the branches connecting two proteins represents the distance between them. The distance between 1theB to 1aye and 1theB to 1o0eA in **A (B)** are 15.95 (17.57) and 28.74 (13.36), respectively.

a couple of seconds. Thus, a database search against PDB using CE would take more than a day.

Table 3.5
Execution time (in seconds)

Grid size[a]	64^3 (voxels)	200^3 (voxels)
Surface triangulation[b]	21	21
Surface voxelization	1	3
3DZD transformation	1	16
Database search[c]	0.43	0.46
Total	24 (s)	41 (s)

[a] The number of voxels where a protein structure is placed. Both grid sizes are used for each proteins. Larger grid sizes can represent surface finer than smaller grid sizes. [b] MSROLL program in Molecular Surface Package (ver. 3.9.3) [95] is used. [c] The benchmark dataset of 2432 proteins used in the current study is searched.

3.2.8 3D-Surfer: A web application for global protein structure comparison[4]

Structure database searches using the 3DZD can be performed through the web at `http://dragon.bio.purdue.edu/3d-surfer/`. Users can search the benchmark dataset with one of the structures in the dataset (i.e., 3DZD of the protein is pre-computed) or by uploading a PDB file to the server.

3.3 Chapter summary

In this chapter, the 3D Zernike descriptor (3DZD) was introduced as a computationally efficient method for searching protein surface for function prediction. Unlike the existing methods for structure comparison and representation, the 3DZD allows

[4] 3D-Surfer is published work in [100].

rapid database searches, which opens up the possibility for real time protein tertiary structure searching on the internet. The search speed can be further increased by pre-screening proteins by their length and/or by multi-resolution search using different orders. A search against the benchmark dataset of 2432 proteins, used in this work, took only 0.46 seconds. This indicates that a search against the current entire PDB database with 45,000 proteins would take 18.5 seconds. A preeminent mathematical property of 3DZD is that it is rotation invariant. This is a significant advantage over spherical harmonics and other methods that need to pose structures on a reference frame for comparison.

Because the 3DZD concerns the protein surface shape and not the main-chain orientation, in principle, proteins found to be similar by the 3DZD do not necessarily have evolutionary relationships, as illustrated in Figure 3.4. However, our benchmark results show that in majority of the cases 3DZD retrieves protein structures of the same fold group (Table 3.2). Since proteins of the same function generally have similar structures or folds, being able to retieve proteins of similar structure, i.e., proteins of the same fold group, demonstrates the utility of the 3DZD as a protein function prediction method by mean of structure comparison. In a practical implementation of a tool for a real-time protein structure search, 3DZD could be used as a rapid primary filter, followed by an option to use a conventional structure comparison method, such as CE, to compute main-chain similarity between a query protein against the top 10 to 20 structures that are retrieved by the 3DZD. The optional uses of a detailed structure alignment method can find cases where the surface shapes are similar but the main-chain orientations are different, e.g., Figure 3.4 **C** and **D**.

4 IMPROVED REAL-TIME PROTEIN STRUCTURE SEARCH TOWARDS APPLICATION TO LOW-RESOLUTION DATA[1]

In the previous chapter, the need for high-throughput protein structure comparison method was addressed. In addition to speed improvement, there are other new challenges brought up by experimental techniques such as electron microscopy (EM), which provide low-resolution structure data. Problems include how to use a low-resolution EM density map for fitting high resolution structures [102–105], refining protein structures into low-resolution density maps [106], and predicting functions of proteins looking at their electron microscopy maps [107].

These problems again emphasize the need for a new generation of structure analysis tools that not only allow fast screening of large structure databases, but that also can handle low resolution structural data in a similar manner to conventional high resolution structures.

This chapter addresses the problem of reducing the discrepancy between the existing structural alignment by CE [29], that was observed in the previous chapter, as well as exploring the applicability of 3D Zernike descriptor (3DZD) to comparing low-resolution data where the atomic coordinates are not known. With this in mind, three new main-chain atom based surface representations are introduced, all of which are described by the 3DZD. Structure retrieval with the three representations, together with the original all-atom surface representations, are tested on two existing structure classifications: CE and the SCOP [108] database. The results show that the newly introduced backbone-atom based surface representations exhibit significantly better agreement with the existing classifications as compared with the all-atom surface representation. Moreover, close examination reveals that the retrieval agreement

[1]This chapter reuses work submitted to GIW2010 [101].

by the main-chain and the all-atom surface representations differ with the surface shapes types of the proteins, and a proper combination of representations further improves the retrieval results. Finally, we show that the proposed representation also allows fast and accurate database searches for EM density maps. This, again, is a property of the rotation and translation invariance of 3DZDs, which enables direct comparisons of EM density maps.

4.1 Materials and methods

4.1.1 Benchmark dataset

To examine structure retrieval performance of the proposed methods, a subset of proteins from the previous dataset that are labeled by both CE and SCOP classifications are used. The selected dataset is composed of 2337 structures. The dataset, in terms of SCOP classification, has 8 class groups, 149 folds groups, 187 superfamily groups, and 279 family groups. We use both CE and SCOP classifications in our study since they have complementary features: the CE classification is automatic, without human intervention and considers main-chain orientation, while SCOP is curated manually and to a certain degree, takes evolution into account.

At this juncture, it is important to note that there is no gold standard for classification of proteins. The structural similarity measured for different representations can greatly differ for distantly related proteins since each representation captures different aspects of structures [27]. As shown in the previous chapter, CE and SCOP do not fully agree, while DALI [30] and CE have poorer agreement than CE and the 3DZD. Each method has its own strength and thus an appropriate method should be selected according to the purpose at hand. For instance, we showed examples of proteins of a similar function that have similar shapes but different sequences and backbone structures. In this chapter, structure retrieval experiments are performed on the CE and the SCOP protein structure classifications, since, presumably, main-

chain structure comparison and search against SCOP are one of the well accepted protein structure analyses.

4.1.2 Computing four types of protein surfaces

For a protein structure, four different surface representations are computed: one that uses all heavy atoms (AASurf), one that uses backbone conformation with all heavy atoms in main-chain, i.e. Cα, C, N, and O atoms (CACNO), one with backbone Cα, C, and N atoms excluding the oxygen atom (CACN), and one with backbone Cα atoms only (CA). For the set of extracted atoms, the surface is generated using the MSMS program [110]. MSMS rolls a probe sphere on the protein atoms and defines the surface as the path of the center of the probe. The radius of the probe sphere is set to 1.5Å for AASurf, CACNO, and CACN. A radius of 2.0Å is used for CA to generate a smoother representation. The generated surface is then mapped on a 3D grid as described in section 3.1.2.

Figure 4.1**A** through **D** show the surface generated from the four representations. Figure 4.1**E** shows the 3DZDs of the four representations for protein 1hdm-A. It can be seen that there is little difference in the 3DZD of CACNO, CACN, and CA as compared to the 3DZD of AASurf.

4.1.3 Evaluation method

The database retrieval performance of the four surface representations is evaluated with precision-recall curves. Precision-recall curves are often confused with the receiver operator characteristic curves which are often used to evaluate the performance of binary classifiers. Although there is a relationship between the two curves, a precision-recall curve is considered being a better measure when the dataset is skewed [109]. The dataset used have groups with the number of proteins varing from 3 to 179 proteins, thus precision-recall curves are used in the retrieval evaluation. For each protein in the dataset, proteins are ranked according to the 3DZD Euclidean

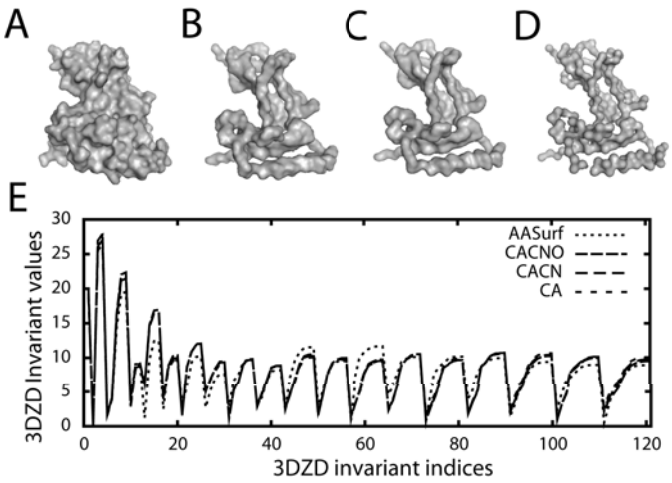

Figure 4.1. Four surface representations of protein 1hdm-A. **A** is surface representation using all-atoms (AASurf). **B** is a backbone representation using all heavy atoms in main-chain (CACNO). Next two are simplified backbone representations which are composed of the atoms Cα, C, and N (CACN) in **C** and only the Cα atoms in **D**. **E** shows the 3DZD invariant values of the four representations.

distances (L2 norms). Then, the precision and the recall values are computed at each distance threshold value. The precision is defined as the fraction of the retrieved proteins of the same group with the query among all proteins retrieved above the distance threshold. The recall is defined as the fraction of the retrieved proteins of the same group with the query among all the proteins in the same group. Finally, we calculate the average precision and recall for each distance threshold. The precision-recall curves of different representations are evaluated by the area under curve (AUC). As we employed in the previous chapter, we also apply pre-filtering of the proteins by their sequence length. For a query, a protein in the database is filtered out if it is longer than 135% or shorter than 65% to the length of the query protein.

4.1.4 Combining two descriptors: AASurf and CACNO

Among the three backbone representations, CACNO has been chosen since there are no significant differences observed in the performances among the three backbone representations (see Results) and since it is customary to choose a full heavy-atom representation of a protein as a backbone. The distances measured with the 3DZDs of AASurf and of CACNO are linearly combined as follows:

$$d_w^2(\mathbf{y},\mathbf{x}) \;=\; \frac{w_{yS}\sum_{i=1}^{m^1}(y_{Si}-x_{Si})^2 + w_{yB}\sum_{i=1}^{m^2}(y_{Bi}-x_{Bi})^2}{m^1\times w_{yS} + m^2\times w_{yB}} \tag{4.1}$$

where $\mathbf{y}$ and $\mathbf{x}$ are the two proteins compared, Si and Bi are the index of 3DZD invariants of AASurf and CACNO, w_{yS} and w_{yB} are weights for AASurf and CACNO of the query protein y, and m^1 and m^2 are the number of invariants in the 3DZD of AASurf and CACNO, respectively. In this study, the 3DZDs of AASurf and CACNO are set to same size, i.e. m^1=m^2=121. Eqn 4.1 is asymmetric since the weights, w_{yS} and w_{yB} depend on the query protein, $\mathbf{y}$.

The weights for AASurf and CACNO of a query protein are determined by two shape characteristics of the query protein: 1) the existence of a tail-like structure and 2) the sphericity. We define a tail as an elongated region in the structure that is longer than three amino acids located further than two times of the radius of gyration (RG) of the protein from the center of gravity of the protein. The radius of gyration is defined as follows:

$$RG(\mathbf{x}) \;=\; sqrt\left(\frac{1}{N}\sum_{j=1}^{N}(\mathbf{x}_j-\mathbf{cog})^2 + \frac{3}{5}R^2\right) \tag{4.2}$$

where N is the number of atoms in protein $\mathbf{x}_j$, $\mathbf{cog}$ is the center of gravity of protein $\mathbf{x}_j$, and R is radius of a pseudo atom in which 1.5Å is used [111]. The sphericity measures how well a protein structure fits to a sphere:

$$Sphericity(\mathbf{x}) \;=\; \frac{RS(\mathbf{x}) \;-\; RG(\mathbf{x})}{RG(\mathbf{x})}, \tag{4.3}$$

where $RS(\mathbf{x})$ is the radius of a sphere that has the same volume as the protein all-atom surface representation computed by the MSMS program. A larger positive value indicates that a protein is spherical.

4.2 Results

4.2.1 Retrieval performance

The database retrieval performance of the four surface representations was first tested on the CE classification. Figure 4.2**A** and 4.2**B** show results with and without the prefiltering by protein lengths. All three backbone surface representations, CACNO, CACN, and CA, show similar performances, which is significantly better than that of AASurf. Note that AASurf showed much higher performance than DALI (see in chapter 3). The length based prefiltering slightly improves the AUC by around 0.03 for all the backbone representations and by 0.06 for the AASurf representation. The prefiltering by protein lengths improves the retieval performance. On the other hand, one cac see that the improvement by the length filter is marginal, which implies that it is not common to observe two proteins of different sizes with a similar shape.

In Figure 4.2**C**, **D**, **E** and **F**, the retrieval performance based on the four hierarchical SCOP classifications, i.e. the class, the fold, the superfamily, and the family, are shown. The length-based filtering was not applied. Retrieval improvement of the backbone representations over the AASurf becomes more obvious as lower levels classification in the SCOP hierarchy is used, showing that the backbone structure similarity is more important in the family classification which is the lowest level in the SCOP hierarchy. In all the six graphs in Figure 4.2 except for Figure 4.2**C**, CACN has the best retrieval performance followed by CACNO, CA, and AASurf. Although CACN performs the best, the difference in performance by CACNO and CA is marginal: the average AUC difference is 0.003. Since the three backbone representations show almost identical performance, only the results of CACNO, along with AASurf, will be shown for further analyses.

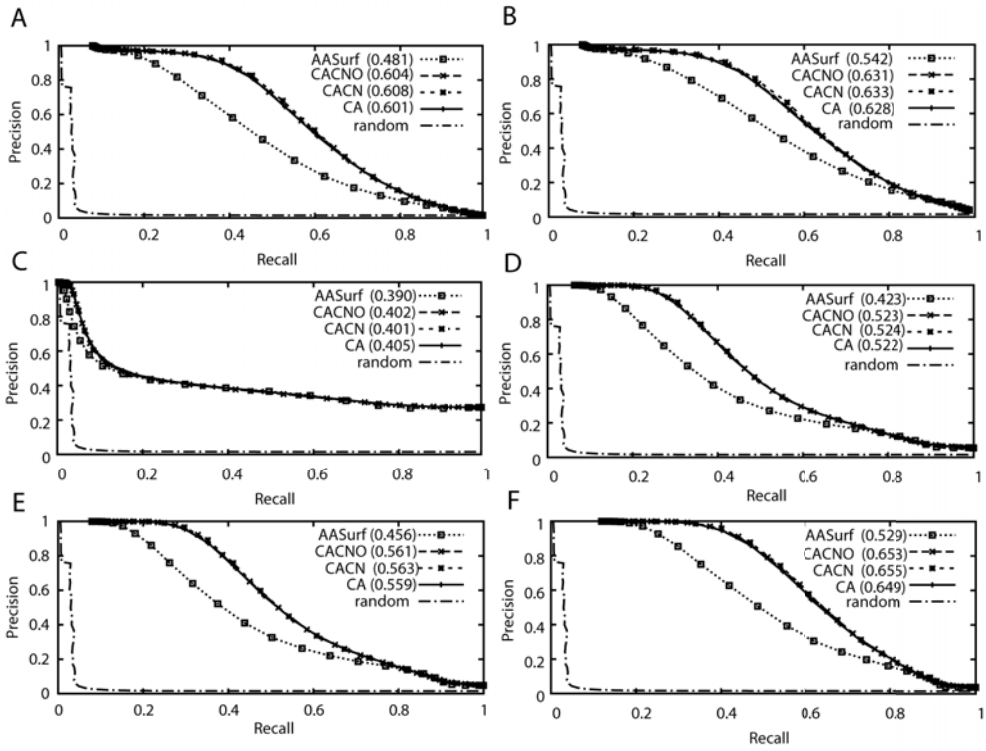

Figure 4.2. Precision-recall curves on the CE and the SCOP classification. The L^2 norm is used to compute distance between two 3DZDs. **A**, **B**, results on is the CE classification. In B, length filtering is applied. **C**, **D**, **E** and **F** are precision-recall curves on the SCOP class, fold, superfamily, and family classification, respectively, as the base truth. The length filtering is not used. The AUC values for the curves are shown inside brackets.

4.2.2 Evaluation of AASurf and CACNO

A close examination of the individual cases in the database retrieval reveals interesting trends with regard to the performance of the AASurf and CACNO representations. In Figure 4.3, the effect of the sphericity (Fig. 4.3**A**) and the tail-like structures (Fig. 4.3**B**) on the retrieval performance of individual proteins are examined. For the retrieval evaluation, the CE classification is used as the base truth. The y-axis shows the difference in the AUC of the precision-recall curves by AASurf minus that of CACNO. Thus, a positive value indicates that AASurf performs better than CACNO and a negative value indicates the opposite.

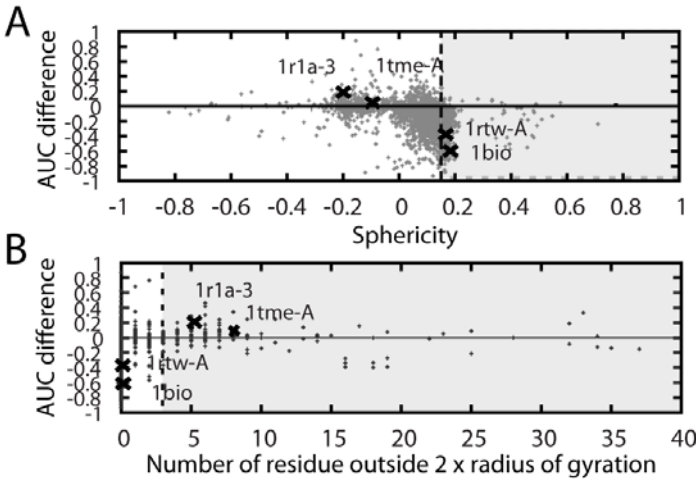

Figure 4.3. Effect of the sphericity and the tail-like structure to the precision-recall AUCs of AASurf and CACNO representations. **A**, the sphericity, **B**, the number of the residues which reside further than two times of the radius of gyration (tail-like structures).

Figure 4.3**A** indicates that CACNO tends to perform better than AASurf for spherical proteins (proteins with positive sphericity value). This means that there are groups of proteins that are similar in terms of AASurf but that are distinguished in terms of backbone similarities. For spherical proteins, AASurf shape alone does not clearly separate proteins of the same group from spherical proteins from the other groups. On the other hand, AASurf performs better for proteins with a tail-like structure which are generally the flexible regions of the proteins (Fig. 4.3**B**). The AASurf is better in tolerating small changes on the tail-like region of proteins.

Figure 4.4 shows the AUC difference between the AASurf and the CACNO representations of individual proteins. On the x-axis, proteins are ordered such that proteins of the same classification are located next to each other. For many proteins, the two representations do not make much difference (51.8% of the proteins have an AUC difference of less than 0.1), however, there are proteins for which the two representations show a large difference in the AUC values. Four examples are presented: the first two proteins, 1tme1 in Group 25 and 1r1a3 in Group 64, are cases

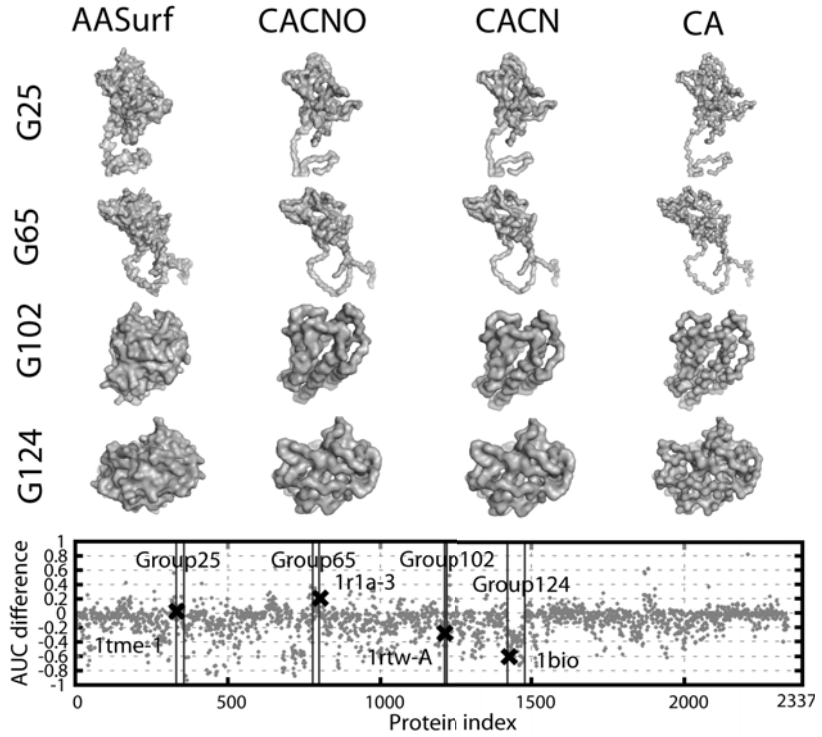

Figure 4.4. Differences in the database retrieval performance of the AASurf and the backbone representations. The graph at the bottom shows the precision-recall AUC difference of AASurf to CACNO. A positive value indicates that performance of the AASurf is better. The first two proteins, 1tme1 in Group 25 (G25) and 1r1a3 in G65, are examples where AASurf shows better performance, while the latter two, 1rtwA in G102 and 1bio in G124, are cases where CACNO performs better. 1tme1 has the AUC difference of 0.0656, the sphericity (sp) of -0.104, and has 8 residues farther than two times of the radius of gyration from the center. 1r1a3 has AUC difference: 0.146, sp: -0.199, and 5 tail-like residues. 1rtwA has AUC difference: -0.387, sp: 0.152, and no tail residues. 1bio has AUC difference: -0.597, sp: 0.165, and no tail residues.

where AASurf performs better, while CACNO performs better in the latter two cases, 1rtwA in Group 102 and 1bio in Group 124. The data points for these four proteins are specified in Figures 4.3 and 4.4. Obviously the first two examples have long tails, whereas the latter two are spherical structures with no tails.

The observed difference betweem the performance of AASurf and CACNO in the individual cases prompted us to combine the two representations so as to improve the retrieval performances. The results for combined representations are shown in the next section. Concretely, spherical proteins (those which have the sphericity value of larger than 0.15, the gray regions in Fig. 4.3**A**) and proteins with tail-like structures (those which have more than 3 residues in a tail, the gray regions in Fig. 4.3**B**) are given different weights for combining the distances of the two representations.

4.2.3 Improving retrieval performance

Next, we examine retrieval results of combining AASurf and CACNO distances using Eqn. 4.1. Query proteins with tail-like structures are given a higher AASurf weight (w_{yS}) and ones with a high sphericity are given a higher CACNO weight (w_{yB}). All others are given fixed weights for AASurf (0.4) and CACNO (0.6). With the threshold value of three residues for the tail-like structure and 0.15 for the sphericity, 383 and 150 structures fall into the category of structures with tails and spherical structures, respectively, in the dataset. Table 4.1 shows the database retrieval using the combination of AASurf and CACNO (named Surf(ace)-Back(bone) representation) with different weight values.

Among the different weight values tested, the weight combination of 0.3 and 0.7 performed the best. In this combination, the weights for AASurf (w_{yS}) and CACNO (w_{yB}) are set to 0.7 and 0.3, respectively, for query proteins with a tail-like structure. On the other hand, if a query protein has no tail-like structure and is spherical, w_{yS} is set to 0.3 and w_{yB} is set to 0.7. Otherwise w_{yS} and w_{yB} are set to 0.4 and 0.6, respectively. Overall, the 0.3/0.7 weight combination results in an AUC increase of

Table 4.1
AUC improvement using weighted distance

Representation	weight	P-R AUC	Improvement*
AASurf	-	0.481	-
CACNO	-	0.604	-
Surf-Back	0.5/0.5	0.605	0.001 (0.124)
	0.4/0.6	0.617	0.013 (0.136)
	0.3/0.7	0.619	0.015 (0.138)
	0.2/0.8	0.612	0.008 (0.131)

* Improvement precision-recall (P-R) AUC of Surf-Back : difference of Surf-Back AUC to CACNO AUC. Difference of Surf-Back AUC to AASurf AUC is shown in parentheses.

0.015 and 0.138 when compared with the retrieval results using CACNO and AASurf, respectively. The overall improvement of Surf-Back is not significant compared to the performance of CACNO. However, improvement of Surf-Back can be further highlighted when looking at the performances of individual fold groups. Out of 185 fold groups in the dataset, 116 groups show improvement by Surf-Back. There are 53 groups where CACNO shows better performances and 16 groups where AASurf performs better.

Figure 4.5 shows precision-recall curves of three fold groups. In the case of Group 48 (Fig. 4.5**A**), for which AASurf performs better than CACNO, Surf-Back improves the AUC by 0.018 compared to AASurf and 0.086 compared to CACNO. Group 74 (Fig. 4.5**B**) is an opposite example where CACNO performs better than AASurf. Surf-Back improves the AUC by 0.154 and 0.067 compared to AASurf and to CACNO, respectively. However, the linear combination of AASurf and CACNO does not improve cases where one representation performs significantly worse than the other. Fold group 124 in Figure 4.5**C** shows such an example. Surf-Back performs

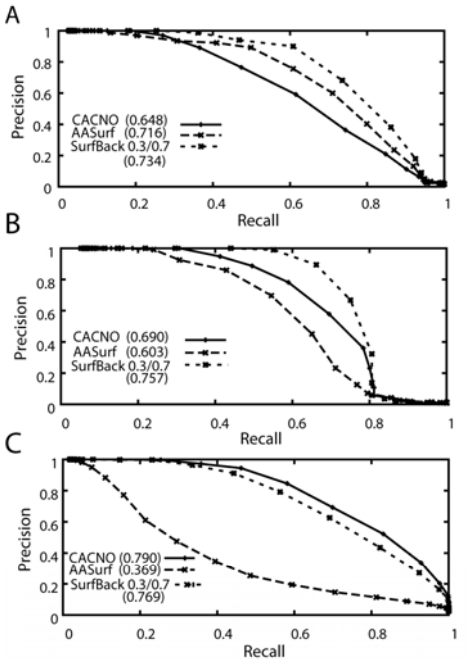

Figure 4.5. Examples of retrieval performances with the AASurf and CACNO combination. **A**, **B**, and **C** are precision-recall curves of Group 48, 74, and 124. AUC value for each precision-recall curve is shown in brackets.

significantly better than AAsurf but deteriorates the performance of CACNO by an AUC value of 0.021.

4.2.4 Towards application in comparing EM density maps

Finally, we show that 3DZDs are readily applicable to comparison of EM density maps. In Figure 4.6, low-resolution structures extracted from EM density maps and reconstructed structures of the 3DZDs are compared. The EM density maps are generated with the pdb2mrc program which simulates EM density of protein structures [112]. The grid interval size is set to 1Å and three different resolutions (10, 15, and 20Å) are used for simulating EM density maps. The derivation of the 3DZD reconstruction is described in [36]. In Fig. 4.6**A**, the original AASurf representations

of proteins are reconstructed from their 3DZDs of three different orders (20, 15, and 10). For Fig. 4.6**B**, 3DZDs of the CACNO representations are used for the reconstruction. Figure 4.6 shows that surface representation by 3DZD is similar to EM isosurfaces and thus would be suitable for describing isosurfaces of EM density maps at varying levels of resolutions.

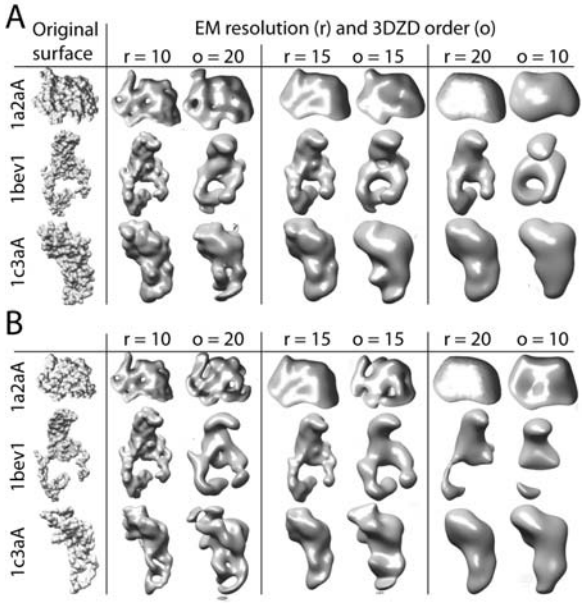

Figure 4.6. Comparisons of surfaces constructed from EM maps and reconstructed molecular surfaces from 3D Zernike moments. Three different EM resolutions (r=10, 15, and 20) and comparable surface shape reconstructed from the 3DZDs (order o = 20, 15, and 10) are shown. **A**, the 3DZDs are computed for AASurf representation of proteins, from which the surfaces are reconstructed. **B**, the CACNO representation is used to compute the 3DZDs.

Based on agreements between the EM isosurface and the structures coded by 3DZD, we investigate further as to whether 3DZDs can be employed for searching similar EM density maps of protein structures (Table 4.2). To perform this experiment, we prepared datasets of EM density maps as follows: First, simulated EM density maps of the 2337 representative protein structures are computed with

pdb2mrc for two different resolutions, 10 and 15Å. Then, for each EM density map of a protein, two 3DZDs are computed, one using voxels with a high density value range (e.g., $9\tilde{1}1$ for the resolution of 10) and another one using voxels with a low density value range (e.g., $5\tilde{8}$ for the resolution of 10; see Table 4.2). Thus, in total, four 3DZDs from the four EM isosurfaces are prepared for each protein. Voxels with the high density value in an EM density map are located at the core of the protein and thus its isosurface resembles the CACNO representation of the protein structure. On the other hand, a isosurface at the lower density value in an EM density map is more similar to an AASurf representation of a protein structure.

Table 4.2 summarizes the retrieval performance of surfaces extracted from simulated density maps described by 3DZD. Here the CE classification is used as base truth. To our surprise, the AUC values shown in Table 4.2 are as good as that of AASurf representation of protein structures shown in Table 4.1 and Fig. 4.2**A** (0.481). Among the parameters tested in Table 4.2, the 3DZDs of the order of 20 computed for EM density maps of 15Å resolution shows the highest AUC value, 0.489. Indeed, this is higher than the structural retrieval results of AASurf (Table 4.1). We believe these are very encouraging results, since it is now shown that the simulated EM density map of a relatively low resolution (15Å) can be as accurately compared as the high-resolution protein tertiary structures by using the 3DZD method. Recent EM techniques can solve protein structures at much higher resolution, such as $4\tilde{6}$Å [113]. The 3DZD should work even better for EM maps with a higher resolution.

4.3 Chapter summary

This chapter examined the applicability of the 3DZD for two important tasks in structural bioinformatics. The first task is real-time protein structure database searching. In contrast to work discussed in the previous chapter in which the 3DZD was used to represent an all-atom surface of protein structures (called in AASurf representation in this chapter), we examined backbone amino acid-based surface

Table 4.2
AUC for database retrieval of EM isosurfaces

		AUC	
EM resolution	Density range*	3DZD order 15	3DZD order 20
10	5 $\tilde{8}$	0.427	0.451
	9 $\tilde{1}$1	0.446	0.454
15	7 $\tilde{1}$1	0.466	0.489
	12 $\tilde{1}$5	0.460	0.480

* Voxels with the density value range are chosen as input of computing 3DZD.

representations (e.g., CACNO). The backbone-based representations showed higher precision-recall AUC of 0.123 (Table 4.1) than AASurf when the retrieval performance was evaluated in terms of the agreement to the CE and SCOP classifications. However,in individual cases, for flexible structures, where there exist tail-like structures, AASurf performs better. Combinations of AASurf and CACNO showed further improvement over CACNO.

The second task applied the 3DZD to the comparison of low-resolution structural data of simulated EM density maps. Visually, isosurfaces of the density maps and molecular surfaces represented by 3DZDs look very similar to each other. Indeed, the 3DZD is well suited for database retrieval based on low-resolution structures, achieving comparable accuracy to high-resolution structure database retrieval in identifying proteins of the same fold to the query protein. This is the most comprehensive study done in identifying fold class of proteins by comparing simulated EM density maps of proteins. It is noteworthy that the 3DZD can identify proteins of the same fold with simulated EM maps even at 15Å resolution. Using EM maps of a higher resolution, which have now become increasingly available, retrieval accuracy is likely to get better. In this study, we compared simulated EM maps of single proteins as the proof

of concept that 3DZD is suitable for handling EM maps. We expect this work will stimulate further investigations for applying 3DZD or similar descriptors for comparing EM maps of multiple protein complexes and other low-resolution structure data, such as electron tomography.

5 RAPID GLOBAL COMPARISON OF PROPERTIES ON PROTEIN SURFACE [1]

Previous chapters demonstrated that the 3D Zernike descriptor (3DZD) efficiently captures protein surface shape similarity and is capable of real-time protein structural searching of large datasets. This chapter proposes the use of the 3DZD method for quantitative comparison of properties defined on protein surfaces. Commonly, proteins are compared and classified based on their sequence [115] or tertiary structure [27, 29]. However, if the goal is the elucidation of protein function, a more intuitive approach would be to directly compare physicochemical properties, such as electrostatics or hydrophobicity, can be mapped on to a protein surface. This is also supported by the fact that such properties have a significant role in influencing molecular interactions [116].

Several methods have been proposed for comparison of physicochemical properties of proteins. The Carbo and Hodgkin indices [117,118], which compute inner products of electrostatic potentials of two proteins, have been used to compare electron densities, electrostatic potentials, and electrostatic fields of small molecules [117–120]. Wade et al. [121] have developed a method which compares the electrostatic potential on the protein 'skin' which is surrounding region outside certain distance of the protein. Their method was used to analyze the relationship of electrostatic potentials and biological function of several proteins and ligand molecules, including the pleckstrin homology domain family, blue copper proteins, and proteins in the ubiquitination pathway [121–124]. Pawlowski and Godzik [125, 126] introduced a method that maps protein surface properties onto a unit sphere. This method is advantageous in the sense that it is less sensitive to noise and that it can map multiple properties.

[1] This chapter is reuse of a published work in [114]

However, as is the nature of a sphere mapping, cavities on the protein surface are not properly represented. Moreover, a major drawback of these approaches is that protein structures have to be pre-aligned in order to determine corresponding regions. The pre-alignment step is often time consuming and may not yield a unique solution, especially in cases where structural similarity is low. The multi-resolution attributed contour tree method represents relative positions of electrostatic potentials on protein surfaces as a tree and does not needs any prior alignment [127]. However, its performance is offset by the high time complexity owing to the tree matching, and it has been found to be only as effective as the Carbo index.

The chapter is organized as follows. First, a short description of the data set and methods is given, followed by a simple example of unit spheres to demonstrate that different property patterns can be discriminated by the 3DZD. Then the study is extended into the biological context, showing that 3DZDs can quantitatively compare electrostatic and hydrophobic properties of evolutionarily diverse protein families.

5.1 Materials and methods

5.1.1 Surface representation

Shape, electrostatic potential, and hydrophobicity values of protein surfaces can be used to generate descriptors for selected proteins. Surface shapes are calculated using the Connolly molecular surface package (MSP) [45], and input grids are generated as described in section 3.1.2. Values of other physicochemical properties, such as the electrostatic potentials, are also assigned only to the surface voxels. The electrostatic potentials are computed by APBS [128] and hydrophobicity values are taken from the eF-site database [129]. The resulting voxels with property values are considered as a 3D function, $f(\mathbf{x})$, which is expanded into the 3DZD as described in section 2.5.2.

5.1.2 Extracting 3DZD for physicochemial properties

The surface electrostatic potentials 3DZD and surface hydrophobicity 3DZD are computed separately for the pattern of positive values and for the negative values, then later combined in the following way. First, voxels with a positive electrostatic potential value are kept but all the other voxels with a negative electrostatic potential value are reset with zero. Then 3DZD for the positive value pattern of the cubic grid is computed. Next, voxels with a negative electrostatic potential value are kept but all the other voxels are reset with zero. Then 3DZD of the negative values pattern is computed. Second, the two 3DZDs, one for voxels with positive values and another for voxels with a negative values, are combined, yielding a descriptor with $2 \times 121 = 242$ invariants. This is because the normalization process in computing 3DZD from 3D Zernike moments does not differentiate between positive and negative values but only considers the magnitude of non-zero values in the 3D space, as will be seen in Figure 5.1. Finally, we normalize the descriptor by the norm of the descriptor. This normalization is found to reduce the dependency of 3DZD on the number of voxels used to represent a protein.

5.2 Results

5.2.1 Clustering of color patterns on spheres

Previous chapters have shown that 3DZDs can be used for identifying global molecular-surface shape similarity. Here, to examine the effectiveness of 3DZD in distinguishing different surface properties, we cluster nine unit spheres with different color patterns. The surface of a unit sphere is equally partitioned into eight sections and voxels of each section are assigned a value of either 11.0 (blue) or 21.0 (red). The nine spheres, S0 to S8 are clustered according to the complete linkage clustering method using CC [Eq. 2.6] of 3DZDs [Fig. 5.1**A**]. These spheres are primarily clustered according to the number of positive voxels relative to the number of negative

voxels, that is, S0 to S8 are arranged in the decreasing order of blue sections on the spheres. However, an interesting result is observed with S3, S4, and S5. The total area of blue sections of the three spheres is the same. However, S4 is more similar to S5 in that both S4 and S5 are partitioned into two areas while S3 is partitioned into four. In contrast, when the spheres are encoded with 3DZDs that are computed without seperating the positive and negative properties, 3DZD only recognizes the contrast of patterns of the positive and the negative values but not the values themselves [Fig. 5.1**B**]. Thus, it is not able to distinguish S0 from S8 or S1 from S7. Our approach is able to obtain the clustering of the nine spheres by combining two 3DZDs separately computed for blue and red sections.

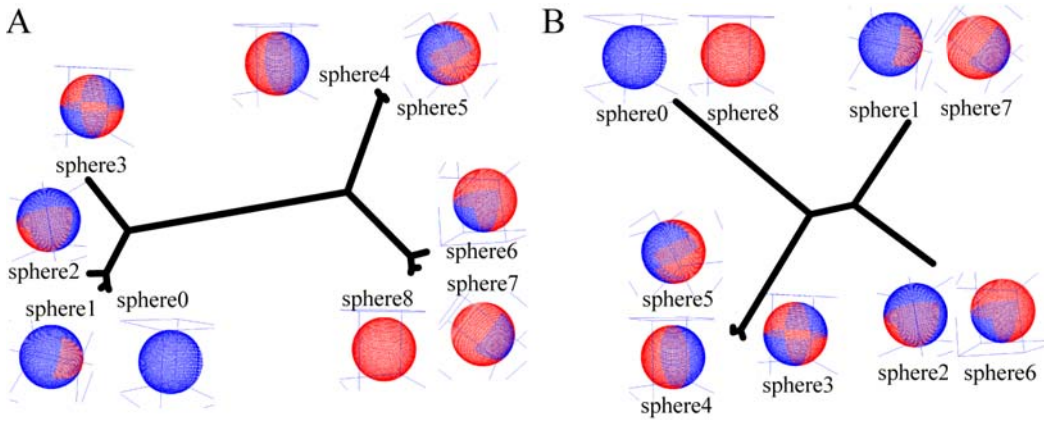

Figure 5.1. Analysis of coloring patterns on spheres using complete linkage clustering. **A** shows the clustering results when the 3DZD of the areas of positive values and the negative values are separately computed and concatenated (thus 242 invariants are used); **B** shows the results using the original 3DZD (i.e., 121 invariants are used).

5.2.2 Distance distribution of random protein pairs and the globin famliy

Next, in order to obtain an idea of the distance distribution of proteins using the 3DZD, we have performed a comparison study on 184 representative proteins and 43

globins. The representative set consists of 184 protein structures, each of which are arbitrarily selected from different CE fold groups of the dataset described in section 3.1.1. The 43 globin structures are selected from globin family structures in the SCOP database [97] that have less than 70% sequence identity with each other.

The CC [Eq. 2.6] and the EUC [Eq. 2.5] of the surface shape, the electrostatic potential, and the hydrophobicity are shown in Figure 5.2 and Table 5.1. The globin family is used because they are known to have a conserved fold pattern but have a wide variation in sequence [130] and function [131,132]; thus, structures in the globin family have high similarity in surface shapes but are expected to have more diversity in electrostatic potential and hydrophobicity than the structures.

Table 5.1
Range of 3DZD distances using representative protein set and globins.

Property	representative set(min,max)		globin set(min,max)	
	CC	EUC	CC	EUC
surface shape	(0.008, 0.723)	(0.059, 0.546)	(0.000, 0.354)	(0.001, 0.366)
hydrophobicity	(0.019, 0.993)	(0.107, 0.760)	(0.027, 0.274)	(0.099, 0.384)
electrostatic potentials	(0.011, 1.829)	(0.090, 1.241)	(0.043, 1.275)	(0.151, 0.787)

The CC of surface shape of the representative proteins ranges from 0.008 to 0.723 with the peak at value of 0.095 [Fig. 5.2**A**]. The majority of the protein pairs have a small distance because they are globular and compact. The hydrophobicity has a narrower distribution [Fig. 5.2**B**] than that of the electrostatic potential. This is partially due to the residue-based assignment of hydrophobicity values using the Kyte-Doolittle scale [133], which assigns a hydrophobicity value to each type of residue that gets mapped to multiple surface points unlike the electrostatic potential values that are assigned per surface point. The CC of the electrostatic potentials [Fig. 5.2**C**]

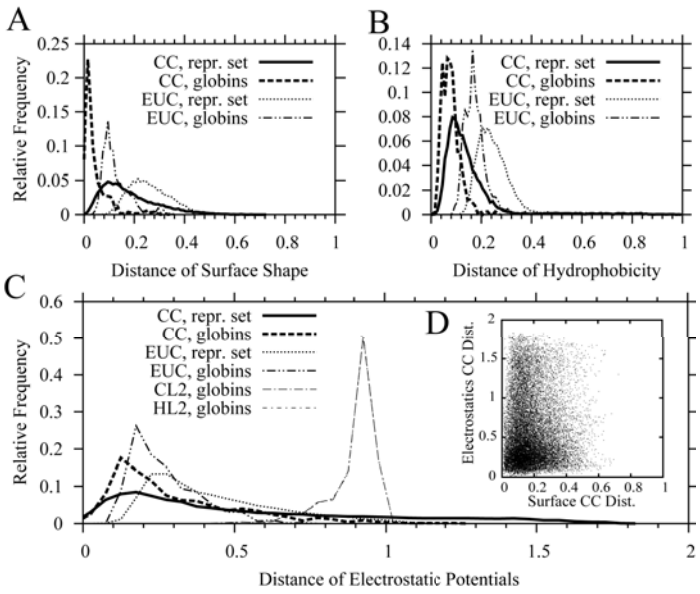

Figure 5.2. Distribution of the 3DZD distances using the representative protein set and the globin family. **A** shows distance histograms of protein surface shape. **B** shows distance histograms of surface hydrophobicity. **C** shows distance histograms of the surface electrostatic potentials. The distribution of the CL2 and HL2 are also shown. HL2 and CL2 happened to have the almost identical distribution. The (min, max) value of the CL2 (HL2) of globins is (0.473, 1.019). **D** correlation of the distance of surface shape and electrostatic potential of proteins in the representative set.

does not have a strong correlation to that of the surface shape [Fig. 5.2**D**]. This result is advantageous for comparing proteins because a combination of shape and electrostatics may therefor provide a hierarchical classification of protein surfaces.

Compared to the representative set of proteins, the surface shapes of globins are significantly more similar to each other [Fig. 5.2**A**]. This agrees with the low root mean square deviation (RMSD) between the globin pairs which range from 0.57 to 3.66 Å computed by the CE. In contrast, the electrostatic potentials of globins show higher variability, i.e, the range of electrostatic potentials distances of the globins cover 68% of the distance range covered by the representative set where as the range of shape distances of the globins cover only 26% of the distance range covered by the

representative set (coverage are computed using CC ranges). This is suggestive of the diversity of functions within the globin family [Fig. 5.2**C**], although much less diverse than that of the representative set. The distance distributions of modified Hodkins and Carbo indices, CL2 and HL2, are also shown in Figure 5.2**C**. The similarity indices by Hodkins and Carbo were calculated using the 'similar' program in the APBS package. To make the range consistent with EUC and CC, we modified the distance as follows:

$$CL2 \;=\; 1 \;-\; \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum (A_i)^2 + \sum (B_i)^2}} \tag{5.1}$$

$$HL2 \;=\; 1 \;-\; \frac{\sum_{i=1}^{N} A_i B_i}{\sum (A_i)^2 \;\sum (B_i)^2} \tag{5.2}$$

where A_i is the electrostatics potential value of position i in the electrostatic potential surface grid A and N is the grid size. HL2 and CL2 are modified L^2 inner products of voxel values which equal 0 when identical, 1 when unrelated, and 2 when opposite [127]. Note that both CL2 and HL2 rely on superimposition, which in the case of the representative set could not be performed, given the poor structural similarity. The CL2 and HL2 have a skewed distribution near 1.0, which means that there is very low similarity between the compared globin proteins, implying that these two metrics are sensitive to the difference in the electrostatic potentials of the globin proteins. CL2 and HL2 distributions for structural superimposed globins are shown in Figure 5.2**C**. Therefore, CL2 and HL2 may be suitable to compare electrostatics of very closely related proteins, but are not suitable for provide meaningful similarity metrics for more general use.

Using the representative protein set, we have examined how the order of 3DZDs affects the distances of the surface electrostatic potentials. Figure 5.3 shows histograms of CC distances between the electrostatic potentials of the representative proteins using different orders. As the order of 3DZD gets higher the histograms

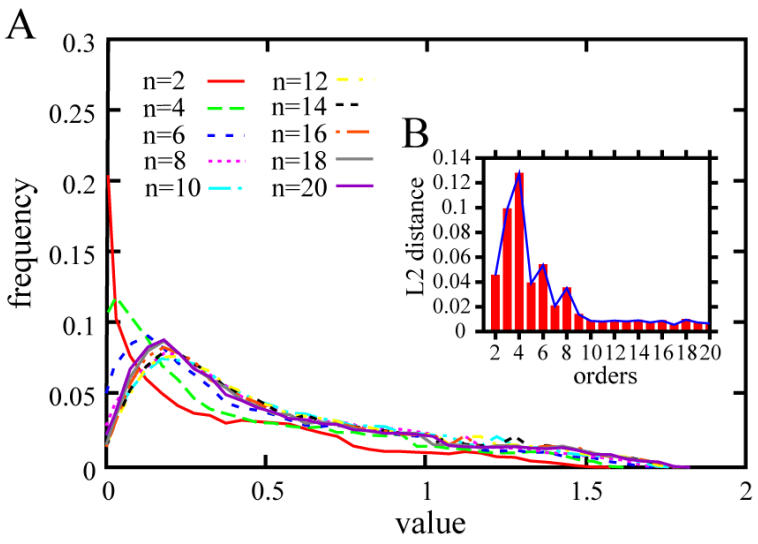

Figure 5.3. Histograms of CC distances between the electrostatic potentials of the representative proteins using different orders. **A** histograms of the CC for 10 different orders, ranging from 2 to 20, are plotted. **B** the difference of pairs of histograms of adjacent 3DZD orders, i.e., difference between the histograms of 3DZD order 1 and 3DZD order 2, that of order 2 and 3, ..., and that of order 19 and 20, are shown. The difference of two histograms is computed as average difference of frequencies at each bin.

almost converge. Therefore, it will not be necessary to use higher order than 20 in describing the surface electrostatic potentials.

Figure 5.4 shows examples of the surface electrostatics and 3DZD of several globin proteins. First, the pair of globins that exhibit the largest electrostatic potential difference according to 3DZD are shown [Fig. 5.4**A**]. 1h97A and 1hbg are monomeric hemoglobins from different organisms which are known to display extremely different oxygen affinity because of the different disposition of amino acids in their heme binding pockets [134, 135]. This apparent difference in the surface electrostatics [Fig. 5.4**A**, top panel] is captured by the 3DZD [Fig. 5.4**A**, middle panel]. The higher invariant values in the first 121 invariants of 1hbg indicates dominance of the positive electrostatic potential on the surface. Likewise, 1hbg has more hydrophobic regions

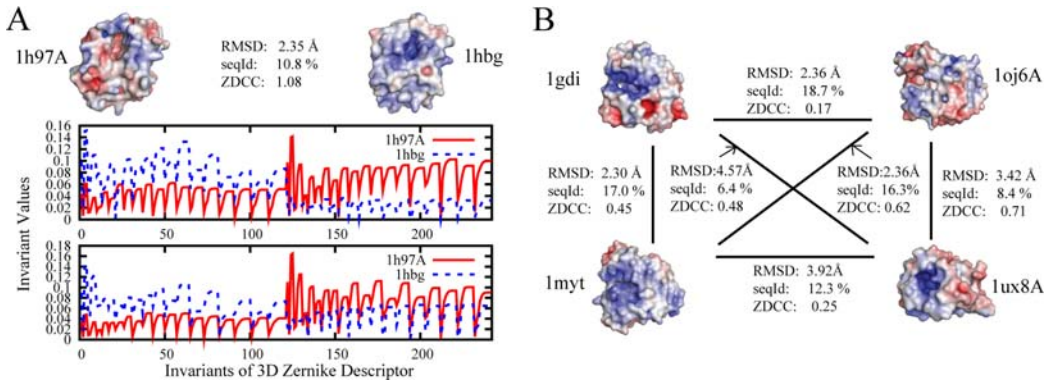

Figure 5.4. Electrostatic potential distances of globins. **A** shows the pair of globin proteins with the largest CC: Monomeric hemoglobin of Paramphistomum epiclitum (1h97A) and Glycera dibranchiata (1hbg). Positive and negative electrostatic potentials on the protein surface are represented in blue and red, respectively. The RMSD (Å) of the main-chain conformation; the sequence identity(%) (SeqId); and the CC of the surface electrostatic potentials by 3DZD of the two proteins are shown. The 3DZD of the electrostatic potentials and the hydrophobicity of the two proteins are shown in the middle and the bottom panel, respectively. **B** show four globin proteins of diverged evolutionary distance. Leghemoglobin from *Lupinus luteus* (1gdi), human brain neuroglobin (1oj6A), myoglobin from yellowfin tuna (1myt), and group II truncated hemoglobin from *Bacillus subtilis* (1ux8A) are shown with the mutual RMSD, seqId, and the 3DZD CC of the electrostatic potentials.

than 1h97A, which is conveyed by the higher values in the first 121 invariants in the third panel of Figure 5.4**A**. Figure 5.4**B** further shows examples of globin proteins of diverse functions. 1gdi is a leghemoglobin from *Lupinus luteus*, that regulates the oxygen concentration in the nitrogen-fixing aerobic bacteria [136]. 1myt is a monomeric myoglobin from yellowfin tuna, which is iron- and oxygen-binding protein found in the muscle tissue of yellowfin tuna. 1oj6A is human brain neuroglobin, whose role is to sustain oxygen supply at highly oxygen-demanding and metabolically active cells like neurons [137]. It has a more negative surface electrostatic potential, which produces a large 3DZD distance value for 1myt and 1ux8A. 1ux8A belongs to the group II truncated oxygen-avid hemoglobin from *Bacillus subtilis*. Thus, trun-

cated hemoglobin is about 20 residues shorter than full length globins [132]. The four globins shown here are all monomeric, but have a distant evolutionary relationship (i.e., the sequence identity between them is low), a varied range of affinity to oxygen and also are located in different environments. These differences coincide with the relatively large distance of surface electrostatic potentials measured by 3DZD.

5.2.3 Thermophilic and mesophilic proteins

Thermophilic proteins have substantially higher thermal stabilities as compared to their mesophilic orthologs, and the underlying principle for their higher stability has been a subject of intense discussion for the past few years [138–140]. Among the different properties examined, electrostatic contributions, especially surface electrostatics, has been identified as possibly one of the major stabilization factors [138,140]. Therefore, a method for robust and quantitative comparison of surface electrostatics may lead to a better understanding of the thermostability of proteins. Here, as a demonstration that 3DZD can cluster proteins into groups of similar physicochemical properties, surface electrostatics of a total of 14 thermophilic and mesophilic proteins from three families are compared: the dihydrofolate reductase (DIR) family, the glutamate dehydrogenase (GDH) family, and the TATA box binding protein (TBP) family.

The three DIRs [Fig. 5.5**A**] are interesting examples where similarity based on surface electrostatics is not inferred from either sequence or structure. 1dyjA has a larger region with negative electrostatic potentials (colored in red), which is represented by a 3DZD CC distance that is larger than that of 1aoeA and 1cz3A. These characteristic surface electrostatics of 1dyjA are not obvious by sequence or structure similarity, as the largest sequence identity is observed between 1dyjA and 1aoeA and the smallest RMSD is observed between 1dyjA and 1cz3A. The CL2 and HL2 again fail to provide meaningful values.

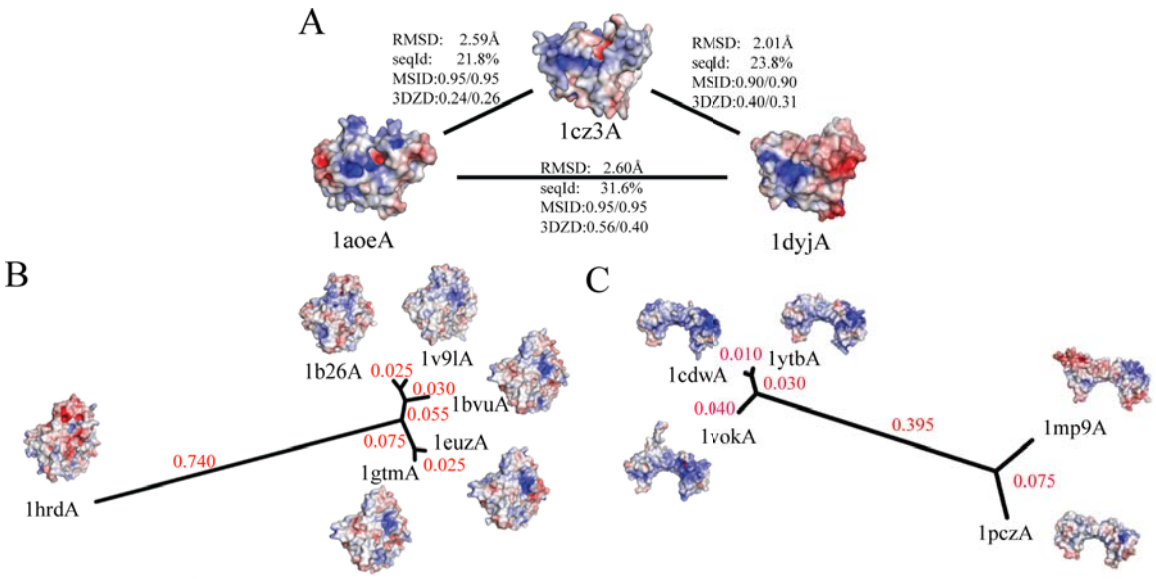

Figure 5.5. Surface electrostatic potential comparisons of the thermophilic proteins and their mesophilic homologues. **A** shows RMSD, SeqId, modified similarity index based distances (MSID: HL2/CL2), and 3DZD distances (3DZD: CC/EUC) of the surface electrostatic potentials of three DIR family proteins. 1cz3A from a thermophilic organism, *Thermotoga maritima*; 1aoeA and 1dyjA from mesophilic organisms, *Candida albicans* and *Escherichia coli*, respectively. **B** shows 3DZD CC distances of the surface electrostatic potentials of proteins in GDH family. 1hrdA is a mesophilic protein from *Clostridium symbiosum*, and the rest are thermophilic proteins: 1b26A (*Thermotoga maritima*), 1v9lA (*Pyrobaculum islandicum*), 1bvuA (*Thermococcus litoralis*), 1euzA (*Thermococcus profundus*), and 1gtmA (*Pyrococcus furiosus*). Complete linkage clustering is used. The distance is indicated in red on branches. **C** Proteins of TBP family: two thermophilic proteins, 1mp9A (*Sulfolobus acidocaldarius*) and 1pczA (*Pyrococcus woesei*) with three mesophilic proteins, 1ytbA (*Saccharomyces cerevisiae*), 1cdwA (human), and 1vokA (*Arabidopsis thaliana*).

Figure 5.5**B** shows that the 3DZD of the surface electrostatic potentials clearly discriminates a mesophilic homolog of the GDH family (1hrdA) from the other five thermophilic proteins. All five thermophilic proteins exhibit significant sequence similarity ranging from 87% to 43% and sequence similarity between the mesophilic protein and the five thermophilic proteins ranges from 33% to 36%. Although se-

quence information can distinguish the two types of proteins, surface electrostatics show more distinction between the 1hrdA with the five other thermophilic proteins with the average CC of 0.714. The thermophlic proteins are very similar in terms of the surface electrostatic potentials, having an average CC of 0.108. The classification of thermophilic and mesophilic homologs of the TBP family is shown in Figure 5.5**C**. The three mesophilic proteins, 1ytbA, 1cdwA, and 1vokA are well clustered with significantly small CC distances (less than or equal to 0.04). They also have high sequence identity range between 81% to 85%. The sequence identity for the two thermophilic proteins 1mp9A and 1pczA is 44.8% while that for 1mp9A and 1ytbA (mesophilic) is 45.0%. Sequence identity alone does not show clear separation between the thermophilic and mesophilic proteins. Similarly, clustering results using HL2 for GDH [Fig. 5.6**A**] and TBP [Fig. 5.6**B**] also do not clear separation between the the two types of proteins. Figure 5.6**A** indicates that some thermophilic protein pairs (e.g., 1euzA and 1v9lA) are as distant from each other as they are to 1hrdA. Figure 5.6**B** shows that the two thermophilic proteins, 1mp9A and 1pczA are very distinct; indeed they are more distant than 1mp9A and 1vokA. 3DZD, on the other hand, provides a clear delineation of the protein families.

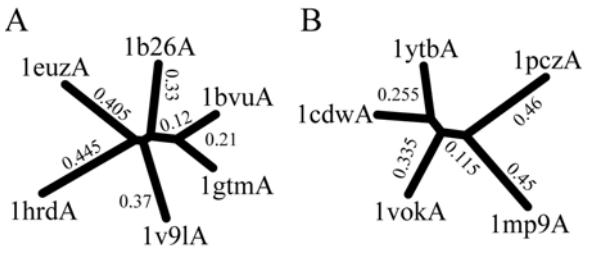

Figure 5.6. Complete linkage clustering of surface electrostatic potentials for GDHs and TBPs using HL2 measure. The HL2 distance is shown on branches. (A) GDH family. 1hrdA is a mesophilic protein and all others are thermophilic proteins. (B) TBP family. 1ytbA, 1cdwA, and 1vokA are mesophilic protein and 1pczA and 1mp9A are thermophilic proteins.

5.2.4 Computation time

Comparison of two 3DZDs is much faster than HL2 and CL2 indices, as the latter considers the whole grid, while the former only evaluates coefficients. Pairwise comparison of 3DZD takes about 0.05 sec. while HL2 or CL2 takes 50 sec on average when the grid size is set to 193. Preprocessing consists of running APBS and computing the 3DZD and can be done offline. Calculating electrostatics for a protein using APBS takes around 1.5 min (also needed in computing HL2 and CL2), and computing a single 3DZD from an APBS output file takes under a 1 min. Since the preprocessing can be done offline and once computed they do not need to be recomputed. In practice, when a new protein (one that is not in the database) is compared to proteins in a database of size N, 3DZD takes 2.5 $min.$ $+ 0.05 \times N sec.$ to find similar structures in the database while HL2 or CL2 takes 1.5 $min. + 50 \times N sec.$. 3DZD is becomes more efficient than HL2 or CL2 when comparing a protein to dataset larger than three.

5.3 Chapter summary

This chapter introduced 3DZD for fast quantitative comparison of physicochemical properties defined on protein surfaces. Using 3DZD, similarities based on properties such as the electrostatics and hydrophobicity can be quantified. The performance of 3DZD is better than those of CL2 and HL2 in since it can provide meaningful distances with less computational effort. Applications of 3DZD can be further extended for comparisons of other properties, such as residue conservation. I believe that the method introduced here opens up a way to develop new methods for quick real-time protein surface-based function assignment [142], which are as fast as those routinely used global and local sequence motif based annotation.

6 LOCAL COMPARISON OF SEGMENTED PROTEIN SURFACES [1]

In the previous chapters, representation and comparison methods for global protein surfaces using the 3D Zernike descriptor (3DZD) were examined. This chapter and the following chapter will focus on local regions of protein surfaces. In particular, this chapter introduces a local-structure based function-prediction method. A typical local-structure based function-prediction approach can be divided into two parts: 1) prediction of characteristic local sites, usually ligand binding pockets, on a given protein, and 2) comparison of local sites against a database of known functional sites to predict the function of the protein. A number of methods have been proposed that use local structure as a key feature for predicting protein function. Since a small ligand molecule usually binds to a protein at a surface pocket region, determination of pockets in the protein surface can identify active sites of enzymes in many cases [142]. Therefore, if a protein is known to bind a ligand molecule, the binding site itself can be predicted by identifying the pockets [143,144]. There are several methods available for ligand binding site predictions that use the shapes of protein structures. Some of the methods are SURFNET [145], POCKET [146], PHECOM [147], PocketPicker [148], LIGSITE [149], VisGrid [142], PocketDepth [150], and CAST [144]. Several methods consider additional information, such as sequence conservation [151–153] and energetics [154–156].

Algorithms used for comparing local sites are closely related to the type of representation. In the Catalytic Site Atlas [157], AFT [78], and SURFACE [158], where a local site is represented as a set of few residue positions, the root mean square deviation (RMSD) of equivalent amino acid residues is computed. In SiteBase [159], atoms in ligand binding sites are compared using geometric hashing. Another func-

[1] Parts of this chapter are reuses of published works in [114, 164]

tional local site database, eF-seek [31,40], represents a protein surface as a graph with nodes characterized by local geometry and electrostatic potentials, and hence uses a maximum subgraph algorithm for seeking similar sites. Thornton and her colleagues explored the use of spherical harmonics in representing and comparing protein pockets [72, 160]. They compared ligand surface shape with pocket sizes [72] and also did pocket-to-pocket comparison. A more recent method introduced by Hoffmann and colleagues [161] applies a convolution kernel method on surface atom positions and charges within 5.3Å of the binding ligand. However, pose normalization is needed prior to the matching.

This chapter focuses on the second step of local geometry-based function prediction. Two approaches of representing and comparing ligand binding pocket (LBP) properties will be introduced. After a pocket region is extracted from the protein surface, the first approach directly represents the 3D pocket with 3DZDs and compares it with 3DZDs in a precomputed dataset. The second approach further segments the LBP to smaller patches, with individual patches represented with 3DZDs. Then a match between two sets of patches, each from a LBP, is computed with bipartite matching algorithm and is evaluated to measure how similar the two LBPs are. The first approach works better when there is high conservation of overall pocket characteristics. However, when only partial region in the LBP are conserved, the second approach provides a relaxed method of pocket comparison. For simplicity, the first approach is referred to as the "global 3DZD method" and the second approach is referred to as the "local 3DZD method."

Two different datasets of LBPs are employed. The first dataset is used to test the accuracy of representing the physicochemical properties of the LBPs using the global 3DZD method. The second dataset is used to compare the performance of the proposed methods with the spherical harmonic based method [72], which is described in section 2.4.3.

6.1 Materials and methods

Both global and local 3DZD methods start by computing protein surface properties and extracting pocket regions. The global 3DZD method compares the 3DZDs of the whole pockets while local 3DZD method further segments the pockets to patches and compares the set of patches. The overview of the local 3DZD method is provided in Figure 6.1 and description of each step is provided in the following sections.

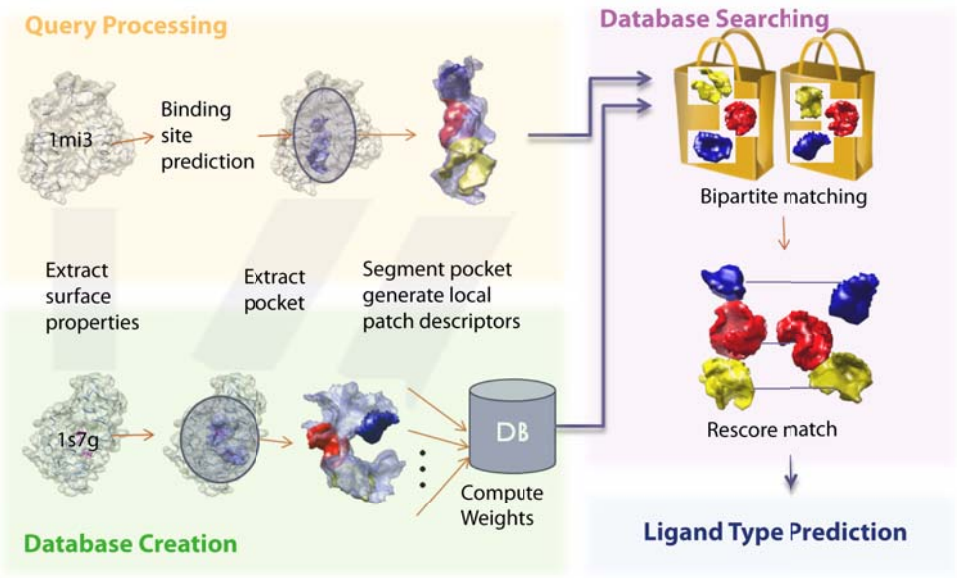

Figure 6.1. Flow chart of local 3DZD method. The local 3DZD method is composed of four parts. The query processing fist computes the protein surface properties, then, either using the ligand position or the predicted binding locations, extracts the binding pockets of the proteins, and then further segments the pocket region to local patches. The database creation process has additional step of weight training using the known binding surfaces. The database searching process consists of a bipartite matching step and rescoring step. The ligand type prediction is made using the retrieval output.

6.1.1 Datasets

There are two datasets used in this chapter. The first dataset consists of nineteen TIM barrel proteins. One TIM barrel protein is selected from each family classified in Table 2 of Nagano et al. [162]. The Nagano dataset is used to verify how the 3DZDs discriminate between the physicochemical properties of LBPs. For this simple analysis, a LBP of a TIM barrel protein is defined as the surface region that is closer than 3.5 Å to any atom of its ligand.

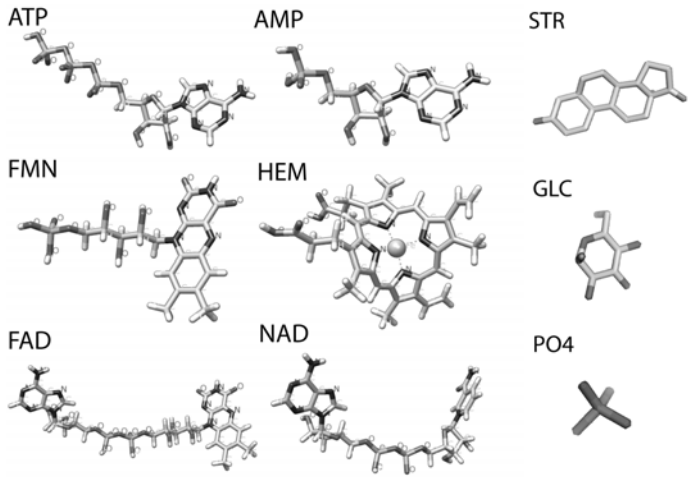

Figure 6.2. Nine ligand structures that are each bound to one of the proteins in the dataset.

The second dataset consists of 100 proteins from Kahraman et al. [72]. This dataset is used to benchmark the proposed methods with the method proposed by Kahraman et al. The Kahraman dataset consists of 100 protein structures, each of which binds to one of the following nine ligands: adenosine monophosphate (AMP), adenosine-5'-triphosphate (ATP), flavin adenine dinucleotide (FAD), flavin mononucleotide (FMN), alpha- or beta-d-glucose (GLC), heme (HEM), nicotinamide adenine dinucleotide (NAD), phosphate (PO4), or 3-beta-hydroxy-5-androsten-17-one (AND) and estradiol (EST), which are two types of steroids (STR). The structures of the nine

ligands are shown in Figure 6.2. The Kahraman dataset consists of proteins selected from different homologous families in the CATH database (i.e., H-level in CATH) so that they are not homologous and their tertiary structures have been solved by X-ray crystallography. The PDB IDs of ligand binding proteins in the dataset are listed in Table 6.1.

6.1.2 Protein surface property calculation

Four protein surface properties are used in this study: shape, hydrophobicity, electrostatic potential, and visibility. Two properties, surface shape and electrostatic potential, are computed with the Adaptive Poisson-Boltzmann Solver (APBS) program [128]. APBS evaluates the electrostatic properties of bio-molecules by solving the Poisson-Boltzmann equation. APBS first defines the solvent accessible regions of the protein and calculates the electrostatic potential of each voxel (3D grid points). The protein surface is computed as the boundaries of solvent accessible and solvent excluded regions computed by APBS and the surface electrostatic potential is the electrostatic potentials of the computed surface region. The Kyte-Doolittle hydrophobicity scale [133] is used to compute the protein surface hydrophobicity. In the Kyte-Doolittle hydrophobicity scale, residues are assigned values between -4.5 (hydrophilic) and 4.5 (hydrophobic). To obtain the surface hydrophobicity, each point on the surface is assigned the hydrophobicity value of the closest residue, then values are smoothed by averaging over hydrophobicity values assigned to its neighboring surface points. Surface visibility adapts the visibility computation proposed in VisGrid [142]. Visibility in this chapter is computed as follows. First for each surface grid point of a protein, a shell of sphere with radius equal to 80% of the specified patch radius (5Åused) is drawn. Then, the grid points that are inside the protein and outside the protein are counted. The ratio of the number of points inside the protein to the number of total points is the visibility value of the surface grid point. The range of visibility value is 0 to 1; 0 indicating an isolated point and 1 indicating a completely

Table 6.1
The ligand pocket benchmark dataset from [72]

Binding ligand	Average size[a](Å)	Seed size[b] atom	Seed size[b] interval	Number of PDB	PDB entries
AMP	8.8	23.7	24.9	9	12asA 1amuA 1c0aA 1ct9A 1jp4A 1khtB 1qb8A 1tb7B 8gpbA
ATP	9.5	29.5	29.6	14	1a0iA 1a49A 1aylA 1b8aA 1dv2A 1dy3A 1e2qA 1e8xA 1esqA 1gn8B 1kvkA 1o9tA 1rdqE 1tidA
FAD	11.0	44.1	42.9	10	1cqxA 1e8gB 1eviB 1h69A 1hskA 1jqiA 1jr8B 1k87A 1poxA 3grsA
FMN	9.7	27.7	27.7	6	1dnlA 1f4v 1ja1A 1mvlA 1p4cA 1jqiA
GLC	8.5	15.2	15.6	5	1bdgA 1cq1A 1k1wA 1nf5C 2gbpA
HEM	10.2	36.9	36.0	16	1d0cA 1d7cA 1dk0A 1eqgA 1ew0A 1gweA 1iqcA 1nazE 1np4B 1po5A 1pp9C 1qhuA 1qlaC 1qpaB 1soxA 1po5A 2cpoA
NAD	10.1	36.8	35.1	15	1ej2B 1hexA 1ib0A 1jq5A 1mewA 1mi3A 1o04A 1og3A 1qaxA 1rlzA 1s7gB 1t2dA 1toxA 2a5fB 2npxA
PO4	7.4	9.7	8.5	20	1a6q 1b8oC 1brwA 1cqjB 1d1qB 1dakA 1e9gA 1ejdC 1eucA 1ew2A 1fbtB 1gypA 1h6lA 1ho5B 1l5wA 1l7mA 1lbyA 1lyvA 1qf5A 1tcoA
STR	9.2	22.2	23.2	5	1e3rB 1fdsA 1j99A 1lhuA 1qktA

[a] The average distances from the pocket center to the pocket surface computed over the same ligand types. [b] Average number of seeds computed with the surface residue atom seed selection method (atom) and even interval seed selection method (interval)

buried point. The visibility values are computed in order to distinguish between a convex patch and a concave patch.

6.1.3 Ligand binding pocket extraction by ray casting

If the exact position of a binding ligand is known, surface region within 3.5Å of any atom in the binding ligand can be considered as the exact location of the binding site. However, if the ligand binding position is not known, i. e., for a binding ligand prediction, than taking the exact binding surface region is not possible. For the second dataset, pockets are selected such that ligand binding site prediction can be incorporated in the analysis to make the method available for a real structure base function prediction. Since ligand binding region predictions, such as LIGSITE prediction [149], identify the center locations of predicted pockets, the LBP extraction process adapts a ray casting method that casts rays from the selected centers. That is, rays are cast from the predetermined pocket centers and surface positions that are encountered first by the rays are selected as the pocket. The extraction process requires the position of the known/predicted ligand position which is used to compute the pocket centers. In this work, the binding site locations are obtained from the atomic positions of ligands that are bound to each of the ninety proteins in pdb files. The ligand atoms are segmented into three sections and the three atoms that are located ate the centers of the three sections are taken as the centers of the rays. Then spatial vectors $[x, y, z]$ that originate from the three pocket centers and terminate at the surface positions that are within 15 Åto the pocket centers are computed. The vectors are hashed in three hashes of size H^3 (H=61 used), one for each center. The hash function is defined on the discretized unit spatial vectors $[x', y', z']$ that spreads evenly over a H^3 hash as follows:

$$Hash\_index \;=\; (x' \times H + y') \times H + z'. \tag{6.1}$$

For each hash bin, a surface vector (position) of the minimum magnitude (the point that has the smallest distance from the center in the direction of the ray) is selected. Then, the selected surface points that are not near other selected points are removed and if there are any holes in the pocket the holes are filled. Extracted pockets can be directly used to generate 3DZD (order 20 is used for the whole pocket and 15 is used for local patches) for the global 3DZD method or further segmented for use in the local 3DZD method.

6.1.4 Local surface patch extraction

Given a protein surface region of interest, the next step in the local 3DZD method is local surface patch extraction. A local surface patch is a single surface region that is within a specified (5Å or 6Å is used) distance from the selected center called a "seed". The seed selection method and radius of the sphere that bounds the patches are closely related with speed and precision of the proposed local 3DZD method. Two seed selection methods were tested. The first method of seed selection selects the surface points that are closest to each surface atom but no closer than 3Å from other seeds that are already selected. Surface atoms are defined as the atoms that are within 3.5Å to the surface of the proteins. The second seed selection method selects surface points that are a specified distance away (6Å used) from each other such that seeds are evenly distributed over a protein surface. For both methods the first seed selected affects the selection of other seeds in a protein. The selection of the first seed is inconsistant from proteins to proteins which can be a cause noise. However, we try to overcome this by using large size of seeds (around 25 for a pocket).

Local surface patches are then created for each seed point by extracting continuous surface points within 5Å (or 6Å) around the seed point. Four inputs are generated for each local patch to obtain 3DZDs: $f_{shape}(x)$, $f_{hyd}(x)$, $f_{ele}(x)$, and $f_{vis}(x)$. The inputs are simply the mapping of properties, shape, hydrophobicity (hyd), electrostatic

potential (ele), and visibility (vis), onto the local surface patch. Order 15 is used to generate 3DZD for local patches. The local 3D Zernike descriptor (lzd) of the ith seed of a protein P, lzd_i^P, is composed of a seed coordinate, $s_i^P = (x_i^P, y_i^P, z_i^P)$, and four normalized 3DZDs, $zd_{i,shape}^P$, $zd_{i,hyd}^P$, $zd_{i,ele}^P$, and $zd_{i,vis}^P$. The local surface patch descriptor of a protein P, $lspd^P$, is a list of lzds for each of the seeds in the protein: $lspd^P = [lzd_0^P, lzd_1^P, \ldots, lzd_n^P]$, where n is the number of seeds in the protein P.

6.1.5 Pocket comparison and ligand binding prediction

There are three steps in the comparison process of a query protein A and a database protein B: 1) measuring the dissimilarity of local surface patch pairs, 2) matching local patches in A to B, and 3) scoring the matches.

1. Local surface patch dissimilarity of two lzds ,lzd_i^A and lzd_j^B, for ith seed in protein A and jth seed in protein B, is computed as follows:

$$ds_{lzd}(lzd_i^A, lzd_j^B) = \sum_{t \in \{shape,hyd,ele,vis\}} ws_{j,t}^B \times l2(zd_{i,t}^A, zd_{j,t}^B) \tag{6.2}$$

where $ws_{j,t}^B$ is the weight for each property $t \in \{shape, hyd, ele, vis\}$ for database protein and $l2(zd_{i,t}^A, zd_{j,t}^B)$ is the Euclidian distance between two normalized 3DZDs, $zd_{i,t}^A$ and $zd_{j,t}^B$. The weights depend on the compared protein in the database and utilize the average (avg) and standard deviation (std) of Euclidian distances ($l2$) between the 3DZDs, i.e. $zd_t(s_i^P)$s. A weight of ith seed in protein P is defined as follows.

$$ws_{i,t}^P = \frac{mask_t \times \frac{1}{avg_t + 2std_t}}{\sum_{a \in \{shape,hyd,ele,vis\}} mask_a \times \frac{1}{avg_a + 2std_a}}. \tag{6.3}$$

The $mask_t$ is either 1 if property t is used or 0 if not used. There are two approaches used for computing the distance statistics. The first approach simply computes the Euclidean distance of 3DZD for every pair in the dataset, for each property type t, and computes average and standard deviation of the Euclidean distances, avg_t and

std_t. The second approach groups the local surface patches by the ligand atom that is closest to each of the local patch. Then the Euclidean distance statistics for each property type t are computed for each group of patches. That is, avg_t and std_t in equation 6.2 are replaced with $avg_t^{l,a}$ and $std_t^{l,a}$, where l is the ligand type and a is the atom id (as defined in the PDB entry). The first approach used when the number of proteins is too small to justify the uses of the statistics when patches are grouped as the second approach. That is, the first approach is used to compute the weights for STR binding proteins, which are composed of three EST binding proteins and two AND binding proteins. The second approach is used to compute weights for all other ligand binding proteins. The distance statistics computed for the Kahraman dataset are presented in the Results section.

2. The second step in the comparison process is matching local patches of protein A and of protein B. That is, the local surface patches in protein A are matched to the local surface patches in protein B so that the resulting match has the minimum possible dissimilarity measure. This problem is the minimization version of the weighted bipartite matching problem which can be approximately solved by the auction algorithm [163]. Since the original auction algorithm is used for the maximum weight complete bipartite matching problem, a modification has been made to two parts to solve minimum distance selective bipartite matching. Complete matching refers to a match that pairs all the patches in a protein to another protein without overlap. The selective matching refers to a match that may not pair all the patches in a protein to another protein. In the modified auction algorithm presented, minimization by distance is done by computing a "weight" as the negative distance to some large value, while selectivity is introduced by ignoring matches if the distance of two seeds is larger than a given threshold distance value. Pseudocode of the modified bipartite matching is provided below.

Modification was done to the auction algorithm for bipartite matching that was developed by Demange et al. [163] as follows:

input lspdA and lspdB {local surface patches of query protein A and datbase protein B}
Set $\delta \leftarrow 1/(n_A + 1)$ {n_A is the size of lspdA }
Set $td \leftarrow$ threshold distance value
Store all lspdB i to queue $Q_i \leftarrow i$
for $j = 1$ to n_A **do**
  Set $p_j \leftarrow 0$ and $pair_j \leftarrow -l$
end for
while Q is not empty **and** number of iteration is less than $10 \times n_A$ **do**
  Set $i \leftarrow Q.front()$ and Dequeue Q
  Find j that maximizes $w_{ij} - p_j$ where w_{ij} is LARGEVALUE - d_{ij}
  {d_{ij} is Euclidean distance of 3DZD}
  if $w_{ij} - p_j \geq 0$ **and** $d_{ij} < td$ **then**
    Enqueue current pair of j $pair_j$ into Q
    Set $pair_j \leftarrow i$ and Update $p_j \leftarrow p_j + \delta$
  end if
end while
Output pairs of ($pair_j$, j) for all $pair_j$ not equal to $-l$

3. After a local surface patch matching is obtained, the match is again rescored using two scoring terms: the weighted average dissimilarity of lzds of the match, and the weighted relative position difference of the match. Pocket size information that has been useful in many of the previous studies [72,161,164] can also be incorporated.

The first scoring term is the weighted average distance of 3DZD values ($avgZd$) for evaluated properties. Computed on a query protein A and a protein in database B, $avgZd$ is defined as follows:

$$avgZd(lspd^A, lspd^B, m^{A,B}) \;=\; \left(\frac{n_A}{N}\right)\left(\frac{1}{N}\sum_{(i,j)\in m^{A,B}} ds_{lzd}\left(lzd_i^A, lzd_j^B\right)\right), \qquad (6.4)$$

where n_A is the number of patches in a protein A, ds_{lzd} is distance of descriptors as defined in Eq. 6.2, and $\frac{n_A}{N}$ is an additional weighting factor that penalizes the match when the number of matched pairs is smaller than the queried patch number.

Relative position difference is a simple measure of the difference between the relative distribution of local surface patches in protein A to local surface patch in protein B. Relative position difference (rpd) of a match, $rpd(lspd_A, lspd_B, m^{A,B})$ is defined as follows:

$$\begin{aligned} & rpd(lspd^A, lspd^B, m^{A,B}) \\ & \quad = \left(\frac{n_A}{N}\right)\left(\frac{2}{N/(N-1)}\sum_{i=0}^{N-1}\sum_{j=i+1}^{N}|l2(s^A_{m^A_i}, s^A_{m^A_j}) - l2(s^B_{m^B_i}, s^B_{m^B_j})|\right), \end{aligned} \tag{6.5}$$

where $m^{A,B}$ contains N pairs of local surface patch indices (m^A_i, m^B_i) and s^A and s^B are the seed coordinates for proteins A and B. The pocket match distance ($pocketMd$) is then computed as linear combination of the two terms with weights w_1 and $1 - w_1$, where $0 \leq w_1 \leq 1$, as follows:

$$\begin{aligned} pocketMd(m^{A,B}) &= w_1 \times avgZd(lzd^A_i, lzd^B_j, m^{A,B}) \\ &+ (1 - w_1) \times rpd(lspd^A, lspd^B, m^{A,B}). \end{aligned} \tag{6.6}$$

The weight values of $w_1 = 0.3$ and $1 - w_1 = 0.7$ is used in the study. The pocket size difference is an optional term, that have been found to increase comparison performances in [72, 164] can be incorporated in the final scoring function that is defined as follows:

$$pocketSd(A, B) = \left|\frac{n_A - n_B}{n_B}\right|, \tag{6.7}$$

Pocket size difference can be linearly combined with match score, $pocketMd$, with weights w_2 and $1 - w_2$, where $0 \leq w_2 \leq 1$, as follows:

$$w_2 \times pocketMd(m^{A,B}) + (1 - w_2) \times pocketSd(A, B). \tag{6.8}$$

The weight values of $w_2 = 0.1$ and $1 - w_2 = 0.9$ are used in the study.

6.1.6 Pocket retrieval performance evaluation and ligand type prediction score

The retrieval performances of pocket comparison methods, using leave-one-out cross validation, are evaluated by the receiver operating characteristic (ROC) curve. Concretely, given a query pocket, it is compared with all other pockets in the dataset and the top k pockets in the database are retrieved. Then they are evaluated by computing false positive (x-axis) and true positive (y-axis) rate. The value of k is varied from 1 to $N-1$ where N is the number of proteins in the dataset. The false positive rate of a set of retrieved pockets for a query is defined as the ratio of the number of retrieved pockets of a different ligand (i.e., false positives) relative to the total number of pockets of a different ligand (i.e., false positives and true negatives) in the dataset. The true positive rate is the ratio of the number of correctly retrieved pockets (i.e., true positives) relative to the total number of pockets of the same type in the dataset. The false positive rate equals true positive rate, on average, in the random cases. Individual ROC curves are compared by the area under curve (AUC) value of the curves.

Given the rank of a protein in the matches for a query protein, the ligand type can be predicted using the *Pocket_score* defined in our previous work [164]. Using the k (k=10 is used) closest pockets to a query, the scoring function of a ligand type F for a query protein P is defined as follows:

$$Pocket\_score(P,F) \;=\; \sum_{i=1}^{k}\left(\delta_{l(i),F}\log\frac{n}{i}\right)\frac{\sum_{i=1}^{k}\delta_{l(i),F}}{\sum_{i=1}^{n}\delta_{l(i),F}}, \tag{6.9}$$

where $l(i)$ denotes the ligand type (AMP, FAD, etc.) of the ith closest pocket to the query, n is the number of pockets of the type F in the database, and the function $\delta_{l(i),F}$ is equal to 1 if the ith protein is of type F, and is 0 otherwise. The first term considers top k closest pockets to the query, with a higher score assigned to a pocket with a higher rank. The second term normalizes the score by the number of pockets of

the same type F included in the database. The ligand with the highest $Pocket\_score$ is predicted to bind to the query pocket.

6.2 Results

6.2.1 Binding sites of TIM barrel proteins

The TIM β/α barrel are one of the most prevalent folds adopted by a variety of enzymes [162]. The active sites of TIM barrel enzymes, which are usually located at the C-terminal region with a cluster of loops of the barrel, show wide ranging pattern in terms of electrostatics. This is also reflected in the nature of the ligands that are bound [165–167]. As a demonstration that 3DZD can effectively compare local surface electrostatics, the ligand binding sites of 19 TIM barrel fold enzymes of different families, that bind to different ligands, have been classified. Figure 6.3 shows the 19 LBPs clustered into three groups, two of which have negative electrostatic potentials while the other has a dominant positive potential.

All of the bound ligands in group 3 have one or more phosphate groups, which complements binding sites with positive electrostatic potentials. In contrast, enzymes in groups 1 and 2 bind ligands with positively charged groups (e.g., aminopteridine and purine) or metal ions (e.g., Mg^{2+}, Mn^{2+}, and Zn^{2+}) in the binding pockets. Despite groups 1 and 2 both having negative potentials, strong peaks in the first few invariants of group 1 differentiate it from group 2. The pattern of the invariant (one with the peak) are common among sphere-like binding sites in group 1.

6.2.2 Reconstruction of local surface shape from the 3DZD

How well 3DZDs capture shape of local surface patches are examined by reconstructing the original shapes from the 3D Zernike moments. In Figure 6.4, three local patches of protein 1kvkA (PDB ID) are reconstructed using four different resolution levels: orders 5, 10, 15, and 20. As seen in the first row of Figure 6.4, the three local

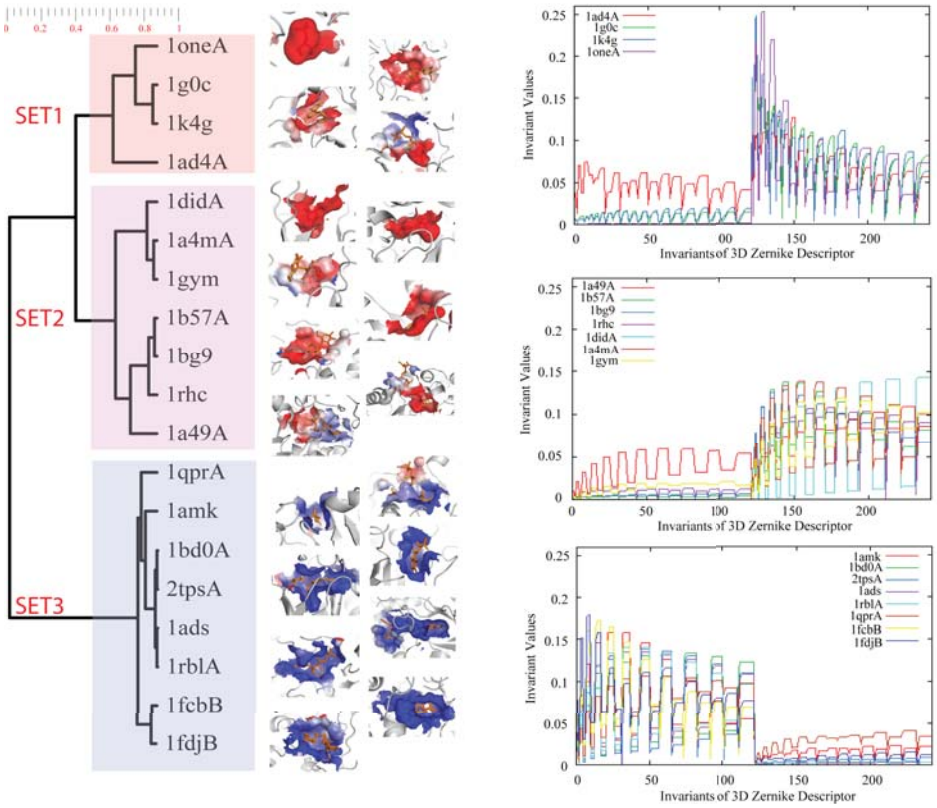

Figure 6.3. Binding site electrostatic potential of 19 TIM barrel structures: The tree structure shows complete linkage clustering using 3DZD CC of electrostatic potentials on the ligand binding interface. Ligand binding sites are computed as surface regions within 3.5 Å of the bound ligand. Two clusters are formed using a CC threshold of 1.22: Set1 + Set2 and Set3. Three clusters are formed using a CC threshold of 0.61: Set1, Set2, and Set3. The three plots on the right show 3DZD invariants for each of the sets.

patches have different shape: local patch 1 has a wrap-like shape; local patch 2 has a pocket shape; and local patch 3 is saddle-like. 3DZD captures these differences well as shown in clear distinction of the descriptors in Figure 6.4**B**. Figure 6.4**B** shows 3DZDs of the order $n = 20$ (121 invariants). An 3DZD of a smaller order is a subset of the invariants in the 3DZD of higher order. For example, the 3DZD that corresponds to the order $n = 15$ is the first 72 invariants out of 121 invariants that corresponds to 3DZD of order $n = 20$. Figure 6.4 shows that reconstruction results of 3DZD with order $n = 15$ is as good as those of order $n = 20$. In previous chapters, we used the order of 20 for describing global protein surface shape. On the other hand, order $n = 15$ is sufficient for local surface patches because they are less complex than that of global surface. Thus, the order $n = 15$ is used to describe local patches.

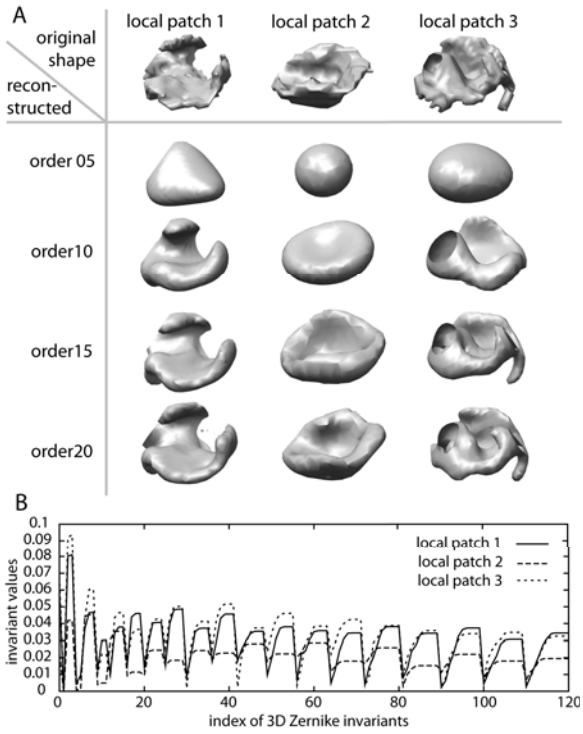

Figure 6.4. Local surface shape reconstruction from the 3DZDs. Three local patches from protein 1kvkA are reconstructed using four different resolutions, the order of 5, 10, 15, and 20.

6.2.3 Ligand local property variance

The Kahraman dataset consists of proteins that are non-homologous in terms of both sequence and structure within the same ligand binding group [72,168].Figure 6.5 shows the mapping of average 3DZD Euclidian distances of protein surface patches that are grouped by the closest ligand atom to the ligand atoms that represent each group. The maps are computed for each of the four properties: shape, hydrophobicity, electrostatic potential, and visibility. The figure shows that even for surfaces that are bound to a same ligand type, some regions are more conserved than other. Also, the conversed regions can be different for each of the four properties. This support the rationale for segmenting pockets to local patches that can accommodate the difference in the conservative of the surface properties for sub-regions within the pocket. More detailed discussion of the local 3DZD capturing the local similarities is given in the following sections.

6.2.4 Retrieval performance difference of global 3DZD and local 3DZD methods

The assumption of the local 3DZD method was that some ligands are flexible such that, although global pocket surface properties are not conserved, there are local regions in the pocket that may be conserved. Thus, segmenting a global pocket to set of local patches will help in identifying pocket similarities. The leave-one-out retrieval performance using the Kahraman dataset was performed and ROC curves plotted as shown in Figure 6.6. For this analysis, shape information with and without pocket size information is used for both global and local 3DZD method. Figure 6.6 clearly shows that segmenting the pockets improves the retrieval result. However, there is little difference (less than AUC 0.05) between local pockets with/without size information and the global pocket with pocket size information. The average size information for each ligand type is provided in Table 6.1. Additionally, the reported value of ROC AUC based on the spherical harmonic method [72] without using size is 0.64 and with the use of size information is 0.77. This is lower than the AUC values

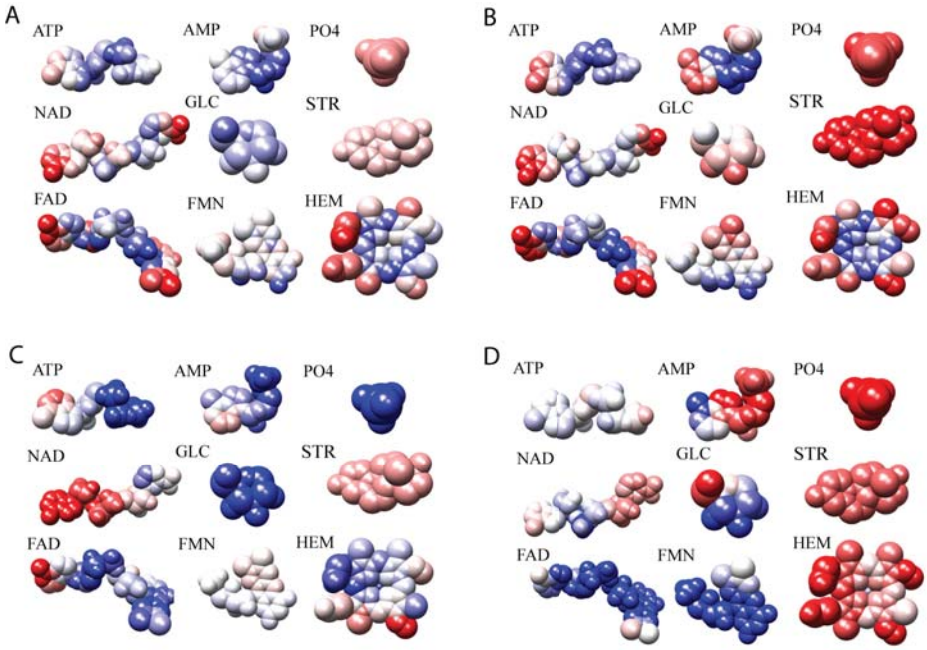

Figure 6.5. Average 3DZD distance values of four protein properties map to closest ligand atoms. **A** shows average 3DZD distances of local patch shape color coded blue (0.2), white (0.3), and red (0.4). **B** shows average distance of visibility color coded blue (0.05), white 0.07, and red (0.09). **C** and **D** shows average distance of hydrophobicity and electrostatic potential color coded blue (0.55), white (0.65), and red (0.75).

for both global and local 3DZD methods when considering the uses of size information seperatly. The retrieval performance for individual ligand types are discussed in the latter sections.

Five parameters were tested in the local 3DZD method. The first parameter considers the method of seed selection. There are two methods, atom position seed selection and even interval seed selection, as described in the Methods section. The average number of seeds for each ligand type is shown in the fourth and fifth column of Table 6.1. The second parameter is the patch radius specifing the maximum distance from the center to the border of a patch. Radii of 5Å and 6Å were tested. The third parameter is the threshold distance used in the bipartite matching process described in section 6.1.5. Results for using $avgZd$ threshold of 0.3 was compared with

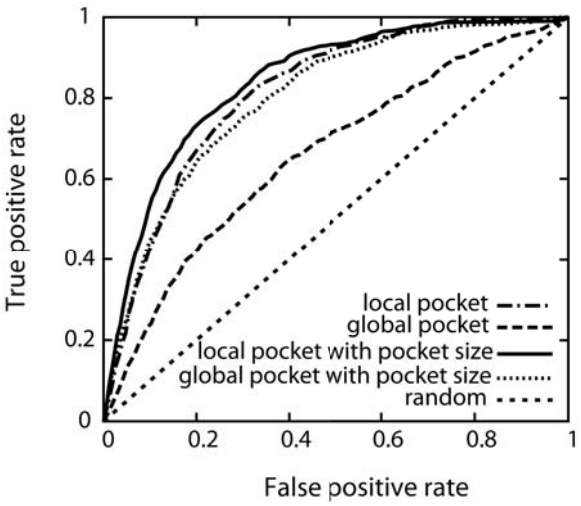

Figure 6.6. ROC curves of pocket shape retrieval using global 3DZD and local 3DZD methods. AUC values for local 3DZD method is 0.81, for global 3DZD method is 0.66, for local 3DZD method with pocket size is 0.84, and global 3DZD method with pocket size is 0.80. The parameters used to generate the local 3DZD method are: atom position seed selection, patch radius 5Å, and no threshold distance used for bipartite matching. ROC computations for Global 3DZD method are drawn from data in [164]

another that does not use any threshold distances. The $avgZd$ threshold value 0.3 is approximately equal to the average plus the standard deviation of the $avgZd$ values resulting from matches not using a threshold value. The fourth parameter considers the use of different combinations of protein surface properties: shape, hydrophobicity, electrostatic potential, and visibility. Combination of surface properties are incorporated by masking values $mask_t$ in ds_{lzd}s (Eq. 6.2 and Eq. 6.3). Fifth parameter is the use of pocket size information (Eq. 6.7 and Eq. 6.8). The ROC AUC values are summarized in Table 6.2.

Table 6.2 shows that there is little performance difference between the combinations of parameters. Selecting seeds according to the surface atom positions has better ROC AUC value than selecting seeds evenly across the protein surface. Changing patch radii from 5Å to 6Å made little difference in AUC values. Using a threshold cut-off distance also did not improve the performance of the matches. However, we

Table 6.2
ROC AUCs using different parameters

Score	Property masking[a]	Atom position seed				Even interval seed			
		Radius 5Å		Radius[b] 6Å		Radius 5Å		Radius 6Å	
		t[c]	non[d]	t	non	t	non	t	non
PocketMd[e]	s- - -	0.78	0.81	0.77	0.81	0.78	0.80	0.77	0.80
	s- -v	0.78	0.81	0.78	0.81	0.79	0.80	0.76	0.79
	s-ev	0.73	0.80	0.74	0.81	0.75	0.79	0.73	0.80
	sh-v	0.78	0.81	0.77	0.81	0.78	0.79	0.76	0.79
	shev	0.74	0.81	0.73	0.81	0.74	0.78	0.73	0.79
PocketMd	s- - -	0.83	0.84	0.83	0.84	0.81	0.82	0.81	0.82
with	s- -v	0.83	0.84	0.83	0.84	0.81	0.82	0.81	0.82
pocket	s-ev	0.81	0.84	0.81	0.84	0.79	0.82	0.78	0.82
size	sh-v	0.83	0.84	0.83	0.84	0.81	0.82	0.81	0.82
	shev	0.81	0.84	0.80	0.84	0.79	0.81	0.78	0.82

[a] Order of the masking is shape(s), hydrophobicity(h), electrostatic potential(e), and visibility(v). s,h,e,or v if the property in the position is used and - if not used to describe the protein patches. [b] Radius of sphere used to extract local surface patches. [c] Threshold distance value 0.3 is used in the bipartite matching. [d] No threshold distance value is used in the bipartite matching. [e] Pocket match distance as described in equation 6.8.

believe that matching with threshold distance will be helpful when the method is applied to comparison such as in comparison of whole protein surface. The threshold cut-off distance is expected to direct the matching algorithm to avoid matching vary different pairs of patches, which is likely to happen when a whole surface is used. Also there is only slight difference in the performance when using different property combinations. Pocket size information improves the AUC values in all cases of parameter combination with average AUC value of 0.051. Overall, ROC AUCs of combinations of parameters ranging between 0.73 and 0.84. With performance being comparable for the combinations of parameters, for further analysis, parameter combination of seed selection by atom position, patch size 5Å, no threshold distance in bipartite matching, and property mask *shev* (use all four surface properties) will be used.

6.2.5 Ligand type retrieval and prediction accuracy of individual pocket types

In this section, retrieval performance and ligand type prediction accuracy of individual pocket types are examined. Table 6.3 summarizes the retrieval accuracy of both the global and local 3DZD methods. First, we compare retrieval performance of global and local 3DZD methods for surface shapes. Looking at the performance of each ligand type, the local 3DZD method improved over global 3DZD method except for the case of FMN. The decrease of AUC value for FMN, however, is only by 0.012. The performance of GLC improved by 0.47 which is a significant improvement. HEM, AMP, PO4, ATP, STR and FAD also have good improvements of 0.22, 0.20, 0.19, 0.15, 0.15, and 0.12 in the order listed. Also there is minor improvement on NAD of 0.02. The improvement in GLC and PO4 contribute mostly from the relative position distance value (*rpd*). The average *rpd* value of PO4 binding protein pairs is 2.41 and is 2.93 for GLC binding protein pairs, which are lower than the average *rpd* value 3.94 for matches computed among the same ligand types. It was expected that segmenting the pockets will contribute to the local 3DZD result being tolerant to flexible ligands where local similarities are preserved. However, improvements of

FAD and NAD, which are the two most flexible proteins, were two of the groups that had the least improvement. Examples of local 3DZD on flexible ligands are given in section 6.2.6. Incorporating hydrophobicity, electrostatic potential, and/or visibility improves the retrieval performance for some of the ligand types but the trend was not consistent. That is, performance difference was not high and all AUC range from 0.80 to 0.81 when size information is not used and all AUC equaled to 0.84 when size information is used. This result is consistent with our previous work [164] where electrostatic potential was incorporated; but there was no improvement.

Table 6.3
Variance of ROC AUCs in ligand type using different combination of protein properties

Ligand Id	global 3DZD	local 3DZD[a]				
	shape only	s- - -	s- -v	sh-v	s-ev	shev
AMP	0.56	0.76	0.75	0.74	0.75	0.77
ATP	0.60	0.75	0.73	0.74	0.73	0.74
FAD	0.71	0.83	0.84	0.83	0.84	0.83
FMN	0.61	0.59	0.58	0.60	0.57	0.58
GLC	0.39	0.86	0.83	0.81	0.81	0.82
HEM	0.73	0.94	0.95	0.95	0.95	0.96
NAD	0.69	0.70	0.70	0.70	0.72	0.71
PO4	0.71	0.89	0.90	0.89	0.89	0.89
STR	0.73	0.89	0.78	0.89	0.77	0.83
ALL	0.66	0.81	0.81	0.81	0.80	0.81

[a] Order of the masking is shape, hydrophobicity, electrostatic potential, and visibility. 1 if the property in the position is used and 0 if not used to describe the protein patches.

Binding ligand were predicted for the query proteins using the *Pocket_score* (Eq. 6.9). The *Pocket_score* is computed for each of the nine ligand types and one with the highest score is predicted to be the binding ligand. Table 6.4 provides success rates of the ligand predictions using the top 10 proteins reretrieved (k=10 in Eq. 6.9) for each query. Overall, both global and local 3DZD method perform far better than the random retrieval. The local 3DZD method using shape information with pocket size shows the best average success rate in the Top3 prediction (84.5%). The success rate differs from ligand to ligand: the prediction for HEM binding pockets show the highest success rates in all cases. HEM binding pockets are characteristically abundant in a local patch with similar properties contributing from symmetrical shape and properties which make it easier for them to be identified. The prediction for PO4 binding pockets also show high success rates, which may be due to their smaller pocket sizes. The prediction for FAD is also among the better performing ligand type. The Local 3DZD method using all four surface properties works the best for FAD. On the other hand, FMN are difficult to identify. For FMN, the prediction failed for all of the pockets in the Top1 and was successful in only one third of them (two out of six FMN binding pockets) in the Top3 ligand prediction results when the size information was incorporated. Individual examination of six FMN binding proteins shows that they have considerably different shapes from each other. That is, some are completely buried (1ja1) whereas others are more exposed. Additionally local similarities show moderate differences, which means there is no local region that is highly conserved in terms of surface properties.

6.2.6 Performance of local 3DZD method on flexible ligands

One of the assumptions for segmenting the ligand pockets was that, for the LBPs where global properties of the pockets vary, local similarities are preserved which can be captured by segmenting the LBPs to smaller patches and comparing the patches instead of the global pocket.

Table 6.4
Summary of the binding ligand prediction on Kahraman dataset

Descriptor	Rank	AMP	ATP	FAD	FMN	GLC	HEM	NAD	PO4	STR	Average
local 3DZD s- - -[a]	Top1	33.3(44.4)	28.6(35.7)	50.0(70.0)	0.0(0.0)	0.0(40.0)	93.8(92.8)	0.0(6.7)	90.0(100)	20.0(40.0)	35.1(47.7)
	Top3	66.7(77.8)	71.4(92.8)	100(90.0)	16.7(60.0)	60.0(100)	100(100)	80.0(60.0)	100(100)	80.0(80.0)	75.0(84.5)
local 3DZD shev[b]	Top1	33.3(66.7)	28.6(42.9)	70.0(80.0)	0.0(0.0)	0.0(60.0)	100(93.8)	6.7(6.7)	80.0(100)	20.0(20.0)	37.6(52.2)
	Top3	77.8(88.9)	64.3(85.7)	100(80.0)	0.0(40.0)	80.0(100)	100(100)	73.3(73.3)	100(100)	60.0(80.0)	72.8(83.1)
global 3DZD[c]	Top1	0.0(0.0)	7.1(21.4)	20.0(50.0)	0.0(0.0)	0.0(0.0)	87.5(87.5)	60.0(60.0)	60.0(100)	0.0(0.0)	26.1(36.1)
	Top3	22.2(77.8)	57.1(100)	70.0(90.0)	40.0(16.7)	0.0(80.0)	100(100)	86.7(100)	95.0(100)	80.0(80.0)	61.2(82.7)
Pocket size[d]	Top1	22.2	7.1	50.0	0.0	0.0	0.0	26.7	100.0	0.0	22.8
	Top3	55.6	78.6	80.0	0.0	0.0	81.2	60.0	100.0	0.0	50.6
Random[e]	Top 1	9.8	13.2	9.7	6.3	5.4	15.4	14.4	19.4	6.2	11.0
	Top 3	28.0	39.7	30.7	20.8	17.0	45.0	42.1	54.5	19.0	33.0

The numbers in parenthesis show AUC values when size information is used. [a] The local 3DZD method using the shape information. [b] The local 3DZD method using shape, hydrophobicity, electrostatic potential, and visibility information. [c] The global 3DZD method using shape information. [d] The pocket size (Table 6.1) is used to retrieve pockets in the dataset. [e] The average of 500 random trials is shown.

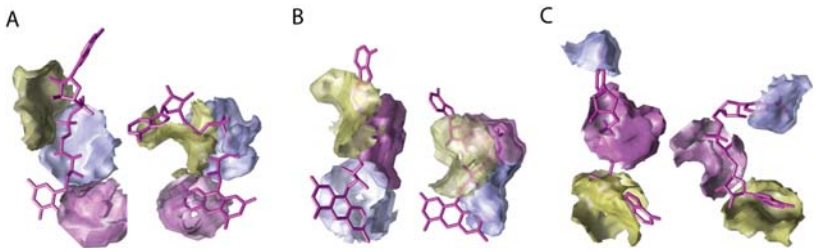

Figure 6.7. Examples of pocket matches of the flexible ligands FAD and NAD. In all three cases, the retrieval rank of local 3DZD method ranks higher than global 3DZD method for the proteins in the right for the query proteins in the left. Protein 1jr8 (**A** right) for query protein 1cqx (**A** right) that both bind to FAD is ranked 3rd for local 3DZD and 31st for global 3DZD, 1k87 (**B** right) for query 1e8g (**B** right) that also binds to FAD ranked 1st for local 3DZD and 18th for global 3DZD, and 1s7g (**C** right) for query 1mi3 **C** right) that both bind to NAD is ranked 2nd for local 3DZD and 16th for global 3DZD.

Example pairs of binding pockets that have different conformations but that are bound to the same ligand types, which their retrieval rank of the local 3DZD method ranks higher than their rank using the global 3DZD method, could be found (Fig. 6.7). However, there were no significant improvements of using the local 3DZD method for the proteins that are bound to the two most flexible ligands, FAD and NAD. The increase in the ROC AUC value for FAD binding proteins was 0.13 and for the NAD binding proteins the increase was only 0.03 (Table 6.3). Moreover, the ligand prediction results for the NAD binding proteins using the local 3DZD method was worse than the results of using the global 3DZD method. There can be several explanations for not having significant improvements for the FAD and NAD binding proteins and one possibility is that the local 3DZD method recognizes ligand sub-components that are shared by several other ligand types. The local 3DZD method recognizing the ligand sub-component is studied in the next section.

6.2.7 Retrieving proteins of same chemical component using local 3DZD method

Although all nine ligands are different, some share the same chemical subcomponents. That is, AMP, ATP, FAD, and NAD all contain adenosine (adenosine moiety) and FAD and FMN both contain flavin (flavin moiety). Since the proposed method captures the local similarities, the database search results in some cases rank the proteins with same chemical moiety higher in the retrieval.

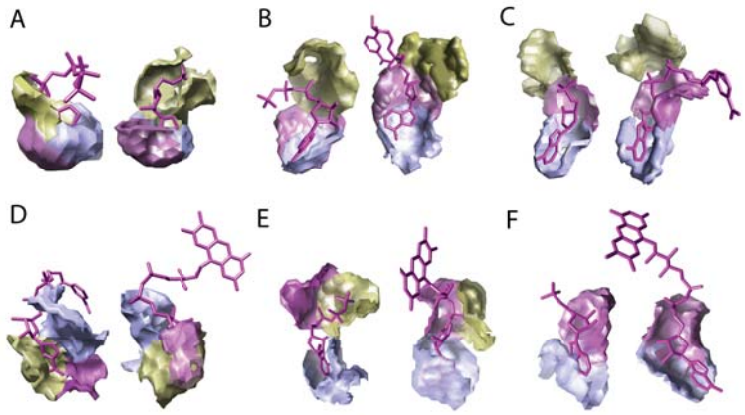

Figure 6.8. Adenosine region matching of AMP, ATP, FAD, and NAD with the local 3DZD method. **A** shows 1b8a (ATP binding) in left and 1kht (AMP) in right. **B** shows 1dy3 (ATP) in left and 1tox (NAD) in right. **C** shows 1kht (AMP) in left and 1ib0 (NAD) on right. **D** shows 1s7g (NAD) in left and 1k87 (FAD) in right . **E** shows 1a49 (ATP) in left and 1k87 (FAD) in right. **F** shows 1amu (AMP) in left and 3grs (FAD) in right. The color codes labels the patch pairs determined by the bipartite matching.

Figure 6.8 shows example match results of how the local 3DZD pairs the adenosine binding regions. The retrieval ranks of proteins with adenosine moieties in Fig. 6.8 are as follows: **A** 1b8a (left) queried on AMP (right) is 1; **B** 1dy3 (left) queried on 1tox (right) is 2; **C** 1ib0 (right) queried on 1kht (left) is 7; **D** 1k87 (right) queried on 1s7g (left) is 4; **E** 1k87 (right) queried on 1a49 (left) is 10; and **F** 3grs (right) queried on 1amu (left) is 27.

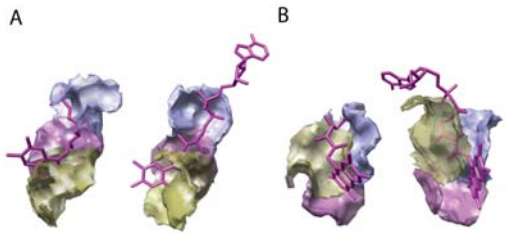

Figure 6.9. Flavin binding region matching with local 3DZD method. **A** shows FMN binding protein 1kht in left and FAD binding protein 1pox in right. **B** shows FMN binding protein 1mvl in left and FAD binding protein 1jqi in right. Flavin binding region matching with local 3DZD method. The color codes labels the patch pairs determined by the bipartite matching.

Figure 6.9 shows two example matches of flavin moiety matching. The retrieval ranks of proteins in flavin moiety in Fig. 6.9 is as follows: **A** 1ja1 (left) queried on 1pox (right) is 12; and **B** 1jqi (right) queried on 1mvl (left) is 17. All pairs of proteins in Fig. 6.8 and Fig. 6.9 rank within 20 in the database retrieval using the local 3DZD method.

The recognition of chemical moieties is an interesting factor that suggests that the local 3DZD method could be used to not only recognize proteins that bind to the same ligand types, but also to ones that bind to ligands within the same chemical moiety. That is, although coupled with other chemical components such as in adenosine connected by phosphate to nicotinamide in NAD and with flavin in FAD or as a subcomponent of a larger ligand as in AMP and ATP, AMP and FAD, and AMP and NAD, the local 3DZD will be able to find a similarity between the same chemical subcomponent. This, however, degrades the ligand prediction accuracy. For example, the AMP binding pocket is often (4 out of 9) recognized as the ATP binding pocket in the Top3 ligand prediction.

6.2.8 Computation time

Evaluated on Intel core i7 at 2.67 GHz and 11GB memory, prediction of a query protein takes on average 2 minutes 29.76 seconds for the local 3DZD method and 31.54 seconds for the global 3DZD method. The prediction process comprises of ligand binding site prediction (bound ligand positions used), protein surface property computation, and computation of the local surface descriptor. Computation time for each step averaged over the dataset of 100 proteins is shown in Table 6.5. As shown in Table 6.5, although the local 3DZD method is sufficiently fast for database searching, the speed of the global 3DZD method is much faster.

Table 6.5
Database search speed of pocket comparison methods

Phase	Task	global 3DZD	local 3DZD
Query preparation	Binding site prediction with LIGSITE	3.12s	3.12s
	Surface property computation with APBS	12.38s	12.38s
	Computation of descriptor	16s[a]	1m 52.96s
Database searching[b]	Dissimilarity computations	0.023s[a]	1.28s
	Ligand prediction	0.02s	0.02s

[a] Taken from previous work in [164] which uses Pentium 4 processor at 3.0 GHz. [b] Total time computed over 100 proteins.

6.3 Chapter summary

This chapter introduced the application of 3DZD to local structure based function prediction. Two methods, both using 3DZD, were introduced to describe local

regions of the protein surface, especially the ligand binding region. The first method, global 3DZD method, simply extracts the LBP region and describes the shape with 3DZD, which has been extensively discussed in our previous publication [164]. The second method, the local 3DZD method, further segments the pocket region to local patches which are described by 3DZDs with additional position information. When the pockets have high similarities, the global 3DZD method provides a fast and efficient method of pocket comparison. However, when only partial regions in the pockets are conserved, comparison of global properties does not provide the best solution. Furthermore, conserved regions have differing geometrical and physicochemical properties, which has also been observed in the works of Kahraman et al. [72, 168]. For these cases, we have shown that the local 3DZD method provides an alternative method of analyzing the similarities of the LBPs. The speed of the local 3DZD method is still reasonable for a database search since structural alignment of pairs of proteins, which is necessary for most structure based comparison methods, is not required.

The local structure based function prediction procedure includes two steps, detection of a potential LBP in a query protein followed by matching of the query pocket against a database of known LBPs. However, performance also depends largely on the accuracy of the LBP detection step. Therefore, establishing a well-coordinated procedure of detecting (predicting) and searching LBPs remains as an important future direction of this work. Another key avenue for improvement is to investigate the inclusion of other features of pockets into the descriptor, such as the degree of residues conservation or the surface residue themselves.

7 TOWARDS ANNOTATING PROTEIN SURFACES: CHARACTERIZATION AND CLASSIFICIATION OF LOCAL PROTEIN SURFACE PATCHES USING SELF ORGANIZING MAP[1]

In the previous chapter, local surface patches of proteins were used to compare ligand binding sites. This chapter takes a further step in generalizing structure based protein function prediction through charactering protein surface with 3D Zernike descriptor (3DZD) of classified local surface patches.

Although ligand binding is a important function for some proteins, it only occurs in a subset of proteins, e.g., enzymes and proteins that use co-factors and substrates. Moreover, binding locations on average cover only 5% of the entire protein surface [142]. Hence, there are portion of surface regions yet to be studied. Ideally, structure based function prediction methods should be more general so that they can describe, compare, and annotate local patches of the whole protein surface. Thus, it is critical to establish computational methods for characterizing and classifying local structures of proteins surfaces.

Toward this goal, local surface patches of proteins from various families are first classified to see how diverse local surface patches are and how many different types of surface patches exist. Then surface patches are characterized by two features, the geometric shape and the electrostatic potential, and represented by 3DZDs. Classification of surface patches is done by the emergent self-organizing map (ESOM) [170, 171], which is a variant of a self-organizing map that can handle a large numbers of neurons. Application of the ESOM to 118,003 surface patches taken from 609 proteins yields 30∼50 clusters. To demonstrate practical the usefulness of this classification,

[1] This chapter is reuses of published work in [169]

classes of local surface patches occurring at the binding sites of six ligand molecules are examined.

7.1 Materials and methods

Figure 7.1 illustrates the procedure used in this study. First, for a set of proteins, surfaces are extracted and the electrostatic potentials on the protein surfaces are computed. Then, protein surfaces are segmented into local patch, using a sphere of a fixed size. Each local surface patches is characterized by the 3DZDs of the surface shape and the electrostatic potentials on the surface. Next, ESOM [170, 171] is used to classify the surface patches in the data set. Three ESOM maps are constructed: the first for classifying surface shape of the patches, the second for classifying the electrostatic potential of patches, and the third for classifying the combination of surface shape and electrostatic potential. The resulting ESOM maps are further clustered to group patches with similar characteristics. We also, compared the binding pockets of six different ligands in terms of cluster group combinations of their binding pockets.

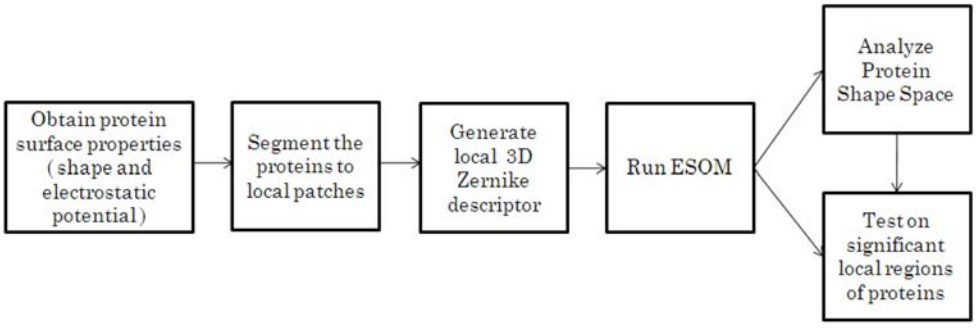

Figure 7.1. Flowchart of protein surface classification method.

7.1.1 Data set

The data set of representative proteins used in this study consists of 609 protein chains, each taken from a different family defined in the SCOP database [108]. These families belong to 72 SCOP superfamilies. These 72 superfamilies were selected so that each superfamily has at least three and no more than twenty families. Selected structures have a crystallographic resolution of 3.0Å or better, have no more than 10 missing residues in the structure solved, have all heavy atom positions solved, are longer than 100 residues, and the structure similarity of each pair is less than a Z-score of 3.8 by the Combinatorial Extension (CE) program [29].

The ligand binding protein data set used later in this study is composed of subsets of proteins, again taken from the work of Kahraman et al. [72]. That is, AMP, ATP, FAD, FMN, NAD, and HEM binding proteins in Table 6.1 are used. The conformations of ATP (adenosine-5'-triphosphate) and AMP (adenosine-5'-diphosphate) are similar. FAD (flavin-adenine dinucleotide) and NAD (nicotinamide-adenine-dinucleotide) are relatively flexible and take diverse conformations when bound to proteins. FMN (flavin mononucleotide) and HEM (haemoglobin) are more distinct in shape compared to the others in the data set. To obtain exact ligand binding surface locations, surface regions that are closer than 3.5Å to the binding ligand atoms are used.

7.1.2 Segmenting protein surface to local patches

Global shape and electrostatic potentials of the protein surface are obtained as described in section 6.1.2. The protein surface is segmented into local surface patches following similar steps as described in the previous chapter. The protein surface segmentation step in this chapter is as follows. First, seed points, or the centroids of the local surfaces, are selected by taking every 100th surface voxel starting from a random point on surface. The average number of seed points in a protein is approximately 194. The local patches are then extracted by taking a surface region that is within

6Å of each seed point. The resulting voxels are considered as an input 3D function, $f(\mathbf{x})$, which is expanded into the 3DZD as described in section 2.5.2. 118,003 local surface patches are obtained from the 609 proteins in the data set. In this work, 3DZDs of order 15 are used to describe the local patches of proteins. This is 72 invariants for shape 3DZD and 144 invariants for the electrostatic potential 3DZD.

For a descriptor that describes both shape and the electrostatic potential, instead of using a linear combination of 3DZDs for shape and electrostatic potential separately, as was done in the previous chapter, a simple concatenation of a 3DZD for shape and 3DZD for electrostatic potential is used. This yields a 3DZD of 216 (i.e. 72+144) invariants. Since the ESOM assumes that the data have a distribution that is close to a Gaussian distribution, log normalization, which is a normalization method that are used to make a distribution look more Gaussian, of the 3DZDs are performed instead of the normalizations described in section 2.5.2.

7.1.3 SOM and ESOM

The self-organizing map (SOM) is a data classification technique which reduces the dimensions of data through the use of self-organizing neural networks [172]. The output of a SOM is a lower dimension map, usually 2D, of neurons which groups similar data items closely together on the output map. Each neuron on the map has a vector of weights, which are trained based on a set of input training data. The weights assigned to each neuron are composed of two components. The first part represents the feature vector of the neuron, which has the same dimension as the input vectors. The second part represents its location in the 2D map. SOM tries to train the neurons so that they are similar to the distribution of the input vectors. The SOM algorithm first initializes the weight vector map. Among several methods available for the initialization, sampling from range of 0 to 1 from Gaussian distribution is that is centered at 0.5 used. The feature weights of neurons and their location in the 2D map are then updated iteratively until change is small or until the iteration

number exceeds the threshold. The process consists of three steps: 1) selecting an input feature vector; 2) finding the neuron that has the feature weight closest to the input; and 3) updating the weights of the closest neuron and its neighboring neurons and updating the position of neighboring neurons so that similar neurons are locate closer to each other in the output map [172].

Emergent SOM (ESOM) is a variation of SOM, which can handle a larger number of neurons (around 4000) and uses boundless maps [170, 173]. It embeds the maps in a finite boundless space such as a sphere or toroid. Two visualization methods of the ESOM maps are used, namely the P-matrix and the U-matrix. The P-matrix visualizes the density in the input data space using the Pareto density estimation. In general, it is suitable for dealing with slowly changing densities and overlapping clusters. The U-matrix visualizes neurons on an ESOM map by a color coding that represents the sum of distances to all immediate neighbors normalized by the largest value of the neighboring neurons. Generally, the U-matrix is appropriate for handling data points which are clearly separated from each other. The ESOM program is available at `http://databionic-esom.sourceforge.net/`. The original paper by Ultsch [173] describes the general ESOM training procedure in details.

The advantage of SOM/ESOM is that it is able to provide an intuitive visualization of the similarity of input data. On the other hand, a potential drawback is that often SOM/ESOMs trained on the same data set do not converge well to a similar map. In this work, to examine the convergence of the obtained ESOM map obtained, neurons in the resulting map are further clustered using the affinity propagation clustering method (see the following section). The training data in this work are feature vectors of the 118,003 local surface patches in the 609 proteins. The feature vector used for local surface patches is the 3DZD describing the local shape, the electrostatic potential, or the combination of the two. As explained in the previous section, the 3DZD of the local shape has 72 values, those of the electrostatic potential have 144 values, and the combination of shape and electrostatic potential have 216 (i.e. 72+144) values. The training data set is inputted to the ESOM program and

trained on 4100 neurons. The iteration is terminated when there is small or no change in the trained neurons or when iteration exceeds 1000 iterations.

7.2 Results

7.2.1 ESOM maps

118,003 local patches from the 609 proteins represented by the 3DZD were analyzed with an ESOM of 4100 neurons. The resulting ESOM maps are boundless toroids, i.e. the right edge of the map is continuously connected to the left edge and the upper edge is connected to the lower edge. Three types of input feature vectors are used to characterize the surface patches: the 3DZD of 1) the surface shape, 2) the electrostatic potential, and 3) a combination of the surface shape and the electrostatic potential (Fig. 7.2). The maps are shown in two color-coding schemes, the P-matrix and the U-matrix. Note that these two matrices visualize the same data, although their appearance is different. Five examples of the local patches that corresponds to the labeled positions in each ESOM map are also shown in Figure 7.2.

The U-matrix maps show the mutual distance of the neighboring neurons. In the U-matrix maps of all three types of input feature vectors, a large blue area and a couple of smaller green areas are observed. The blue area indicates that there is a large region of neurons that have small distances to each other. These neurons are primarily composed of patches of more or less flat shapes. Pocket-shaped patches are mapped to the neurons that have a higher distance (green regions). This is reasonable because it indicates that pocket shapes have more diversity in shape compared to the flat regions.

The P-matrix maps show the density of the data sets that are assigned to each neuron and are shaded from dark to light colors as the density decreases. Compared to the U-marix, more distinct clusters are identified. Local patches of dense regions on the P-matrix (dark colored regions) are relatively flat and have a mixture of the positive and negative electrostatic potential that are near 0.0. In the maps for the

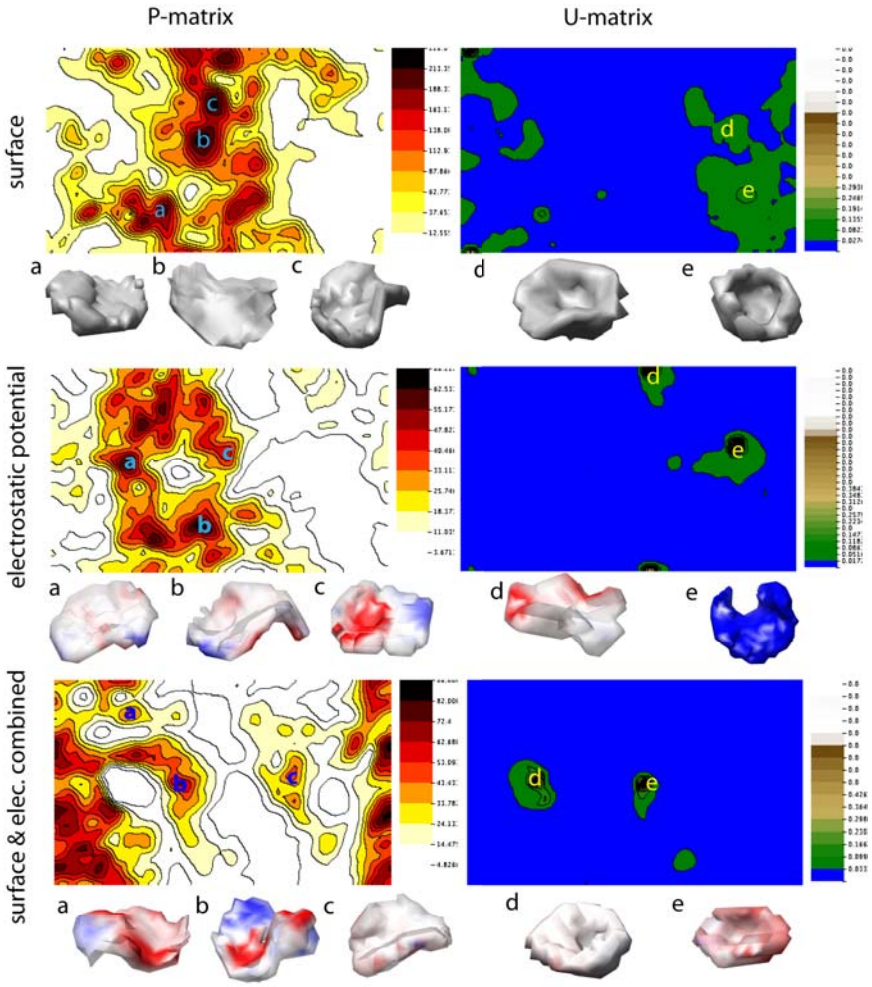

Figure 7.2. ESOM maps of local surface patches. Three types of feature vectors of the local patches are used as input data: the 3DZD of shape, the 3DZD of the electrostatic potential, and 3DZD of the combination. The resulting ESOM maps are visualized as the P-matrix and the U-matrix. For each ESOM map, five examples of the local patches are sampled at the labeled positions. In the surface ESOM map, local patches from 1a2kC (a, b, c), 1b7eA (d), and 1af6C (e) are sampled. In the ESOM map of the electrostatic potential, patches from 1a2kA (a, b, c) and 1aquB (e, f) are shown. In the ESOM map of the combination of the shape and the electrostatic potential, local patches from 1b9mB (a), 1aurB (b, c), and 1a8uB (d, e) are shown.

electrostatic potential (the middle row), clear separation of the patches is not observed except for some cases (e.g., a cluster with a patch from 1aquB at (e) in the U-matrix of the electrostatic potential, which is strongly positively charged). This is because the 3DZD of the electrostatic potential intrinsically convolutes information about the electostatic potential and the shape as the potential is mapped on the surface.

Figure 7.2 shows the overall landscape of the similarity of the surface patches. To obtain more distinct clusters, the 4100 neurons of the ESOM maps were clustered using the affinity propagation clustering method. The affinity propagation clustering method clusters data using a message passing protocol, which has been shown to have a low error rate and to be as fast as other common clustering methods [174]. In the affinity propagation clustering method, number of clusters is influenced by a "preference" parameter. Setting the preference parameter to the median of input distances results in a moderate number of clusters; setting them to the minimum of input distances results in a smaller number of clusters. The ESOM maps for the surface shape (Fig. 7.3**A,B**) and ESOM map of the combination of the shape and the electrostatic potential (Fig. 7.3**C,D**) is used. Setting the preference parameter to the median of input distances results in 369 clusters for the surface ESOM map (Fig. 7.3**A**) and 215 clusters for the surface and the electrostatics combination (Fig. 7.3**C**). When the minimum of the input distances is used as the parameter, the number of clusters is reduced to 48 in the surface ESOM map (Fig. 7.3**B**) and 27 for the surface and the electrostatics combination (Fig. 7.3**D**). These distinct clusters are convenient for labeling a protein surface. Since a local surface patch can be assigned to one of the clusters, a protein can be represented by a set of cluster groups to which the local patches of the protein belong.

7.2.2 Local patch types of ligand binding surfaces

Next, local patches of the ligand binding sites are classified with the resulting cluster groups (Fig. 7.3). Figure 7.4 shows histograms of surface patch clusters

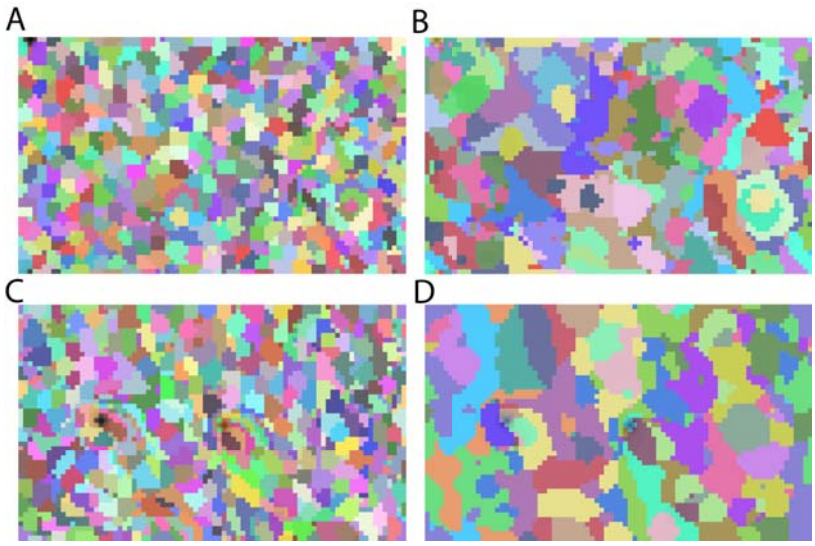

Figure 7.3. Affinity propagation clustering of ESOM neurons. **A** and **B** are clustering results of ESOM neurons using the surface shape, while **C** and **D** are results using the surface shape and the electrostatic potential combination as the feature of local surface patches. In **A** and **C**, the median distance computed between the neurons is used as the preference parameter p. **B** and **D** use the minimum distance as the preference parameter. There are 369 clusters in **A**, 48 clusters in **B**, 215 clusters in **C**, and 27 clusters in **D**. Note that the colors are used just to separate neighboring clusters; clusters in the same color at different locations do not mean similarity.

observed at ligand binding surfaces of the AMP, ATP, FAD, FMN, HEM, and NAD binding proteins listed in Table 6.1. The cluster compositions of each ligand binding surface share some similarities, derived from the fact the binding sites have many pocket-like local shapes. In all of the binding sites, the surface patch cluster 16 and 34 and the surface electrostatic potential clusters 6 and 22 are observed. The molecular shape and the binding sites of FMN are different from the others, which is reflected in the relatively distinct histogram.

Figure 7.5 shows concrete examples of the local patches of the ligand binding sites. They are mapped on the ESOM map of the shape and electrostatic potential combination. The ATP binding site of 1dv2A consists of surface patch clusters 6(c),

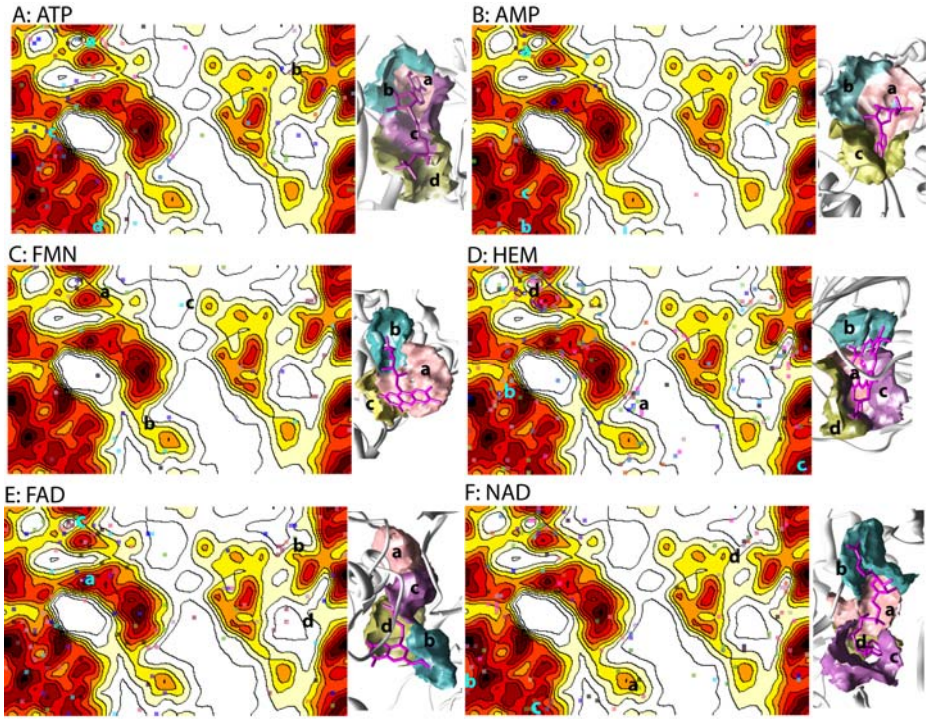

Figure 7.4. Histograms of ESOM neuron clusters of surface patches of binding sites of the six ligand molecules listed in Table 6.1. The clusters are obtained through affinity propagation clustering of ESOM neurons using the minimum distance as the preference parameter, i.e. Fig 7.3**B** when the shape is used and Fig. 7.3**D** when the combination of the surface shape and the electrostatic potential is used as the feature of the surface patches. The x-axis is the index of the cluster and the y-axis is the density of each cluster.

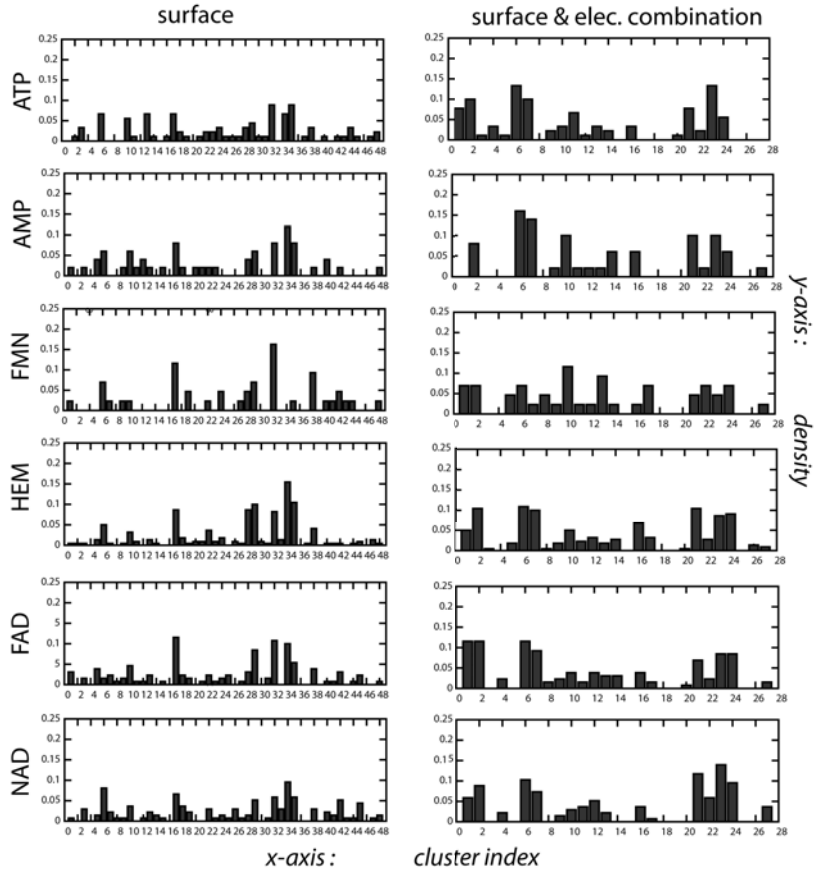

Figure 7.5. Distribution of the ligand binding surface patches on the ESOM map. The ESOM map of the shape and electrostatic potential combination is used (i.e. Fig. 7.2, bottom row). The colored dots specify the location of the neurons the local patches match with. The colors of the dots indicate the source proteins. For each type of ligand, one example of the binding sites is shown on the right side of each ESOM maps. The proteins used are as follows: 1dv2A for ATP binding; 1c0aA for AMP; 1mvlA for FMN; 1np4A for HEM; 1hskA for FAD; and 2npxA for NAD. The labels (a, b, c, d, e) indicate the position of the corresponding local patches on the ESOM map.

13, 16(d), 21(b), and 24(a). The labels shown in Figure 7.4 are in the parentheses. The AMP binding site of 1c0aA is composed of the clusters 6(a), 12(c), and 23(b). The FMN binding site of 1mvlA is composed of the clusters 1(a), 5(c), 10, and 21(b). The heme binding site of 1np4A consists of the clusters 2(c), 6(d), 7(b), 10, and 17(a). The FAD binding site of 1hskA consists of the clusters 6(c), 14(d), 16(b), 20, 23(a), 24, and 27. The NAD binding site of 2npxA is composed of the clusters 1, 2, 7(b), 12(a), 16(d), 22(c), and 23. The binding site patches for all other proteins that are listed in Table 6.1 are also shown in Figure 7.4; visualized with the colored dots. Results showed that a binding site of the same chemical moiety does not always correspond to one cluster group or to set of groups. For example, the adenosine binding site is assigned to the patch 24 in 1dv2A, an ATP binding protein, and to the patch 12 in 1c0aA, an AMP binding protein. This is partly due to the conformational changes of the ligands [72] and also due to the different distribution of the seed points on a protein surface. However, there are cases where binding sites of the same chemical group of ligands are assigned to the same surface patch cluster group, such as the phosphate binding region in 1dv2A and 1c0aA, both of which are assigned to the cluster 6. Overall similarities in local patch grouping can also be found in Figure 7.4, where both ATP and AMP have high peaks at the clusters 6, 7, 23, and 24 in their histograms.

7.2.3 Computation time

The time taken to generate the 3DZDs for each protein depends on the size of the proteins. It takes around two minutes to generate the surface geometry and the surface electrostatic potential information for each protein and an average of 37.12 seconds to compute 3DZDs for the surface patches of a whole protein. The training of an ESOM using the surface patch data with 1000 iterations took approximately two weeks. However, the training is needed only once. Once the ESOM is trained, assignment of patches in a protein to neurons can be done in less than a second. The

analysis was performed on a Linux machine equipped with Intel Core TM2 CPU 6400 at 2.13GHz.

7.3 Chapter summary

In this chapter, the classification of local surface patches of proteins which are characterized by the shape and the electrostatic potential was examined. The use of the 3DZD provides a convenient compact representation of two characteristics of the patches as feature vectors. The ESOM was used for classifying the local patches since it is effective for describing landscape of similarity of data points, whose similarity is rather continuous and not very distinct. Moreover, in order to obtain distinct groups of local surface patches, neurons of the ESOM map trained with the local patches was clustered. The procedure yielded a manageable number of clusters of local patches; 48 clusters for the surface shape, and 27 for the combination of the shape and the electrostatic potentials. I believe that annotating protein surfaces with the classified local patches, where a protein can be described as list of local patch classes, will be applicable to variety of protein surface analyses problems, such as classification, function prediction, and database searches, in analogous ways to protein sequence analyses.

8 CONCLUSION

The importance of computational works in characterization of protein structures has become clear with the increasing number of protein structures of unknown function being solved. However, computational protein structure analysis significantly lags behind sequence analysis in a practical sense, since there are few method that have been developed for large scale global/local structure comparison. To address the necessity for such analysis methods, both global and local protein surface comparison methods using the 3D Zernike descriptor (3DZD) were introduced. In the first chapter a review of existing object comparison methods in the engineering field and a review on how the proteins are compared in the biology field were provided. Among the reviewed methods, the 3DZD had many characteristics that are valuable in representing protein surfaces where the interactions occur.

The global protein surface comparison method was introduced in the following three chapters. In the second chapter, 3DZD was applied for searching protein tertiary structures. Unlike existing methods for structure comparison and representation, the 3DZD allow an extremely rapid database search, which opens up the possibility for a real time protein tertiary structure search on the internet. A search against the benchmark dataset of 2432 proteins used in this work took only 0.46 seconds. A preeminent mathematical property of 3DZD is rotational invariance. This is a significant advantage over spherical harmonics and the multipole representation [175] that require pose normalization prior to comparison of the protein structure. In the third chapter, a comprehensive analysis of four different levels of protein structure representations was done and the information was used to further improve the accuracy of the protein structural search. Evaluated on a data set of 2337 protein chains, in general, three backbone representations coincide well with both CE and SCOP [108] classifications. This is not surprising since both classifications compares the backbone

structure of proteins and not the surface defined on all residue atoms. Looking at individual cases, for flexible structures, where there exist tail-like structures, AASurf performs better. Results also suggest that 3DZD is well suited for representing surfaces generated from low-resolution structural data, such as density maps, as well as structures with specific atomic positions. Although more investigation is needed on the application of the 3DZD to specific uses, 3DZD has been shown to be well suited for representation and comparison of protein structures at low-resolution. The fourth chapter introduced the use of 3DZD for fast quantitative comparison of physicochemical properties defined on the protein surfaces. Using 3DZD, similarities based on properties such as the electrostatic potential and hydrophobicity can be quantified. 3DZD performs better than two traditional density comparison methods, Carbo index [117] based comparison and Hodgkin index [119] based comparison, in its ability to provide meaningful distances with minimal computational effort. Application of 3DZD can be further extended for comparison of the other properties, such as a residue conservation.

In the last two chapters, local protein surface comparison was studied. In the fifth chapter, we have introduced two new ligand binding prediction methods: 1) one using global binding pockets, the global 3DZD method, and 2) one using segmented binding pockets, the local 3DZD method. Both methods have better retrieval performance than the spherical harmonic bases method [72]. The global 3DZD method provides a quick search method while the local 3DZD method allows for comparison of ligand binding pockets with partial similarities. Since the suggested methods do not predict the binding locations of the proteins, establishing a well-coordinated procedure of detecting and searching ligand binding location is left as an important future direction of this work. In the fifth chapter, local surface patches of proteins are classified towards the goal of annotating protein surfaces with the classified patches. The properties used are characterized by the shape and the electrostatic potential. ESOM was used for classifying the local patches. ESOM [170] is effective for observing the landscape of similarity of data points, whose similarity is rather continuous and not

very distinct. Moreover, in order to obtain distinct groups of local surface patches, we have clustered neurons of the ESOM map with the affinity propagation clustering method [174]. The procedure yielded a manageable cluster size of local patches: 48 clusters for the surface shape, and 27 for the combination of the shape and the electrostatic potentials. Although more investigation is necessary for the structure based function annotation, we believe that representatives of the local surface patches clusters are useful means of surface annotation.

Surface shape representation by 3D Zernike descriptors has numerous applications. One possible application is to analyze surfaces of proteins or biological molecules with similar functions but different main-chain or molecular structure, such as binding sites of DNA-binding proteins, or proteins that display structural mimicry. Another application might be for a protein-protein interaction analysis. Biology has entered an informatics era where an efficient reuse of knowledge from existing databases is crucial. In biological sequence comparisons, BLAST [42] and FASTA [176] have enabled fast database searches for more than a decade and have revolutionized biological research. In contrast, comparison of protein 3D structures is still in the realm of pairwise comparison, in which a 3D structure database search may take hours, rendering 3D structure search impractical and hindering the development of novel tools/applications based on fast structure searches. I believe that the methods introduced here provide a foundation for new methods for quick real-time protein surface based function assignment by comparing protein surfaces of known function, which are similarly fast as the routinely used global and local sequence motif based annotation.

LIST OF REFERENCES

LIST OF REFERENCES

[1] J. B. Hagen. The origins of bioinformatics. *Nature Reviews Genetics*, 1(3):231–236, December 2000.

[2] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, 181(4610):662–666, March 1958.

[3] H. Muirhead and M. F. Perutz. Structure of heamoglobin: A three-dimensional Fourier synthesis of reduced human heamoglobin at 5.5Å resolution. *Nature*, 199:633–638, August 1963.

[4] C. A. Ouzounis and A. Valencia. Early bioinformatics: The birth of a discipline – A personal view. *Bioinformatics*, 19(17):2176–2190, November 2003.

[5] M. O. Dayhoff and R. S. Ledley. Comprotein: A computer program to aid primary protein structure determination. In *American Federation of Information Processing Societies '62 (Fall): Proceedings of the December 4-6, 1962, fall joint computer conference*, pages 262–274, New York, NY, USA, 1962. ACM.

[6] M. O. Dayhoff. Computer aids to protein sequence determination. *Journal of Theoretical Biology*, 8(1):97–112, January 1965.

[7] M. Dayhoff. *Atlas of protein sequence and structure*. National Biomedical Research Foundation, 1965.

[8] E. Zuckerkandl and L. Pauling. Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, 8(2):357–366, March 1965.

[9] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155(760):279–284, January 1967.

[10] C. Levinthal. Molecular model-building by computer. *Scientific American*, 214(6):42–52, June 1966.

[11] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, March 1970.

[12] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3):441–448, May 1975.

[13] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74(2):560–564, February 1977.

[14] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, March 1981.

[15] D. J. Lipman and W. R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, March 1985.

[16] R. D. Fleischmann, M.D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, and et al. Whole-genome random sequencing and assembly of haemophilus influenzae Rd. *Science*, 269(5223):496–512, July 1995.

[17] D. Chen. Structural genomics: Exploring the 3D protein landscape. *Biomedical Computation Review*, pages 11–19, 2010.

[18] About PSI: PSI-Nature structural genomics knowledgebase. http://kb.psi-structuralgenomics.org/about/psi.html, retrieved June 10, 2010.

[19] H. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, 10(12):980, December 2003.

[20] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[21] T. Hawkins, S. Luban, and D. Kihara. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Science*, 15(6):1550–1556, June 2006.

[22] M. Chitale, T. Hawkins, C. Park, and D. Kihara. ESG: Extended similarity group method for automated protein function prediction. *Bioinformatics*, 25(14):1739–1745, July 2009.

[23] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5(4):823–826, April 1986.

[24] D. Kihara and J. Skolnick. Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q. *Proteins: Structure, Function, and Bioinformatics*, 55(2):464–473, 2004.

[25] R. A. Laskowski, J. D. Watson, and J. M. Thornton. Protein function prediction using local 3D templates. *Journal of Molecular Biology*, 351(3):614–626, August 2005.

[26] A. Martin, C. A. Orengo, E. G. Hutchinson, S. Jones, M. Karmirantzou, R. A. Laskowski, J. Mitchell, C. Taroni, and J. M. Thornton. Protein folds and functions. *Structure*, 6(7):875–884, July 1998.

[27] K. Mizuguchi and N. Go. Seeking significance in three-dimensional protein structure comparisons. *Current Opinion in Structural Biology*, 5(3):377–382, June 1995.

[28] D. Kihara and J. Skolnick. The PDB is a covering set of small protein structures. *Journal of Molecular Biology*, 334(4):793–802, December 2003.

[29] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 11(9):739–47, September 1998.

[30] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 233(1):123–138, September 1993.

[31] K. Kinoshita and H. Nakamura. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Science*, 12(8):1589–1595, 2003.

[32] V. Venkatraman, L. Sael, and D. Kihara. Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors. *Cell Biochemistry and Biophysics*, 54(1-3):23–32, 2009.

[33] A. Gutteridge and J. M. Thornton. Understanding nature's catalytic toolkit. *Trends in Biochemical Sciences*, 30(11):622–629, November 2005.

[34] C. Winter, A. Henschel, W. Kim, and M. Schroeder. SCOPPI: A structural classification of protein-protein interfaces. *Nucleic Acids Research*, 34(Database issue):D310–314, 2006.

[35] E. R. Jefferson, T. P. Walsh, T. J. Roberts, and G. J. Barton. SNAPPI-DB: A database and API of structures, interfaces and alignments for protein-protein interactions. *Nucleic Acids Research*, 35(Database issue):D580–589, 2007.

[36] M. Novotni and R. Klein. 3D Zernike descriptors for content based shape retrieval. In *Proceedings of the Eighth ACM Symposium on Solid Modeling and Applications*, pages 216–225, Seattle, Washington, USA, 2003. ACM.

[37] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH: A hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, August 1997.

[38] L. Lo Conte, S. E. Brenner, T. J. P Hubbard, C. Chothia, and A. G. Murzin. SCOP database in 2002: Refinements accommodate structural genomics. *Nucleic Acids Research*, 30(1):264–267, 2002.

[39] T. Madej, J. F. Gibrat, and S. H. Bryant. Threading a database of protein cores. *Proteins*, 23(3):356–369, November 1995.

[40] K. Kinoshita, Y. Murakami, and H. Nakamura. eF-seek: Prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape. *Nucleic Acids Research*, 35(Web Server issue):W398–402, July 2007.

[41] R. A. Laskowski, J. D. Watson, and J. M. Thornton. ProFunc: A server for predicting protein function from 3D structure. *Nucleic Acids Research*, 33(suppl_2):W89–93, July 2005.

[42] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997.

[43] L. Sael and D. Kihara. Protein surface representation and comparison: New approaches in structural proteomics. In *Biological Data Mining*, Chapman & Hall/CRC data mining and knowledge discovery series., pages 89–109. Chapman & Hall/CRC Press, USA, 2010.

[44] J. Tangelder and R. Veltkamp. A survey of content based 3D shape retrieval methods. In *SMI '04: Proceedings of the Shape Modeling International 2004* , pages 145–156, 2004.

[45] M. L Connolly. Solvent-accessible surfaces of proteins and nucleic-acids. *Science*, 221:709–713, 1983.

[46] X. Wang. Alpha-surface and its application to mining protein data. In *ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 659–662, Washington, DC, USA, 2001. IEEE Computer Society.

[47] M. Elad, A. Tal, and S. Ar. Directed search in a 3D objects database using SVM. Technical report, HP Laboratories, Israel, 2000.

[48] C. Zhang and T. Chen. Effective feature extraction for 2D/3D objects in mesh representation. In *Proceedings in 2001 International Conference on Image Processing*, pages 935–938, 2001.

[49] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin. Shape distributions. *ACM Transactions on Graphics*, 21(4):807–832, 2002.

[50] M. L. Connolly. Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interface. *Biopolymers*, 25(7):1229–1247, July 1986.

[51] B. S. Duncan and A. J. Olson. Approximation and visualization of large-scale motion of protein surfaces. *Journal of Molecular Graphics*, 13(4):250–257, 1995.

[52] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz. Rotation invariant spherical harmonic representation of 3D shape descriptors. In *SGP '03: Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 156–164, Aire-la-Ville, Switzerland, 2003. Eurographics Association.

[53] T. Funkhouser and P. Shilane. Partial matching of 3D shapes with priority-driven search. In *Symposium on Geometry Processing*, June 2006.

[54] N. Canterakis. 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition. *In 11th Scandinavian Conference on Image Analysis*, pages 85—93, 1999.

[55] H. Laga, H. Takahashi, and M. Nakajima. Spherical wavelet descriptors for content-based 3D model retrieval. *SMI '06: Proceedings of the Shape Modeling International 2006*, 0:15, 2006.

[56] A. Mademlis, A. Axenopoulos, P. Daras, D. Tzovaras, and M. G. Strintzis. 3D content-based search based on 3D Kawtchouk moments. In *3DPVT '06: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission*, pages 743–749, Washington, DC, USA, 2006. IEEE Computer Society.

[57] G. Leifman, S. Katz, A. Tal, and R. Meir. Signatures of 3D models for retrieval. In *4th Israel Korea Bi-National Conference on Geometric Modeling and Computer Graphics*, pages 159–163, 2003.

[58] M. Yu, I. Atmosukarto, W. K. Leow, Z. Huang, and R. Xu. 3D model retrieval with morphing-based geometric and topological feature maps. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 656–661, 2003.

[59] D. Y. Chen, M. Ouhyoung, X. P. Tian, Y. T. Shen, and M. Ouhyoung. On visual similarity based 3D model retrieval. In *Eurographics*, pages 223-232, Granada, Spain, 2003.

[60] R. Ohbuchi, M. Nakazawa, and T. Takei. Retrieving 3D shapes based on their appearance. In *MIR '03: Proceedings of the 5th ACM SIGMM Iternational workshop on Multimedia Information Retrieval*, pages 39–45, New York, NY, USA, 2003. ACM Press.

[61] J. J. Koenderink and A. J. van Doorn. Surface shape and curvature scales. *Image and Vision Computing*, 10(8):557–564, October 1992.

[62] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):433–449, 1999.

[63] P. A. de Alarcón, A. D. Pascual-Montano, and J. M. Carazo. Spin images and neural networks for efficient content-based retrieval in 3D object databases. In *CIVR '02: Proceedings of the International Conference on Image and Video Retrieval*, pages 225–234, London, UK, 2002. Springer-Verlag.

[64] S. J. Pickering, A. J. Bulpitt, N. Efford, N. D. Gold, and D. R. Westhead. AI-based algorithms for protein surface comparisons. *Computers and Chemistry*, 26(1):79–84, December 2001.

[65] C. Hofbauer, H. Lohninger, and A. Aszodi. SURFCOMP: A novel graph-based approach to molecular surface comparison. *Journal of Chemical Information and Modeling*, 44(3):837–847, 2004.

[66] L. Baldacci, M. Golfarelli, A. Lumini, and S. Rizzi. Clustering techniques for protein surfaces. *Pattern Recognition*, 39(12):2370–2382, 2006.

[67] M. Rosen, L. Shuo Liang, and W. Haim. Molecular shape comparison in search for active sites and functional similarity. *Protein Engneering*, 11(4):263–277, 1998.

[68] S. L. Lin, R. Nussinov, D. Fischer, and H. J. Wolfson. Molecular surface representations by sparse critical points. *Proteins*, 18(1):94–101, 1994.

[69] D. Fischer, S. Lin, H. L. Wolfson, and R. Nussinov. A geometry-based suite of molecular docking processes. *Journal of Molecular Biology*, 248(2):459–477, 1995.

[70] I. Halperin, B. Ma, H. L. Wolfson, and R. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–443, June 2002.

[71] M Gerstein. A resolution-sensitive procedure for comparing protein surfaces and its application to the comparison of antigen-combining sites. *Acta Crystallographica*, A(48):271–276, 1992.

[72] A. Kahraman, R. J. Morris, R. A. Laskowski, and J. M. Thornton. Shape variation in protein binding pockets and their ligands. *Journal of Molecular Biology*, 368(1):283–301, April 2007.

[73] D. Ritchie and G. Kemp. Protein docking using spherical polar Fourier correlations. *Proteins*, 39:179–194, 2000.

[74] B. B. Masek, A. Merchant, and J. B. Matthew. Molecular skins: A new concept for quantitative shape matching of a protein with its small molecule mimics. *Proteins: Structure, Function, and Genetics*, 17(2):193–202, 1993.

[75] A. R. Poirrette, P. J. Artymiuk, D. W. Rice, and P. Willett. Comparison of protein surfaces using a genetic algorithm. *Journal of Computer-Aided Molecular Design*, 11(6):557–569, November 1997.

[76] M. Bock, C. Garutti, and C. Guerra. Discovery of similar regions on protein surfaces. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 14(3):285–299, April 2007.

[77] Z. Shentu, M. Al Hasan, C. Bystroff, and M. J. Zaki. Context shapes: Efficient complementary shape matching for protein-protein docking. *Proteins*, 70(3):1056–1073, February 2008.

[78] A. K. Arakaki, Y. Zhang, and J. Skolnick. Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics*, 20(7):1087–1096, May 2004.

[79] F. Ferre, G. Ausiello, A. Zanzoni, and M. Helmer-Citterich. Functional annotation by identification of local surface similarities: A novel tool for structural genomics. *BMC Bioinformatics*, 6:194, 2005.

[80] T. A. Binkowski, L. Adamian, and J. Liang. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *Journal of Molecular Biology*, 332(2):505–526, September 2003.

[81] N. Canterakis. Fast 3D Zernike moments and invariants. Technical report, Ludwigs University Freiburg, ALU Freiburg, Institute for informatics, Freiburg 79110, Germany, 1997.

[82] M. Novotni and R. Klein. Shape retrieval using 3D Zernike descriptors. *Computer-Aided Design*, 36(11):1047–1062, 2004.

[83] A. Kahraman, R. Morris, R. Laskowski, and J. Thornton. Variation of geometrical and physicochemical properties in protein binding pockets and their ligands. *BMC Bioinformatics*, 8(Suppl 8):S1, 2007.

[84] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin, and D. Jacobs. A search engine for 3D models. *ACM Transactions on Graphics* , 22(1):83–105, 2003.

[85] M. Hu. Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory*, 8(2):179–187, 1962.

[86] Y. Sheng and H. Arsenault. Experiments on pattern recognition using invariant Fourier-Mellin descriptors. *Journal of the Optical Society of America A*, 3(6):771–776, June 1986.

[87] D. Casasent and D. Psaltis. Scale invariant optical transform. *Optical Engineering*, 15:258–261, 1976.

[88] C. H. Teh and R. T. Chin. On image analysis by the methods of moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):496–513, 1988.

[89] N. H. Foon, Y. Pang, A. Teoh, B. Jin, and D. N. Chek Ling. An efficient method for human face recognition using wavelet transform and Zernike moments. In *CGIV '04: Proceedings of International Conference on Computer Graphics, Imaging and Visualization*, pages 65–69, Penang, Malaysia, 2004.

[90] M. Asadi, A. Vahedi, and H. Amindavar. Leukemia cell recognition with Zernike moments of holographic images. In *NORSIG '06: Proceedings of the 7th Nordic Signal Processing Symposium*, pages 214–217, Reykjavik, Iceland, 2006.

[91] B. Bayraktar, P. P. Banada, E D. Hirleman, A. K. Bhunia, J. P. Robinson, and B. Rajwa. Feature extraction from light-scatter patterns of listeria colonies for identification and classification. *Journal of Biomedical Optics*, 11(3):3400–3406, June 2006.

[92] J. Yeh, D. Chen, B. Chen, and M. Ouhyoung. A web-based three-dimensional protein retrieval system by matching visual similarity. *Bioinformatics*, 21(13):3056–3057, July 2005.

[93] L. Sael, B. Li, D. La, Y. Fang, K. Ramani, R. Rustamov, and D. Kihara. Fast protein tertiary structure retrieval based on global surface shape similarity. *Proteins*, 72(4):1259–73, September 2008.

[94] A. Murzin, S. Brenner, T. Hubbard, and C. Chothia. SCOP: A structural classification of proteins database for the investigation of sequence and structures. *Journal of Molecular Biology* , 247:536–540, 1995.

[95] M. L. Connolly. The molecular surface package. *Journal of Molecular Graphics*, 11(2):139–141, June 1993.

[96] L. Holm and J. Park. DaliLite workbench for protein structure comparison. *Bioinformatics* , 16(6):566–567, June 2000.

[97] A. Andreeva, D. Howorth, S. E. Brenner, Tim J. P. Hubbard, C. Chothia, and A. G. Murzin. SCOP database in 2004: Refinements integrate structure and sequence family data. *Nucleic Acids Research*, 32(suppl_1):D226–229, 2004.

[98] M. L Sierk, and W.R. Pearson. Sensitivity and selectivity in protein structure comparison. *Protein Science*, 13:773–785, 2004.

[99] J. Felsenstein. PHYLIP – Phylogeny inference package (Version 3.2). *Cladistics*, 5:166, 164, 1989.

[100] D. La, J. Esquivel-Rodriguez, V. Venkatraman, B. Li, L. Sael, S. Ueng, S. Ahrendt, and D. Kihara. 3D-SURFER: Software for high-throughput protein surface comparison and analysis. *Bioinformatics*, 25(21):2843–2844, November 2009.

[101] L. Sael and D. Kihara. Improved real-time protein structure search with application to low-resolution data. submitted in *GIW2010: Proceedings of the 21st International Conference on Genome Informatics*.

[102] J. I. Garzon, J. Kovacs, R. Abagyan, and P. Chacon. ADP_EM: Fast exhaustive multi-resolution docking for high-throughput coverage. *Bioinformatics*, 23(4):427–433, February 2007.

[103] M. Topf, K. Lasker, B. Webb, H. Wolfson, W. Chiu, and A. Sali. Protein structure fitting and refinement guided by cryo-EM density. *Structure* , 16(2):295–307, February 2008.

[104] T. Kawabata. Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a Gaussian mixture model. *Biophysical Journal*, 95(10):4643–4658, 2008.

[105] H. Ceulemans and R. B. Russell. Fast fitting of atomic structures to low-resolution electron density maps by surface overlap maximization. *Journal of Molecular Biology*, 338(4):783–793, May 2004.

[106] F. DiMaio, M. D. Tyka, M. L. Baker, W. Chiu, and D. Baker. Refinement of protein structures into low-resolution density maps using Rosetta. *Journal of Molecular Biology*, 392(1):181–190, September 2009.

[107] J. A Velazquez-Muriel, C. O. Sorzano, S. H. Scheres, and J. M. Carazo. SPI-EM: Towards a tool for predicting CATH superfamilies in 3D-EM maps. *Journal of Molecular Biology*, 345(4):759–771, 2005.

[108] A. Andreeva, D. Howorth, J. Chandonia, S. E. Brenner, T. Hubbard, C. Chothia, and A. G. Murzin. Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Research*, 36(suppl_1):D419–425, 2008.

[109] J. Davis and M. Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240, Pittsburgh, Pennsylvania, 2006.

[110] M. F. Sanner, A. J. Olson, and J. C. Spehner. Reduced surface: An efficient way to compute molecular surfaces *Biopolymers* 38(3):305–320, 1996.

[111] M. Yu. Lobanov, N. S. Bogatyreva, and O. V. Galzitskaya. Radius of gyration as an indicator of protein structure compactness. *Molecular Biology*, 42(4):701–706, 2008.

[112] S. J. Ludtke, P. R. Baldwin, and W. Chiu. EMAN: Semiautomated software for high-resolution single-particle reconstructions. *Journal of Structural Biology*, 128(1):82–97, December 1999.

[113] W. Jiang, M. L. Baker, J. Jakana, P. R. Weigele, J. King, and W. Chiu. Backbone structure of the infectious epsilon15 virus capsid revealed by electron cryomicroscopy. *Nature*, 451(7182):1130–1134, February 2008.

[114] L. Sael, D. La, B. Li, R. Rustamov, and D. Kihara. Rapid comparison of properties on protein surface. *Proteins*, 73(1):1–10, October 2008.

[115] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.

[116] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–1149, 1995.

[117] R. Carbo, L. Leyda, and M. Arnau. An electron density measure of the similarity between two compounds. *International Journal of Quantum Chemistry* , 17:1185–1189, 1980.

[118] E. E. Hodgkin and W. G. Richards. A semiempirical method for calculating molecular similarity. *Journal of the Chemical Society – Chemical Communications*, pages 1342–1344, 1986.

[119] E. E. Hodgkin and W. G. Richards. Molecular similarity based on electrostatic potential and electric field. *International Journal of Quantum Chemistry*, 32(S14):105–110, 1987.

[120] N. Nikolova and J. Jaworska. Approaches to measure chemical similarity – A review. *QSAR & Combinatorial Science*, 22(9-10):1006–1026, 2003.

[121] N. Blomberg, R. R. Gabdoulline, M. Nilges, and R. C. Wade. Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity. *Proteins*, 37(3):379–387, November 1999.

[122] F. De Rienzo, R. R. Gabdoulline, M. C. Menziani, P.G. De Benedetti, and R. C. Wade. Electrostatic analysis and Brownian dynamics simulation of the association of plastocyanin and cytochrome f. *Biophysical Journal*, 81(6):3090–3104, 2001.

[123] K. Schleinkofer, U. Wiedemann, L. Otte, T. Wang, G. Krause, H. Oschkinat, and R. C. Wade. Comparative structural and energetic analysis of WW domain-peptide interactions. *Journal of Molecular Biology*, 344(3):865–881, November 2004.

[124] P. J. Winn, T. L. Religa, J. N. Battey, A. Banerjee, and R. C. Wade. Determinants of functionality in the ubiquitin conjugating enzyme family. *Structure*, 12(9):1563–1574, September 2004.

[125] J. M. Sasin, A. Godzik, and J. M. Bujnicki. SURF'S UP! – Protein classification by surface comparisons. *Journal of Biosciences*, 32(1):97–100, 2007.

[126] K. Pawlowski and A. Godzik. Surface map comparison: Studying function diversity of homologous proteins. *Journal of Molecular Biology*, 309(3):793–806, 2001.

[127] X. Zhang, C. L. Bajaj, B. Kwon, T. J. Dolinsky, J. E. Nielsen, and N. A. Baker. Application of new multiresolution methods for the comparison of biomolecular electrostatic properties in the absence of global structural similarity. *Multiscale Modeling & Simulation*, 5(4):1196–1213, 2006.

[128] N. A Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, 98(18):10037–10041, 2001.

[129] K. Kinoshita and H. Nakamura. eF-site and PDBjViewer: Database and viewer for protein functional sites. *Bioinformatics*, 20(8):1329–1330, 2004.

[130] O. H. Kapp, L. Moens, J. Vanfleteren, C. N. Trotman, T. Suzuki, and S. N. Vinogradov. Alignment of 700 globin sequences: Extent of amino acid substitution and its correlation with variation in volume. *Protein Science*, 4(10):2179–2190, October 1995.

[131] H. Aronson, W. E. Royer-Jr., and W. A. Hendrickson. Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. *Protein Science*, 3(10):1706–1711, 1994.

[132] J. Lecomte, D. A. Vuletich, and A. M. Lesk. Structural divergence and distant relationships in proteins: Evolution of the globins. *Current Opinion in Structural Biology*, 15(3):290–301, June 2005.

[133] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, May 1982.

[134] C. Tarricone, A. Galizzi, A. Coda, P. Ascenzi, and M. Bolognesi. Unusual structure of the oxygen-binding site in the dimeric bacterial hemoglobin from vitreoscilla sp. *Structure*, 5(4):497–507, April 1997.

[135] A. Pesce, S. Dewilde, L. Kiger, M. Milani, P. Ascenzi, M. C. Marden, M. L. Van Hauwaert, J. Vanfleteren, L. Moens, and M. Bolognesi. Very high resolution structure of a trematode hemoglobin displaying a TyrB10-TyrE7 heme distal residue pair and high oxygen affinity. *Journal of Molecular Biology*, 309(5):1153–1164, June 2001.

[136] E. H. Harutyunyan, T. N. Safonova, I. P. Kuranova, A. N. Popov, A. V. Teplyakov, G. V. Obmolova, B. K. Valnshtein, G. G. Dodson, and J. C. Wilson. The binding of carbon monoxide and nitric oxide to leghaemoglobin in comparison with other haemoglobins. *Journal of Molecular Biology*, 264(1):152–161, November 1996.

[137] T. Hankeln, B. Ebner, C. Fuchs, F. Gerlach, M. Haberkamp, T. L. Laufs, A. Roesner, M. Schmidt, B. Weich, S. Wystub, S. Saaler-Reinhardt, S. Reuss, M. Bolognesi, D, De Sanctis, M. C. Marden, L. Kiger, L. Moens, S. Dewilde, E. Nevo, A. Avivi, R. E. Weber, A. Fago, and T. Burmester. Neuroglobin and cytoglobin in search of their role in the vertebrate globin family. *Journal of Inorganic Biochemistry*, 99(1):110–119, 2005.

[138] M. Torrez, M. Schultehenrich, and D. R. Livesay. Conferring thermostability to mesophilic proteins through optimized electrostatic surfaces. *Biophysical Journal*, 85(5):2845–2853, 2003.

[139] A. R. Kinjo and K. Nishikawa. Comparison of energy components of proteins from thermophilic and mesophilic organisms. *European Biophysics Journal*, 30(5):378–384, 2001.

[140] L. Xiao and B. Honig. Electrostatic contributions to the stability of hyperthermophilic proteins. *Journal of Molecular Biology*, 289(5):1435–1444, June 1999.

[141] K. Kinoshita and H. Nakamura. Protein informatics towards function identification. *Current Opinion in Structural Biology*, 13(3):396–400, June 2003.

[142] B. Li, S. Turuvekere, M. Agrawal, D. La, K. Ramani, and D. Kihara. Characterization of local geometry of protein surfaces with the visibility criterion. *Proteins*, 71(2):670–683, May 2008.

[143] R. A. Laskowski, N. M. Luscombe, M. B. Swindells, and J. M. Thornton. Protein clefts in molecular recognition and function. *Protein Science*, 5(12):2438–2452, December 1996.

[144] J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Science*, 7(9):1884–1897, September 1998.

[145] R. A. Laskowski. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics*, 13(5):323–330, 307–308, October 1995.

[146] D. G. Levitt and L. J. Banaszak. POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of Molecular Graphics*, 10(4):229–234, December 1992.

[147] T. Kawabata and N. Go. Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins*, 68(2):516–529, August 2007.

[148] M. Weisel, E. Proschak, and G. Schneider. PocketPicker: Analysis of ligand binding-sites with shape descriptors. *Chemistry Central Journal*, 1(1):7, 2007.

[149] M. Hendlich, F. Rippmann, and G. Barnickel. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics & Modelling*, 15(6):359–363, 389, December 1997.

[150] Y. Kalidas and N. Chandra. PocketDepth: A new depth based algorithm for identification of ligand binding sites in proteins. *Journal of Structural Biology*, 161(1):31–42, 2008.

[151] M. Ota, K. Kinoshita, and K. Nishikawa. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *Journal of Molecular Biology*, 327(5):1053–1064, April 2003.

[152] B. Huang and M. Schroeder. LIGSITEcsc: Predicting ligand binding sites using the connolly surface and degree of conservation. *BMC Structural Biology*, 6(1):19, 2006.

[153] Y. Tseng, J. Dundas, and J. Liang. Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. *Journal of Molecular Biology*, 2009.

[154] A. H. Elcock. Prediction of functionally important residues based solely on the computed energetics of protein structure. *Journal of Molecular Biology*, 312(4):885–896, September 2001.

[155] A. Laurie and R. M Jackson. Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, 21(9):1908–1916, May 2005.

[156] J. An, M. Totrov, and R. Abagyan. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular & Cellular Proteomics*, 4(6):752–761, June 2005.

[157] C. T. Porter, G. J. Bartlett, and J. M. Thornton. The catalytic site atlas: A resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research*, 32(Database issue):D129–133, 2004.

[158] F. Ferre, G. Ausiello, A. Zanzoni, and M. Helmer-Citterich. SURFACE: A database of protein surface regions for functional annotation. *Nucleic Acids Research*, 32(Database issue):D240–244, 2004.

[159] N. D. Gold and R. M. Jackson. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *Journal of Molecular Biology*, 355(5):1112–1124, February 2006.

[160] R. J. Morris, R. J. Najmanovich, K. Abdullah, and J. M. Thornton. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparison. *Bioinformatics*, 21(10):2347–2355, 2005.

[161] B. Hoffmann, M. Zaslavskiy, J. Vert, and V. Stoven. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: Application to ligand prediction. *BMC Bioinformatics*, 11(1):99, 2010.

[162] N. Nagano, C. A Orengo, and J. M Thornton. One fold with many functions: The evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *Journal of Molecular Biology*, 321(5):741–765, August 2002.

[163] G. Demange, D. Gale, and M. Sotomayor. Multi-Item auctions. *The Journal of Political Economy*, 94(4):863–872, August 1986.

[164] R. Chikhi, L. Sael, and D. Kihara. Real-time ligand binding pocket database search using local surface descriptors. *Proteins*, 78(9):2007–2028, July 2010.

[165] V. Z. Spassov, A. D. Karshikoff, and R. Ladenstein. Optimization of the electrostatic interactions in proteins of different functional and folding type. *Protein Science*, 3(9):1556–1569, 1994.

[166] J. D. Madura and J. A. McCammon. Brownian dynamics simulation of diffusional encounters between triose phosphate isomerase and d-glyceraldehyde phosphate. *The Journal of Physical Chemistry*, 93(21):7285–7287, October 1989.

[167] S. Raychaudhuri, F. Younas, P. A. Karplus, C. H. Faerman, and D. R. Ripoll. Backbone makes a significant contribution to the electrostatics of alpha/beta-barrel proteins. *Protein Science*, 6(9):1849–1857, 1997.

[168] A. Kahraman, R. J. Morris, R. A. Laskowski, A. D. Favia, and J. M. Thornton. On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. *Proteins*, October 2009.

[169] L. Sael and D. Kihara. Characterization and classification of local protein surfaces using self-organizing map. *International Journal of Knowledge Discovery in Bioinformatics*, 1:32–47, 2010

[170] A. Ultsch. Maps for the visualization of high-dimensional data spaces. In *WSOM '03: Proceedings of the 3rd International Workshop on Self-Organizing Maps*, pages 225–230, Kyushu, Japan, 2003.

[171] A. Ultsch and L. Herrmann. Architecture of emergent self-organizing maps to reduce projection errors. In *ESANN '05: Proceedings of European Symposium on Artificial Neural Networks*, pages 1–6, Dortmund, 2005.

[172] R. Rojas and J. Feldman. *Neural networks – A systematic introduction.* Springer-Verlag, 1996.

[173] A. Ultsch and F. Moerchen. ESOM-Maps: Tools for clustering, visualization, and classification with emergent SOM. Technical Report 46, Department of Mathematics and Computer Science, University of Marburg, Germany, 2005.

[174] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976 2007.

[175] D. E. Platt and B. Silverman. Registration, orientation, and similarity of molecular electrostatic potentials through multipole matching. *Journal of Computational Chemistry*, 17:358–366, 1996.

[176] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448, April 1988.

VITA

VITA

Sael Lee received her bachelor's degree in computer science from Korea University in Feburary 2005. She came to the Computer Science Department of Purdue University and joined the Bioinformatics Laboratory in August 2005. She was funded by graduate fellowship for the first year of her Ph.D. studies and was funded through a research assistantship for the rest of her studies. She has coauthored seven publications regarding protein surface analysis under the name Lee Sael. She also actively participated in mentoring other graduate students through the Computer Science Korean Graduate Student Association and the Korea University Alumni Association at Purdue.