

**Characterization and Analysis of Protein – Protein
Interfaces**

by

Nurcan Tunçbağ

**A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
the Degree of**

Master of Science

in

Computational Science and Engineering

Koç University

September, 2007

Koç University
Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Nurcan Tunçbağ

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

Assist. Prof. Özlem Keskin

Assoc. Prof. Attila Gürsoy

Assist. Prof. Alkan Kabakçioğlu

Date:

ABSTRACT

The diverse range of cellular functions is performed by the limited number of protein folds existing in nature. One may similarly expect that the number of protein-protein interface architectures would also be restricted. In this study, the recently derived dataset of protein-protein interfaces is analyzed and compared with older datasets to address questions like (i) how many different protein-protein interaction types are expected to exist in nature for necessary biological diversity; (ii) what fraction of interactions is already known toward elucidation of the organization of the cell; and (iii) whether the increase in the number of interface architectures and consequently the functional coverage and interaction map of the PDB are reaching a plateau. The results show that number of protein interfaces increases at a much faster rate compared to the number of folds and is not yet to level off. Functional coverage is also found to steadily increase. As an estimation of this study, the total number of different interfaces will be around 8000 and it will take almost 30 years to discover at the current rate of experiment. Also, despite the diversity of interface architectures, some are more favorable and frequently used, and of particular interest, those are the ones which are also preferred in single chains. Another significant result is that some species, especially eukaryotes, prefer intra chain domain-domain interactions; others, less complex organisms, prefer inter chain interaction. This adaptation may be the result of the crowded traffic in the eukaryotic cells.

This thesis presents the multidirectional analysis and applications of the protein – protein interfaces. We believe in that the dataset of protein – protein interfaces is a rich source for researchers dealing with protein – protein interactions, protein recognition mechanisms, drug design etc.

ÖZET

Canlı hücrelerinde meydana gelen çeşitli fonksiyonlar, doğada var olan sınırlı sayıdaki protein etkileşimlerinin sonucu olarak ortaya çıkar. Benzer olarak, protein-protein arayüzey yapılarının da sınırlı sayıda olması beklenilebilir. Çalışmada, (i) gerekli biyolojik çeşitlilik için kaç çeşit farklı protein – protein etkileşim tipi beklendiği, (ii) hücre içi organizasyonun aydınlatılması için bu etkileşimlerin ne kadar kısmının bilindiği ve (iii) arayüzey yapılarının sayısındaki artış ile bunun sonucu Protein Veri Bankası'nın fonksiyonel kapsamı ve etkileşim haritasının düzlüğe ulaşıp ulaşmadığı sorularını cevaplandırmak için en yeni elde edilmiş olan protein arayüzey veri kümesi incelenmiş, daha önceden elde edilmiş olan protein arayüzey veri kümeleriyle çok yönlü olarak karşılaştırılmıştır. Sonuçlar, arayüzey yapılarının sayısının protein yapılarının sayısıyla karşılaştırıldığında çok daha hızlı bir şekilde arttığını ve henüz bu artışın durmadığını göstermektedir. Bu çalışmanın bir tahmini olarak, toplamda birbirinden farklı 8000 arayüzey yapısının olacağı ve deneysel yöntemlerin şimdiki hızıyla bunların açıklığa kavuşturulmasının 30 yıl alacağı öngörülmüştür. Ayrıca, arayüzeylerin yapısal çeşitliliğine rağmen, bazı arayüzeylerin daha çok kullanıldığı ve buna ek olarak bu yapıların tek zincirli proteinler tarafından da tercih edildiği görülmüştür. Bir başka önemli sonuç da, bazı türlerin, özellikle ökaryotların, zincir içi etkileşimi, daha az gelişmiş organizmaların ise zincirler arası etkileşimi tercih ettiği görülmüştür. Bu durum, ökaryotik hücrelerin karmaşıklığının sonucu oluşmuş bir adaptasyon olarak nitelendirilmiştir.

Bu tez çalışması protein arayüzeylerinin çok yönlü analizini ve bunların uygulama amaçlı kullanılmasını içermektedir. Elde edilen protein arayüzey veri kümesinin, protein-protein etkileşimleri, proteinlerin birbirlerini tanıma mekanizmaları ve ilaç tasarımıyla ilgilenen araştırmacılar için zengin bir kaynak olacağına inanıyoruz.

ACKNOWLEDGEMENTS

I would like to thank **Assist. Prof. Özlem Keskin** (my advisor) and **Assoc. Prof. Attila Gürsoy** for the chance of being a member of their research group, supplying us a productive and comfortable research environment and their support during my work. Also, I would like to thank my thesis committee members for their critical reading and useful comments.

I would like to thank also **the Scientific and Technological Research Council of Turkey (TUBITAK)** for their financial support during my M.S. study.

I would like to thank my friends at Koç University; my former and recent office mates (**Güneş Gündem, Emre Güney, M. Cengiz Ulubaş, Ekin Tüzün, Gözde Kar**), my home mates **Özge Engin and Zeynep Akçay**, my friends in 110 (**Sefer Baday, Ashhan Arslan, again Özge Engin**) and Z20 (**Osman N. Yoğurçeu, Bahar Öndül**), and also all members of **Graduate Student Association** for their support, encouragement and good times during these 2 years.

And finally, I would like to thank **my family** for their continuous support and patience during every step of my education.

TABLE OF CONTENTS

List of Tables	x
List of Figures	xi
Nomenclature	xiii
Chapter 1: Introduction	1
Chapter 2: Literature Review	4
2.1. Types of Protein – Protein Interactions	4
2.1.1 Homo- & Hetero- Complexes.....	4
2.1.2. Non-obligate and Obligate Interactions	5
2.1.3. Transient & Permanent Interactions	5
2.2. Protein – Protein Interfaces.....	6
2.2.1. Interface Properties	6
2.2.2. Hot Spots in Protein – Protein Interfaces.....	8
2.3. Biological Relevancy of the Structures (Identification of the Crystal and Biological Structures).....	10
2.3.1. Protein Quaternary Structure (PQS)	11
2.3.2. Conservation Based Identification.....	11
2.3.3. Amino Acid Composition Based Approach	12
2.3.4. Conserved Domain Interaction Approach	12
2.3.5. NOXclass Algorithm	12

2.4. Protein – Protein Interfaces Datasets	13
2.4.1. PIBASE.....	15
2.4.2. InterPare.....	15
2.4.3. SCOWLP	16
2.4.4. 3DID	16
2.4.5. SCOPPI.....	16
2.5. Similarity between Binding and Folding.....	17
2.6. Domain Fusion and Related Works	18
Chapter 3: Extraction of Results: A New Non-Redundant Interface Dataset	20
3.1. Construction of the new non-redundant data set of protein – protein interfaces	20
3.1.1. Interface definition and extraction of interfaces from the Protein Data Bank..	20
3.1.2. Structural Comparison of Interfaces:	22
3.1.3. Protein – Protein Interfaces Dataset Construction Steps	23
3.1.4. Identification of the crystal interfaces.....	24
3.1.5. Interface Characterization.....	24
3.2. Generation of the Non-redundant Clusters of the Dataset.....	28
3.3 Functional Interaction Network of the PDB Constructed from the Data Set	30
3.4. PRISM Update.....	32
3.4.1 Template Dataset	32
3.4.2. Target Dataset.....	35
3.4.3. Prediction Algorithm	35
Chapter 4: Comparison of the Interface Dataset	36
4.1. Presentation of the Dataset with Numbers.....	36

4.2. Top 10 Clusters in the Dataset	41
4.3 Interface Characterization.....	45
4.5. Web Page of the Interface Dataset.....	47
4.4 Functional Coverage of the Dataset and Comparison with Previous Sets.....	51
Chapter 5: Classification of the Protein Interfaces	54
5.1. Non-redundant set of clusters, Type I, II and III clusters	54
5.1.1. Type I Clusters	54
5.1.2. Type II Clusters.....	55
5.1.3. Type III Clusters	57
5.2. Relationship between Interfaces and Folds.....	58
5.2. Domain Fusion Events in the New Extracted Data Set	64
5.2.1 Overall Domain Fusions in the Dataset	64
5.2.2 Case Studies in the Dataset	66
Appendix A: APPENDIX	75
A.1. Interface Extraction Methods.....	75
A.1.1. Atomic Distance Calculation Method.....	75
2.1.2. Accessible Surface Area Calculation Method	76
A.2. Geometric Hashing Algorithm.....	77
A.3. Webservers, Softwares, Tools, Databases	79
A.3.1. NACCESS.....	79
A.3.2. SURFNET	80
A.3.3. MULTIPROT	80
A.3.4. CLUSTALW	81

A.3.5. CytoScape (network visualization and analysis)	81
A.3.6. VMD (molecule visualization).....	82
A.3.7. Other.....	82
Bibliography	83
Vita	89

LIST OF TABLES

2.1	Comparison of the current protein interface data sets.	14
4.1	The parameters used during the clustering of the interfaces.....	37
4.2	Number of PDB structures, interfaces, interface clusters and protein families for each data set.	38
4.3	Details of the top ten representative interfaces in both 2002 and 2006 datasets	43
4.4	Functional Annotations of the top ten representatives of the datasets 2002 and 2006	44
4.5	Content of the current data set and old data sets.....	46
5.1	Comparison of the most populated folds with the representative interfaces of the most populated clusters.....	59

LIST OF FIGURES

3.1	Representation of the contacting and nearby residues schematically.....	21
3.2	Two chain protein interface definition.....	22
3.3	Schematic representation of the methodology to form the protein-protein interfaces data set.	27
3.4	Generation steps of the interface clusters types.	29
3.5	Preparation of the functional network of the interfaces.....	31
3.6	Flowchart to generate template dataset.....	34
4.1	(A) Comparison of three dataset from view of numbers of clusters, interfaces, PDB structures and SCOP families. (B) Presentation of the increase ratio in protein families, protein interface clusters and PDB entries. This figure shows the changes in proteins and interfaces in 12 year period.	39
4.2	Representation of the exponential increase in the interface number and protein number annually.....	40
4.3	Representation of the enlargement in the cluster sizes.....	41
4.4	Illustration of the representative interfaces of most populated 10 clusters by ribbon diagrams.	42
4.5	The snapshot of the main page of DAPPI. User can query with Interface ID.....	49
4.6	The snapshot of the main page of the individual interface page for 1lasAB.....	50
4.7	Functional interaction network of the proteins coming from the interfaces generated from PDB entries in 2006..	51
4.8	The function interaction network of the old and current data sets.....	53
5.1	Type I cluster examples.	55

5.2	Type II cluster examples.....	56
5.3	Type III cluster examples.....	58
5.4	Comparison of the 1o1pAD with 1enh (3 – helix bundle fold).	60
5.5	Comparison of the 1djrDE with 1ris ($\alpha - \beta$ plait :1ris).....	61
5.6	Two more examples for similarity between folding and binding.	63
5.7	Domain fusion map of the PDB derived from new generated interface dataset.....	65
5.8	Intersection of the general domain fusion map of the structures in PDB (as of February 2006) with the domain fusions available in literature.	66
5.9	Gene fusion examples in urease enzyme..	67
5.10	Schematic representation of the domain fusion.	68
5.11	Gene fusion example in imidazole glycerophosphate synthase enzyme..	69
5.12	Domain fusion events in cytochrome BC1 complex.	71
A.1	Overlap of three interface identification methods.....	77
A.2	Representation of the terminal residues.	79

NOMENCLATURE

<i>PDB</i>	Protein Data Bank
<i>ASA</i>	Accessible Surface Area
<i>PPI</i>	Protein – Protein Interaction
<i>RMSD</i>	Root Mean Square Deviation
<i>SCOP</i>	Structural Classification of Proteins
<i>PRISM</i>	Protein Interactions by Structural Matching

Chapter 1

INTRODUCTION

Most molecular and cellular processes are controlled by protein-protein interactions. Deciphering the mechanism of protein interactions is crucial for the comprehension of the large scale organization of the cells and their biological pathways. Recent studies have discovered thousands of novel protein interactions, changing our perception of the cellular organization. Comparative assessment of large-scale data sets of protein-protein interactions supported by more than one method, estimate that there are around 2400 interactions out of the estimated 80,000 physical, genetic or functional associations between yeast proteins [1].

Protein interactions change during development and in response to external stimuli. The interaction networks they form are dynamic, regulating and supporting each other. Antigen-antibody recognition, enzyme substrate binding, hormone receptor binding, RNA splicing, DNA replication, transcription, signaling pathways are some examples of the diverse and complex biological processes dominated by protein-protein interactions. Pioneering studies on the mechanism of protein-protein recognition provided insights into the properties of different types of protein-protein complexes and the principles of the interactions [2-4].

It was proposed that there are approximately 1000 types of protein folds in nature [5]. It is remarkable how, with such a relatively small number, nature can still perform the immense functional diversity. In a similar vein, it was stated that currently there are 700-800 known protein folds and despite an exponential increase in the number of structures, the increase in folds was leveling off which makes the 1000 fold estimate still hold [6].

Since, in order to carry out their functions, proteins generally associate with each other, the next obvious questions to ask are (i) how many different protein-protein interaction types exist in nature that will allow the diversity in biological processes; and, (ii) what fraction of the interactions is currently known in order to elucidate the organization of the cell. Aloy and Russell estimated that there should be 10,000 distinct structural types of protein-protein interactions. This number does not include antibody-antigen interactions, membrane proteins, protein-peptide complexes and ‘very special’ interactions. As a result, they estimated that in 2004, 1800 of the 10,000 types of interactions were already known and this number should increase at a rate of 250 new interactions per year [6].

Protein-protein interactions (PPI) take place through interfaces. If we consider each of the unique interface architectures as an interaction type, the number of distinct protein interactions should be related to the number of distinct protein interfaces. Thus, identification of distinct protein interfaces can provide information regarding how close we are to the limit in the number of interaction types.

In this thesis study, the nonredundant dataset of protein-protein interfaces and its analysis is presented. Here, also using our currently derived dataset of protein-protein interfaces and its comparison with the older datasets (extracted in 1994 [7] and 2002 [8]) we address the questions posed above. Further, evolution of the protein-protein interfaces is detected by interface cluster types, domain fusions and the similarity between binding and folding. This interface dataset is also used as template for updating the putative protein – protein interactions available in Protein Interactions by Structural Matching (PRISM) [9, 10] after elimination of the crystal structures.

The outline of this thesis study is as follows:

In Chapter 2, the corresponding work in the literature is demonstrated. This chapter includes the most recent interface databases, information about biological and crystal complexes and general aspects about protein interactions.

Chapter 3 is the methods part and illustrates the details of the generation steps of interface dataset. Atomic distance calculation method, clustering method, interface cluster type generation method and interface characterization method are explained in this chapter. In the continuing part of it, template and target dataset generation steps are shown for PRISM update.

Chapter 4 includes the presentation of the dataset with numbers and the comparison of the new generated interface dataset with previous datasets [7, 8] from all aspects, such as secondary structure comparisons, functional coverage comparison.

In Chapter 5, evolution of the interfaces is analyzed. The similarity between folding and binding is shown with some case studies. Also, general domain fusion map and domain fusion events in the dataset are illustrated with some examples.

This thesis ends with a chapter which includes discussion of the results, future directions and conclusion of the study.

Chapter 2

LITERATURE REVIEW

This chapter includes the summary of the detailed literature search about the study presented in this thesis. Here, other protein interface databases with their distinct properties, the more their comparisons are available. Also, works about protein interface properties and all related information are presented.

2.1. Types of Protein – Protein Interactions

Protein – protein interactions can be categorized according to their sequence identity, their lifetime in the cell, and also their stability. In the studies about protein-protein interactions, the characteristics of these types are examined. Functional divergence is also taken into account such as antibody-antigen, enzyme-inhibitor complexes.

2.1.1 Homo- & Hetero- Complexes

Protein-protein interactions can be classified as homo- and hetero- whether binding partners are identical or non-identical. The similarity based classification declares that two proteins can interact through i) identical chains (homo-complex), same surfaces ii) identical chains (homo-complex), different surfaces, and iii) non-identical chains (hetero-complex), different surfaces. In homo-complexes if two chains interact through same surfaces with each other, this complex named isologous homo-complex. Two proteins can also dimerize through different surfaces in a heterologous way [3].

2.1.2. Non-obligate and Obligate Interactions

In addition to the sequential similarities, protein interactions can also be classified according to their stability *in vivo*. Obligate protein interactions are not found stable on their own in the cell; they should interact with its interaction partner to be stable and they are also obligate to function. However, non-obligate complexes can be found stable on their own [3]. In obligate interactions the binding site is more conserved than non-obligate interactions [11]. Each of the obligate protomers is co-localized in the cell and they are generally hetero-complexes. In obligate interactions also the gap volume, gap between two partner chains, to binding site area ratio is smaller than the non-obligate interactions. Also Nooren and Thornton stated that obligate complexes are tightly packed and also binding site of the non-obligate complexes is more planar than obligate complexes [12].

2.1.3. Transient & Permanent Interactions

Transient and permanent interactions are distinguished according to the lifetime of the complex in the cell. Permanent interactions are stable structures and these are present in complex form. However, transient interactions dissociate and associate according the environment and other conditions *in vivo* [3]. Nooren and Thornton 2003 focused on these types of interactions and they stated that weak transient interactions have smaller interface size and they are more planar and polar. On the other side, strong transient interactions are more conserved, large in size and more hydrophobic. When transient and permanent interactions are mapped to obligate vs non-obligate type interaction, Nooren and Thornton stated that obligate interactions are always permanent, on the other hand, non-obligate interactions can be either transient or permanent [12].

2.2. Protein – Protein Interfaces

Protein – protein interfaces are crucial to elucidate binding principles of the proteins. The most important aim in interface analysis is to generate the set of properties which strictly distinguishes binding part from rest of the protein. Numerous studies for this aim have been addressed to detect general patterns of the protein binding process. In the continuing parts (in 2.1.1 and 2.1.2) details of studies about protein – protein interfaces analysis are explained.

2.2.1. Interface Properties

Knowing the binding region properties is critical to predict unknown protein-protein interactions; also information about the binding sites can help drug design. Some principles of protein-protein binding such as shape complementarity, hydrophobicity, hydrogen bonding, electrostatic interactions, residue propensities, conservation etc. have long been studied on different datasets to generate a significant pattern to define protein interaction sites. In the study of Jones and Thornton (1997) [13], 59 different protein-protein interfaces are considered which contain 4 groups; homo dimers, enzyme-inhibitor complexes, antibody complexes and hetero complexes. These complexes are characterized according to some essential properties using structural information of them. They detected six properties of these complexes; size and shape, complementarity, residue interface propensities, hydrophobicity, segmentation and secondary structure, conformational changes. They compared four groups according to their six properties and as a result they stated that homodimers prefer hydrophobicity and they are large in size. Hetero complexes are less hydrophobic than homo dimers. They bind this result the fact that homo complexes are generally permanent complexes. Also, homodimers, permanent hetero complexes and enzyme-inhibitor complexes are more complementarity than antigen-antibody complexes.

In spite of detecting significant differences, they did not find a strict pattern to identify protein-protein interaction in this study [13].

Larsen et al (1998) has also worked about the protein recognition mechanisms on a dataset of 136 homodimeric proteins. They stated that one-third of the interfaces have a distinguishable hydrophobic core which is large, the remaining interfaces have also small hydrophobic patches, polar contacts and water mediated interactions [14].

Amino acid propensity is also an important property. The interaction type of the proteins can be distinguished by the amino acid frequencies alone. In the study of Ofran and Rost (2003), six types of protein-protein interfaces are explored which are homo- vs. hetero- dimers, transient vs. permanent dimers and same domain vs. different domain interfaces. Using only amino acid composition and residue-contact preferences, they have reached 63-100 % accuracy to predict interaction types [15].

In another study, Valdar and Thornton (2001) analyzed six homodimers to detect conservation of the interfaces. They reached the result that interface residues are more conserved than the rest of the surface by a chance higher than random [11]. Caffrey et al. continued to explore the question whether interface regions of the proteins are more conserved than the remaining parts of their surfaces with a larger dataset than the previous works [16]. They considered 64 protein-protein interfaces and used their conservation scores and they showed that individual interface residues are more conserved than other surface residues. However, when they analyzed the surface patches and interface patches they explored that there is not a significant difference between them. Conservation is not enough alone to completely predict the protein binding sites, but it can be combined with other interface properties. In addition, they showed that buried interface residues are more conserved than the partial ones [16].

In the study of Chakrabarti and Janin (2002), the analysis of 70 protein complexes is presented. They considered small binding regions as single patch and large binding regions

as multi patches. Accessible surface area, aminoacid propensities, flatness and the hydrophobicity of the interfaces are analyzed. As a result, they stated that the interfaces and the interior part are similar in terms of the residue frequencies [17].

Bahadur et al (2003) split interfaces into two regions; core region and rim region. Core region is the buried part of the interface, and rim region contains solvent accessible residues. 122 homodimers are evaluated [18]. Patches in interfaces were identified by the same algorithm of Chakrabarti and Janin (2002) [17]. They stated that all interfaces have buried residues containing core region which is surrounded by a rim region. Core region of the interface is similar to interior of the protein in residue frequency. However, rim region is similar to protein surface. These permanent homodimers were compared with the protein-protein complexes. They stated that interfaces of the homodimers are larger and more hydrophobic than protein complexes. The hydrophobicity of the homodimers come from the large constitution of the core region of the interface. When the core regions of the homodimers and protein complexes are compared they are not significantly different from each other in terms of amino acid propensity [18].

Nooren and Thornton (2003) have evaluated transient protein interactions. In their data set they considered 16 weak transient, 23 strong transient homodimers. Transient complexes have interfaces which are similar to crystal contacts. They have characterized weak transient interfaces as flat, small and polar contact regions [12].

2.2.2. Hot Spots in Protein – Protein Interfaces

The binding energies on the protein interfaces are not distributed uniformly. Some key residues can contribute the large part of the binding free energy. These residues are called “hot spots”. They are critical residues in stability of the protein complexes and protein function. In the protein interfaces analysis hot spot residues are important. Numerous studies have dealt with the hot spot characterization and identification to gain more

knowledge about protein binding mechanisms. Experimental method to identify hot spot residues is Alanine Scanning Mutagenesis which is based on the fact that if a residue has a significant drop in binding affinity when mutated to alanine then it is a hot residue. Thorn and Bogan (2001) [19] deposited hot spots from alanine scanning mutagenesis experiments, in the ASEdb database. BID [20] is also a database of experimental hot spots which collects all available experimental data about hot spots in protein interfaces. However, they cover only a small portion when compared to available protein-protein interactions. In the lack of experimental information about the hot spot, researchers focused on computational methods to introduce several approaches to detect hot spots [21]. Some of them have developed energy based methods to predict hot spots [22-24]. Molecular dynamics studies have also been used to investigate the energetic contributions of interface residues [25-27]. As an alternative to energy based methods, conservation is an important property to detect hot spots. Structurally conserved residues and hot spots correlate with each other significantly [28-30]. These hot spots are also found to be buried and tightly packed with other residues [29] resulting in densely packed clusters of networked hot spots, called '*hot regions*'.

Another residue conservation based method is available in HotSprint. Only residue conservation is not sufficient to identify hot spots. Hot spots are buried and some amino acids are seen frequently as hot spot; such as ARG, TYR and THR [31]. By combining these statements, HotSprint uses these three properties (conservation, ASA, residue propensity) to detect hot spots. When checked with available experimental hotspots, HotSprint reaches an accuracy of 76%. HotSprint uses a scoring function to identify an interface residue as hot spot, called pScore.

In the light of these works, we can say that although there are not strict rules about the binding mechanism of the proteins, significant properties give clues about the protein – protein recognition. By the help of the properties like binding site size, residue frequency,

shape complementarity, conservation, hot spots, hydrophobicity etc., binding sites can be predicted with a high accuracy. Also, this leads to characterization of the binding surfaces of the proteins. These analyses about the binding sites and characterization of protein interfaces intensify our knowledge about the protein-protein recognition mechanisms and help to improve new methods for computational prediction and experimental identification of new protein-protein interactions and to design new drugs etc. If we combine the studies explained above, as a general view, protein interfaces are more conserved regions when compared to surface region and have critical residues on them named hot spot. Also, they are similar to interior region of the proteins from the view of physical and chemical properties.

2.3. Biological Relevancy of the Structures (Identification of the Crystal and Biological Structures)

Protein structure models in PDB are determined by experimental techniques such as Nuclear Magnetic Resonance (NMR), X-ray crystallography, etc. Among the diverse techniques for structure determination, X-ray crystallography is the most conventional method. However, X-ray crystallography does not produce always biologically relevant protein complexes. In other words, not all the complex structures in the PDB [32] are biologically relevant. Many of the contacts are formed as a result of crystallization process. These crystal packing interactions may cause noise in analyses. For this reason, the biologically meaningless contacts should be distinguished. A number of studies addressed the problem of distinguishing between biological and crystal packing contacts. In the continuing part, methods to determine the crystal interfaces are presented.

2.3.1. Protein Quaternary Structure (PQS)

PQS is the protein quaternary structure file server. In PQS, only the X-ray crystal entries of PDB are considered to differentiate crystal and biological complexes. Henrick and Thornton defined crystal packing interfaces by assigning a cutoff value (400 \AA^2) in the buried surface area [33]. This approach is based on the assumption that crystal structures have smaller interfaces in size when compared with biological interfaces. In PQS, if an interface has an interface size smaller than 400 \AA^2 , it is identified as a crystal contact [33]. Other studies, based on the interface size, used different cutoff values to distinguish crystal contacts. The most important feature to differentiate crystal packing contact is the interface size. The size of the crystal contacts are much smaller and less hydrophobic than biological interfaces [34].

However, interface size is not the only criterion to define an interface as either crystal or biological. There are opposite cases which have large interface sizes, but are biologically irrelevant; such as crystal contact in porcine adenylate kinase having an ASA of 2600 \AA^2 , pancreatic ribonuclease crystals with 1800 \AA^2 interface size [35].

2.3.2. Conservation Based Identification

Based on the assumption that the biological interfaces are more conserved than non-biological interfaces, Valdar and Thornton (2001) suggested that interfaces can be distinguished by residue conservation. They combined both the size and conservation information of the binding sites, and they achieved an accuracy of 98.3% on their training set. As a result, they conclude that biologically relevant binding sites are more conserved than the rest of the surface regions of that protein [36]. The difference between biological and crystal contacts is that biological interactions are more conserved and larger than crystal interactions in size [34, 37].

2.3.3. Amino Acid Composition Based Approach

The amino acid composition of the surface region of proteins differs from interface composition. Carugo et al. (1997) stated that if the binding site amino acid composition is similar to the rest of the surface of that protein, then this interaction is possibly crystal packing. They showed that there is no significant difference between amino acid compositions of the protein surface and the binding site in crystal packing interactions. In other words, according to this identification method, the binding region of the biological interactions are composed differently from the crystal packing [37].

2.3.4. Conserved Domain Interaction Approach

In this method, all interacting domain pairs are generated. By using structural alignment, these interacting domain pairs are clustered to obtain unique interface geometries. If two or more members in this resulting clusters have similar interface location, Shoemaker et al stated that these interfaces contain a conserved binding mode (CBM). In brief, from the large scale experimental data, the set of conserved domain – domain interaction model is generated in this method. According to these conserved modes they distinguished biological interactions. They checked the accuracy of their method on all globin interacting pairs and reached an accuracy of 90% without false positives [38].

2.3.5. NOXclass Algorithm

NOXclass is an algorithm to determine protein quaternary structures and to predict interaction types of known 3D structures of proteins. Zhu et al. [39] proposed that the amino acid and chemical composition of the interfaces are also useful in identification of crystal packing interfaces. In the NOXclass study, first they prepared the training set which contains three types of interactions (crystal, obligate, non-obligate) from various sources. In the training set, there were totally 106 crystal packing, 75 obligate and 62 non-obligate

interactions. To characterize protein interaction types, they used 6 different interface properties: Interface area, interface area ratio, amino acid composition, correlation between surface and interface regions, gap volume index, and conservation score of the interface. Using a combination of the properties in a support vector machine application, NOXclass distinguishes biological and non-biological interfaces reaching an accuracy of 91.8 % based on three parameters (interface area, interface area ratio and area based amino acid composition). As a result of the analysis, they found that the size of the interface is the most important feature to distinguish biological contacts. By using only interface area, they have reached an accuracy of 93% to separate biological and crystal interfaces. Also, biological interfaces were shown to be more conserved than the crystal packing interfaces [39].

2.4. Protein – Protein Interfaces Datasets

Protein interfaces have long been studied at both the protein level and the domain level. They have been represented as interface data sets and deposited into databases such as PiBASE [40], InterPare [41], SCOWLP [42], 3did [43], SCOPPI [44]. In Table 2.1, the comparison of three recent interface databases and the one used here is illustrated. The first row lists the databases. The next rows show the attributes of the databases, interface level (where the interfaces are extracted from, either chains or domains), and some key aspects of the databases. In the continuing part detailed information about these databases is supplied.

Table 2.1 Comparison of the current protein interface data sets.

	SCOWLP	InterPare	PIBASE	DAPPI
Reference	Teyra et al. 2006	Gong et al. 2005	Davis et al. 2005	Updated version of Keskin et al. 2004
Availability	http://www.scowlp.org/	http://www.interpare.net/	http://alto.compbio.ucsf.edu/pibase/	http://prism.cccb.ku.edu.tr/interface
Interface Level	Domain level interfaces	Domain level interfaces Inter-Intra Contact	Domain level interfaces Inter-Intra Contact	Chain Level
Interface Extraction Method	Atomic Distance	Atomic distance, ASA, Voronoi	Atomic Distance	Atomic Distance
Query Type	Search by PDB ID	PDB ID, sequence, keyword	Domain ID, PDB ID	Interface ID
Redundancy Removal	No redundancy removal	No redundancy removal	Crystal Structures (PQS)	Crystal Structures (NOXclass)
Clustering	No structural clustering	No structural clustering	Clustering according to complex, binding site and interface topology	Structural clustering by Geometric Hashing Algorithm
Interface Characterization	Peptidic interfaces, Water mediated interacting residues Interface area, residue characterization (hydrophilic, hydrophobic, wet spot)	Interface, surface, interior residues in pdb file format. Only visualization for all three methods Statistics of amino acid propensities for intra and inter interfaces.	Interface area (polar, nonpolar), cluster number, inter or intra domain interaction.	Homo-hetero dimer, interface area, gap volume index, conservation, hot spot data, residue propensities, cluster number, interface size, GO annotations, interaction type

2.4.1. PIBASE

PIBASE is a comprehensive database of structurally determined protein interfaces formed between domain pairs. The resource of the database comes from two types of data i) protein structures taken from PDB and PQS databases, ii) domain definitions obtained from SCOP and CATH domain classifications. Interatomic distances are calculated using a distance threshold (6.05 Å by default to allow also the water mediated contacts). Calculated distances are combined with the domain definitions. The generated domain interfaces also eliminated according to an ASA threshold (300 Å²) and duplicated domain-domain interactions. The dataset characterized by various geometric, physicochemical and topologic properties. Secondary structure topology of the interfaces is generated and the clustering procedure is performed according to the secondary structure fingerprints. PIBASE also provides the contact topology of the domains, polar vs nonpolar ASAs, etc. Users can query the database by PDB ID, SCOP classification or domain topology fingerprint [40]. PIBASE is available at <http://alto.compbio.ucsf.edu/pibase>

2.4.2. InterPare

InterPare is a protein domain interaction interface database, provides both inter (between chains) and intra (same chain) chain interfaces. As a methodology, InterPare uses three methods to generate domain domain interfaces: i) the atomic distance calculation method, ii) ASA, iii) the Voronoi Diagram, a computational geometry method. InterPare provides user the results of all three methods, also a visualization tool to distinguish the surface, interior and interface regions of the proteins. Each individual protein assigned with the amino acid properties of the surface, interior and interface parts. General statistics about the amino acid propensities of the inter interfaces and intra interfaces are illustrated comparatively. Also, the overlap of three interface identification methods is shown in the statistics parts. All three methods cover each other with high percentages. In InterPare,

there are options to query by PDB ID, sequence or keywords. The results of all three methods are presented on the web page [41]. InterPare is available at <http://interpare.net>

2.4.3. SCOWLP

SCOWLP is a database of detailed interface information by adding peptidic interfaces and solvent mediated contacts. The web-server allows visualization of the interfaces, also detailed structural analysis and comparison of the protein interfaces. The generated interfaces are domain level interfaces. To extract interfaces, atomic distance calculation method is used. By adding the peptidic ligand interfaces and solvent mediated contacts, the database of interfaces is enriched. Solvent mediated contact residues are named “wet spot”. User can query by PDB ID or SCOP domains [42]. Web server is available at <http://www.scowlp.org>

2.4.4. 3DID

3DID is a database of domain – domain interactions. Database contains GO based annotations. Users can query database by domain information, by sequences in FASTA format or by GO accession codes. On the output page, for a queried domain user can retrieve a network of interacting domains with the queried domain as source. Also, it lists the PDB structures containing this domain. For each domain – domain interaction, server calculates a score which identifies the interaction [43]. User can access the server at <http://gatealoy.pcb.ub.es/3did/>

2.4.5. SCOPPI

SCOPPI is a database of structurally classified protein interfaces. Here, to extract protein interfaces atomic distance calculation method is used and for domain classification SCOP is used. Besides the multiple sequence alignment, they also applied structural

alignment to the SCOP families. Binding region of a domain to another domain, named faces, are clustered according to sequence and structure. They concluded that one domain family has more than one face type. According to their work two interacting faces generates an interface type. They observed approximately 8400 interface types. SCOPPI can be queried by PDB ID, SCOP ID, GO annotation, and keywords [44].

2.5. Similarity between Binding and Folding

Folding and binding are similar processes conceptually. In the protein interfaces analysis, a similarity pattern between interfaces and interior regions of the proteins is noticed [11, 14, 16, 17]. In the study of Tsai and Nussinov (1997), they searched whether the main driving force, the hydrophobicity, in protein folding is also dominant in protein-protein binding process. They used 362 nonredundant interface structures to explore hydrophobic effect in binding. They stated that hydrophobicity is an important effect in protein associations; however it is not as strong as in protein folding [45]. The folding of one chain differs from interaction of two chains with the absence of the chain connectivity. In the folding process, intra chain recognitions are dominant. However, binding depends on recognition of inter chains. Tsai and Nussinov (1997) detected this similarity between folding and recognition. They considered two chain interfaces as single chain and explored the hydrophobic folding units. The hydrophobic units on the interfaces are also the driving force for binding like in single chain protein folding. They explained the binding process with analogy to the protein folding. In conclusion, they stated that compactness and hydrophobic effect have critical roles both in folding and binding and both of them stem from hydrophobicity [4, 45].

In one of the studies about the similarity between folding and binding, the high frequency vibrations (HFV) residues – critical residues determined according to Gaussian Network Model – of the conserved residues on the interfaces and in the core region of the

proteins are compared. As a result of the comparison of the interfaces and cores, they showed that core region and interface part show similar vibrations. In principle, similarity between folding nuclei and the hot spots on interfaces, derived by experimental and computational methods, is detected and both of them vibrate high frequencies. These results imply the validation of the statement that “folding and binding are similar processes” [46].

2.6. Domain Fusion and Related Works

Domains are described as the primary building blocks of proteins; combination of these functional groups results in completely different functions of proteins. Protein domains can be found in multiple protein structures, and they participate in intermolecular interactions. Interaction of them outcomes stable complexes which provide certain biological functions. As a result, protein – protein interaction analysis reduces to analysis of domain-domain interactions. In evolutionary constraints domains fuse from one chain to another in proteins. Gene or domain fusions are observed when two separate proteins in one organism appear as a single homologous fusion protein in another organism. Similarly, two domain pairs that interact may belong to a single chain in one species, whereas the same domain pairs belong to different proteins in another species [28]. Fused sequences are termed Rosetta stone proteins. For example, if A-B complex is a Rosetta stone protein that proposes the functional relation between A and B and describes the fact that their interaction has a better-than-random chance.

There are several studies focusing on the gene or domain fusion events in the organisms. In one of them, Kummerfeld et al considered proteins from 131 genomes and noticed that 2869 groups of multi-domain proteins exist as monomer in some species; multimers in others. Also, they stated that domain fusions are 4 times common than inter-chain domain-domain interaction. Domain fusions are reliable resources for prediction of

protein-protein interactions. Chia and Kolatkar generated their own domain fusion dataset. According to these intra-chain domain-domain interactions they filtered their putative protein-protein interactions [47]. In another study, Hua et al generated a domain fusion map within 30 entirely sequenced genomes. They reached the result that Rosetta-stone proteins, proteins formed by fused domain pairs, contain at least one α/β fold. The domain fusion maps illustrate an evolutionary history which implies the evolution of some multiple domains from a series of domain fusion events [48]. Yanai et al delineated also gene fusion events and as a result of the analysis of 30 microbial genomes, they stated that fused genes are generally in the same functional category. In other words, functional association of the genes can be extracted from gene fusions. Being nonrandom events make domain fusions reliable sources for prediction of protein-protein interactions and functional associations [49].

Chapter 3

EXTRACTION OF RESULTS: A NEW NON-REDUNDANT INTERFACE DATASET

This chapter contains methods used in this study to generate a protein – protein interfaces dataset and characterization of it. First, an interface extraction method is presented. Then structural comparison and clustering methods are described. In the continuing part, interface properties are elucidated. And finally, PRISM update steps are described.

3.1. Construction of the new non-redundant data set of protein – protein interfaces

As implied in Chapter 2, interface datasets are crucial to elucidate the protein recognition mechanisms. For this purpose, the non-redundant dataset of protein interfaces is generated and each individual interface is characterized with its physical and chemical properties.

3.1.1. Interface definition and extraction of interfaces from the Protein Data Bank

An interface can be defined as the set of amino acids which represents a region that links two polypeptide chains in a protein structure by non-covalent interactions. Residues interacting with each other across the binding region form the interface between two chains. Interface residues are selected according to the closeness of two residues, one from each chain. We defined two types of residues in two-chain interfaces; interacting and nearby residues. If the distance between any two atoms belonging to two residues, one

from each chain, is less than the sum of their van der Waals radii plus a tolerance 0.5 Å, these two residues are defined as interacting. If the distance between a non-interacting residue and interacting residue in the same chain is smaller than 6 Å, the non-interacting residue is flagged as a nearby residue. Nearby residues are important for information relating to the architecture of the interface and are convenient in structural alignment of the interfaces.

Figure 3.1 shows C^α atoms of both nearby and contacting residues on the interface named 1axdAB. As stated above, nearby residues are important for the continuity in the interface architectures. Especially in clustering step, structural alignment of the interfaces is achievable by means of nearby residues.

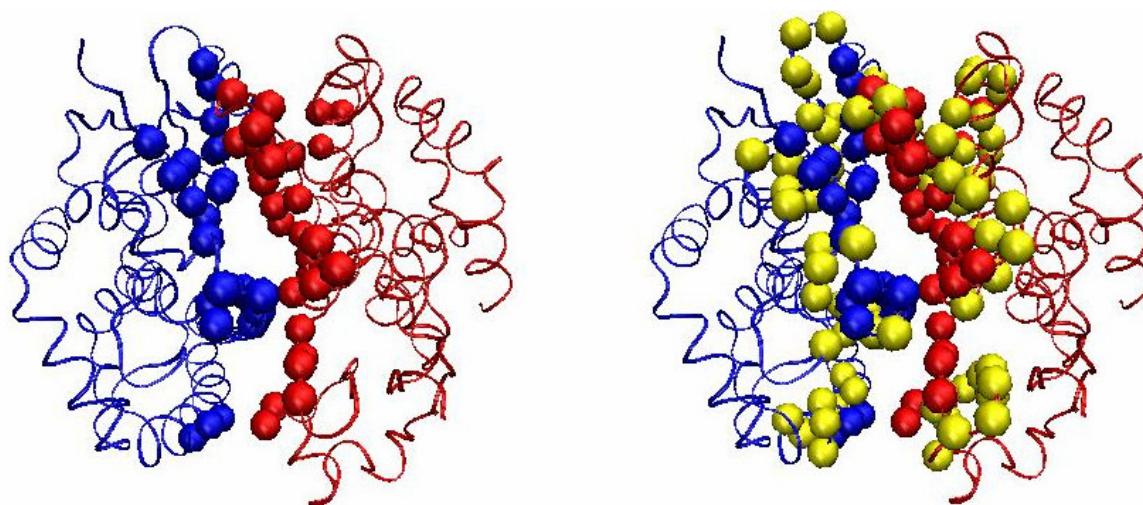


Figure 3.1 Representation of the contacting and nearby residues schematically. In the left figure, only contacting residues of each partner chain of 1axdAB are shown. In the right part, the both nearby and contacting residues are shown. Spheres are the C^α atoms on the interface (1axdAB). Yellow colored atoms are C^α atoms of the nearby atoms, blue atoms are C^α atoms of the contacting residues on chain A, and red atoms are C^α atoms of the contacting residues on chain B.

Interfaces are denoted as in the previous works [7, 8]. If the PDB name of a protein structure is 1fwi and there is an interface between the chains A and C, then this interface is named 1fwiAC. **Figure 3.2** presents an example of interfaces among three chains of a protein complex (pdb id: 1fwi; chain A colored yellow, chain B red, and chain C in cyan). Two interfaces are formed between chains A-C and B-C. Chains A and B are not close enough to form an interface.

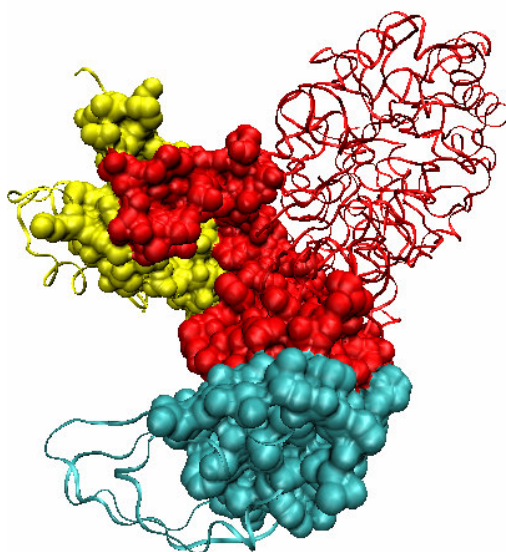


Figure 3.2 Two chain protein interface definition. Ribbon diagram and interface representation of the protein 1fwi is displayed. The chains A, B and C are colored yellow, red and cyan. Totally, two interfaces are formed between these three chains. First interface is between chains B and C. The second one is between chains A and C. 1fwi is shown as ribbon diagram, also the contacting and nearby residues of these two interfaces are shown as spheres.

3.1.2. Structural Comparison of Interfaces:

Since protein interfaces (i) have a discontinuous structure with two separate chains, (ii) may contain isolated residues and (iii) the residue order may differ between different interfaces with similar architectures, their comparison with others necessitates a sequence–

order-independent structural comparison technique. Here, the Geometric Hashing is used for alignment. Details of this algorithm are available in **Appendix A**.

Clustering of the data set is based on two measures, homogeneity and separation. Homogeneity implies that elements inside a cluster should be highly similar to each other; while separation implies that elements from different clusters should have low similarity. We have applied iterative clustering based on these principles. After each clustering cycle, the similarity definition is relaxed. In the first cycle, the first interface entry in the interface list forms a new cluster. If the similarity between the first and second entries is above the similarity threshold, the second interface entry is a member of the first cluster else it forms a new cluster. Next, the third interface is compared to existing clusters. The first cycle continues until all interfaces are assigned to a cluster. At the end of the cycle, the similarity between the cluster members and their representative should be above the current threshold.

3.1.3. Protein – Protein Interfaces Dataset Construction Steps

The current interface data set was generated following the flow chart in **Figure 3.3**. Initially, binary interfaces are extracted according to the atomic distance constraints between residues (detailed in section 3.1.1). Next, they are compared structurally using the Geometric Hashing algorithm and clustered. Detailed information about the structural clustering algorithm is available in the previous works and also in **Appendix A**. As a result, the structurally clustered interface dataset is generated. However, this dataset includes also the crystal structures in PDB. To be able to distinguish the crystal interactions, NOXclass algorithm is run for all interfaces in the dataset. NOXclass not only separates the interfaces as biological and crystal, but also it identifies the biological interfaces as obligate complex or non-obligate complex. This interaction type identification is obtained from the second step outputs of NOXclass. Besides the NOXclass outputs for the interaction type, the

interface properties are also generated: Gap volume of the interfaces (SURFNET), average conservation score of the interfaces (Rate4Site) and accessible surface area of the interfaces (NACCESS). At the end of this flowchart, each interface is registered along with its properties.

3.1.4. Identification of the Crystal Interfaces

After the generation of the interfaces with atomic distance calculation method, interfaces having less than 10 residues are eliminated according to the fact that crystal interfaces are small. A more rigorous identification of the crystal complexes is next carried out. NOXClass is used for this purpose. Detailed information about NOXclass is available in Chapter 2. The method uses six interface properties: interface area, ratio of interface area to protein surface area, amino acid composition of the interface, correlation between amino acid compositions of interface and protein surface, interface shape complementarity, and conservation of the interface, calculated by Naccess (for the first four parameters), Surfnet (for shape complementarity), Consurf (for conservation). NOXclass uses a two-stage Support Vector Machine (SVM).

3.1.5. Interface Characterization

In the dataset, all interfaces are characterized according to their properties. The ASA, gap volume, amino acid propensities, domain classification, conservation and computational hotspot information and interaction type as homo- or hetero- are extracted to be able to understand interactions more clearly.

i. Buried ASA Calculation

NACCESS is used to calculate the accessible surface area of the interfaces. Buried ASA of an interface gives also clues about crystal structures. Buried ASA is calculated as

the difference between complex ASA of two chains and sum of monomer ASA of these two partner chains. Probe size for calculations is default 1.4 Å. If we suppose that xxxxAB is an interface formed between chain A and chain B of the protein xxxx, then interface ASA is defined as: $\Delta ASA = (ASA_A + ASA_B) - ASA_{AB}$

ii. Gap Volume

Gap volumes of the interfaces are calculated by SURFNET with default parameters. Gap volume index gives the normalized value of gap volume with interface ASA and it is calculated as the ratio of gap volume to interface ASA. Gap volume indexes are important in determining obligate – nonobligate interactions. If gap volume index is small interaction type is probably obligate.

iii. Residue Propensity

To measure the individual residue frequencies, interface propensities are calculated. Propensities of 20 amino acids are calculated for the interface residues including nearby residues. The propensity of an amino acid on the interface is computed as follows:

$$P_i = \frac{(n_i / N_i)}{(n / N)}$$

where n_i is the number of residues i ($i = \text{Ala, Tyr, His ...}$) on the interface, N_i is the number of residues in whole partner chains. n is the total number of residues on the interface, N is the total number of residues on whole partner chains.

iv. Conservation and Computational Hotspots

Interfaces are more conserved than surfaces of the proteins. To observe the conserved residues, HotSprint database is used. Also, HotSprint gives computational hotspot information according to 3 different models (details are in Section 2.3). The computational

hotspots are defined as conserved residues which have a conservation score equal or greater than 7 in this thesis study. Besides the interface properties, conservation details are used also in NOXclass runs.

v. Domain Classification of the Interfaces

Whole chains of the interface producing proteins are mapped to their SCOP domains of version 1.71. This domain classification is used in the continuing part to explore domain fusions in the dataset and to generate interface cluster types (Type I, II and III).

vi. Homo-, Hetero- Interfaces

The FASTA sequences of two partner chains of the interfaces are compared with each other. If two interface producing partner chains are identical with each other (100% homologous) then this interfaces are identified as homo- complexes. Interfaces coming from non-identical chains are defined as hetero- complexes.

In addition to these properties, crystal-biological interaction, obligate-nonobligate interaction, size of the interface, cluster name of that interface and GO annotations are also identified in the presentation of the current dataset. Each individual interface is characterized in a detailed way and demonstrated in the web page of the dataset at <http://prism.cccb.ku.edu.tr/interface> for both the users who want to perform statistical analysis of the dataset and the users who want to deal with an individual protein – protein interface.

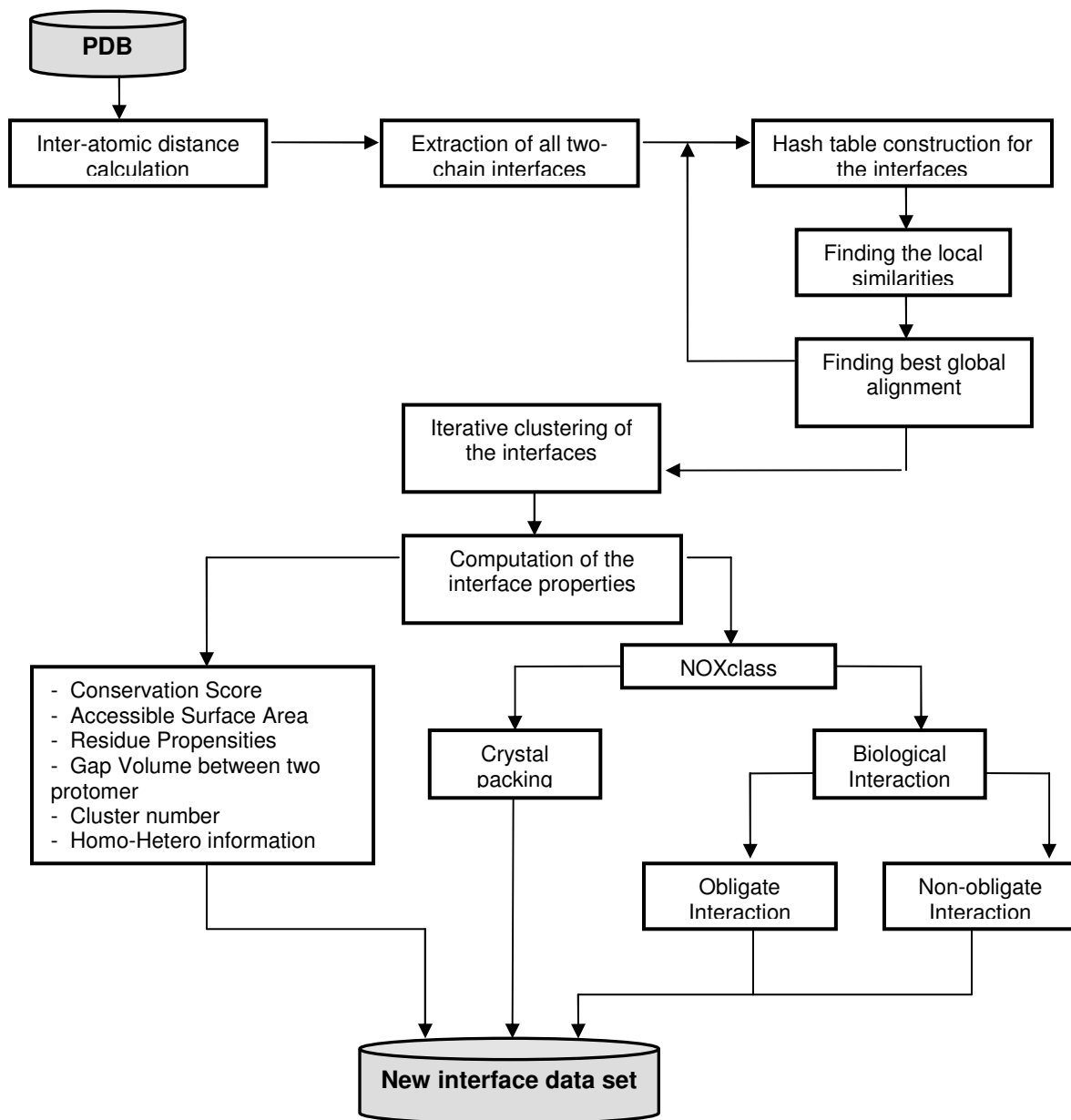


Figure 3.3 Schematic representation of the methodology to form the protein-protein interfaces data set.

3.2. Generation of the Non-redundant Clusters of the Dataset

To generate the nonredundant set of interfaces, members of each cluster are compared with each other sequentially. As shown in **Figure 3.4**, the elimination of the redundant interfaces is an iterative process. Members of each individual cluster are compared sequentially with each other starting with its representative interface.

CLUSTALW [50] with BLOSSUM matrix is used for this sequence comparison. The similarity threshold is set as 50%. During the comparison of two interfaces, if both partner chains of an interface are similar more than 50% to both partner chains of the compared interface, then the similar one is eliminated from the dataset. A threshold of 5 members is set to obtain relevant clusters, which implies at least 10 chains in a cluster. As a result of this filtering process, dataset of the interfaces become not only structurally but also sequentially nonredundant. The nonredundant dataset is further compared by Multiprot [51] to distinguish cluster types. This structurally and sequentially nonredundant set contains three types of clusters: i) Type I clusters; similar interface architectures coming from similar folds, ii) Type II clusters; similar interface architectures coming from dissimilar folds, iii) Type III clusters; one side similar architectures coming from dissimilar folds. Domains are classified at the super family level of SCOP version 1.71 [52].

Type I Clusters

Type I clusters are generated from the sequentially and structurally nonredundant set according to the fold of overall chain. Besides the structural similarity between members of the clusters, if partner chains are coming from the same super families these clusters are defined as Type I clusters.

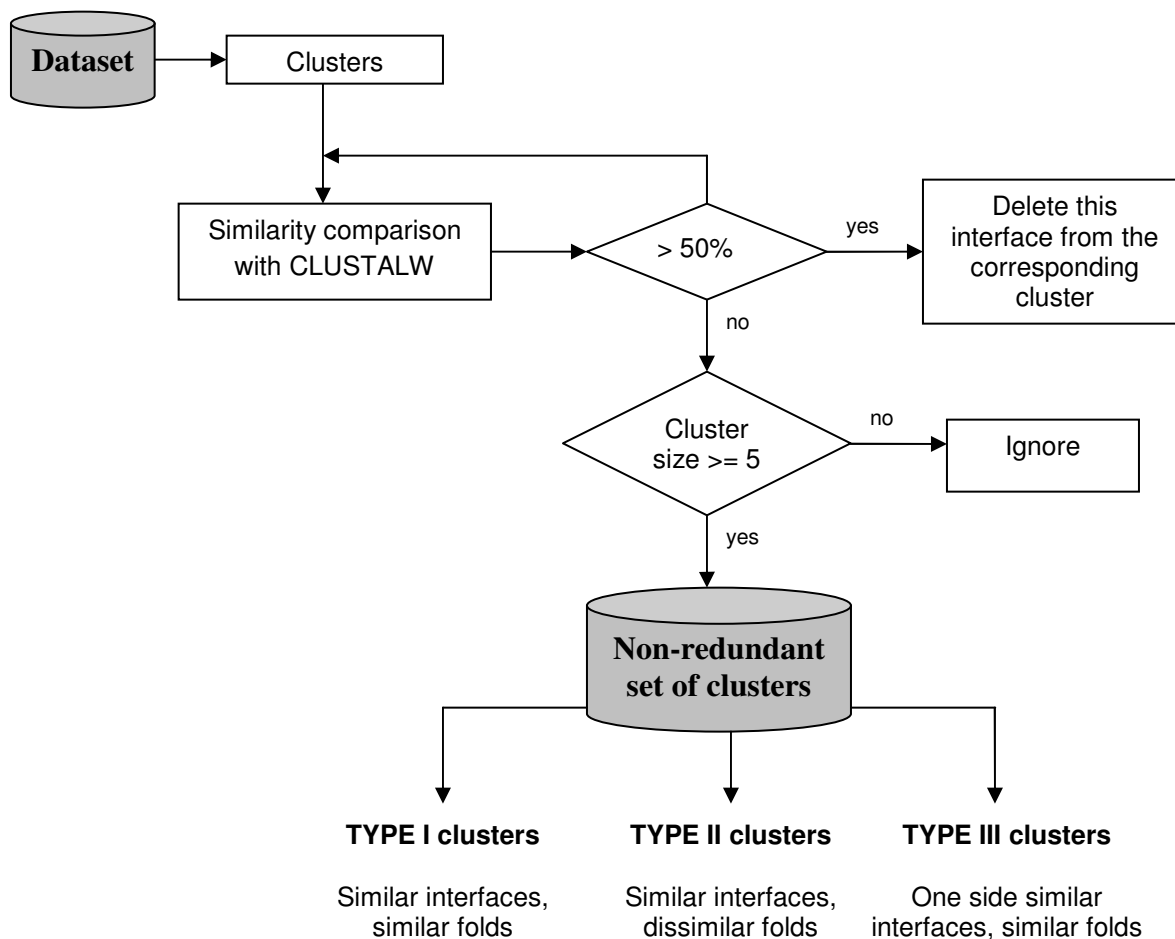


Figure 3.4 Generation steps of the interface clusters types.

Type II Clusters

Type II clusters contain structurally similar interfaces. However, global folds of their chains are not same. In other words, member interfaces are not generated from same domain – domain interactions. To obtain Type II clusters, the SCOP domain information of global folds are compared. If they are coming from different folds they are identified as Type II cluster. This type of clusters implies that differently folded proteins can interact through similar binding sites.

Type III Clusters

Type III clusters contain one chain conserved interfaces. In their interface structures, one side of the interface is always structurally conserved; the interacting partner of this conserved structure differs. The threshold to define this type is set as having an RMSD smaller than 3.5 Å between at least 10 matched residues. This type of clusters shows that one partner can interact with various partner chains.

3.3 Functional Interaction Network of the PDB Constructed from the Data Set

To analyze the functional coverage of the PDB, we generated the functional interaction network. The network was constructed by assigning GO molecular functions to proteins that constitute the interface representatives. From PDB, the file “gene_association.goa_pdb”, containing GO IDs of PDB chains, is downloaded. We start with the representative interfaces. Each interface chain is annotated by GO IDs extracted from “gene_association.goa_pdb” file. Then, each GO ID is mapped to the first level molecular functions described in GO annotations. First level molecular functions with their function IDs in parentheses are as follows: (# 1)Antioxidant activity, (#2) Binding, (#3) Catalytic activity, (#4) Chaperon regulator activity, (#5) Chemo attractant activity, (#6) Chemo repellent activity, (#7) Energy transducer activity, (#8) Enzyme regulator activity, (#9) Molecular function unknown, (#10) Motor activity (#11) Nutrient reservoir activity, (#12) Obsolete molecular function, (#13) Protein tag, (#14) Signal transducer activity, (#15) Structural molecule activity, (#16) Transcription regulator activity, (#17) Translation regulator activity, (#18) Transporter activity, (#19) Triplet codon-amino acid adaptor activity. If the function of a representative interface is unknown, the second member in the same cluster is used. The interface is identified as a functional representative interface and the function is used for that cluster.

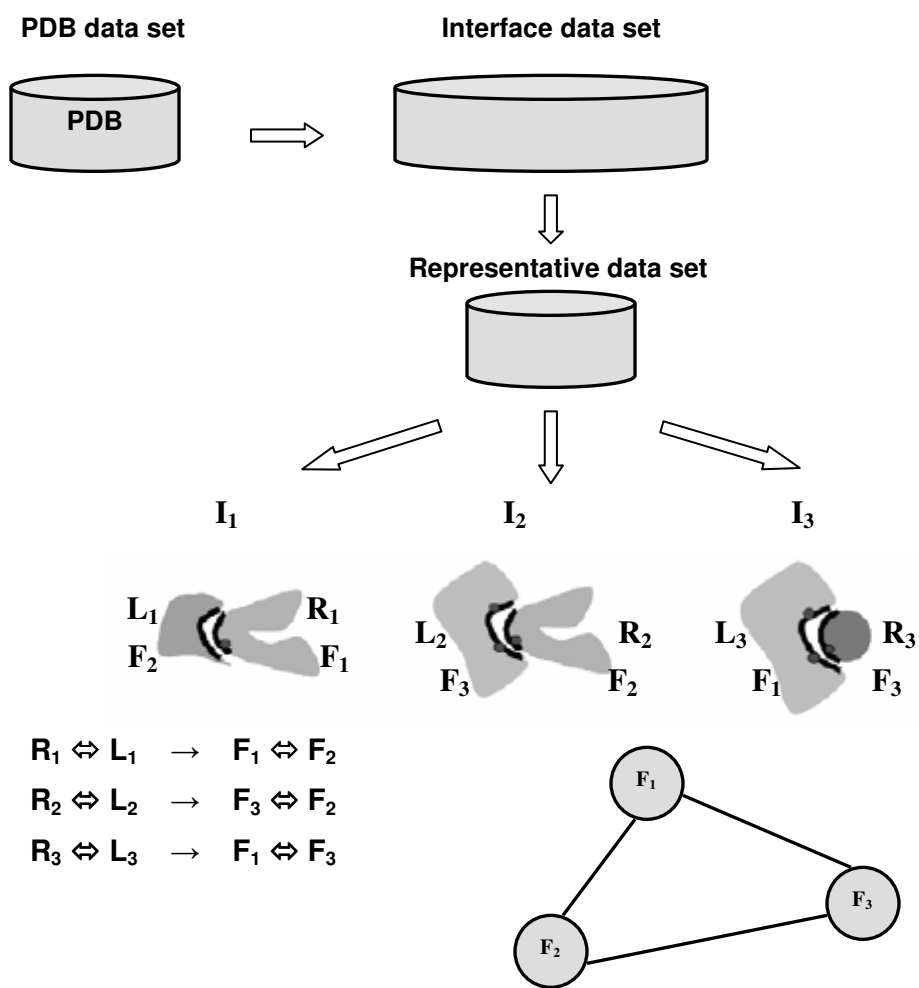


Figure 3.5 Preparation of the functional network of the interfaces. Each protomer coming from interfaces are annotated by GO annotations. As a result, the interaction network of these 19 functional classes is formed from representative interface data set. In the scheme, letter I represents interface, R represents the right partner of the interface; L represents the left partner of the interface. Also, F represents the functional class of that chain.

Figure 3.5 illustrates the procedure for the generation of the functional network. Here, if two chains interact with each other through an interface, the functions of the chains

should also be related. For example, supposing that the interface I_1 is composed of the partner chains L_1 and R_1 . The functional annotation of L_1 is F_1 and the functional annotation of L_2 is F_2 . Because these two chains interact with each other through the interface I_1 , we proposed that the functional classes F_1 and F_2 interact also with each other and an edge is put between these two functional classes. By applying this procedure to the whole dataset, the functional interaction network of the dataset is constructed. The more, besides the current dataset, the functional networks of the previous datasets are also constructed to detect the improvement in the functional coverage of the datasets and to compare them with each other.

3.4. PRISM Update

PRISM is a web server for the querying, visualization and analysis putative protein-protein interactions derived from known protein structures in PDB. Putative interactions between proteins are predicted with an efficient algorithm using structural and evolutionary similarities. The algorithm seeks possible binary interactions between proteins (targets) through similar known interfaces (templates). To improve the PPI predictions, the template and target datasets are updated. As a result, the putative interactions are also increased and improved. In the continuing parts, the new template and target datasets are

3.4.1 Template Dataset

Template dataset is a subset of protein-protein interfaces dataset. In the previous version of the PRISM, template dataset was constructed with a different procedure. First the similar chains containing interfaces are eliminated. Then, the structurally conserved hotspots are defined by using Multiprot. As a result, 67 template interfaces were generated. In the new version of PRISM, template dataset contains much more interfaces than the previous version. Template interfaces are extracted from overall representative interfaces

according to their biological relevancy. The flowchart for the template dataset generation is represented in **Figure 3.6**. If the representative interface is coming from crystal contact according to the NOXclass outputs, then the members of its corresponding cluster are analyzed whether biological or non-biological interaction. If the member is coming from biological contact then this interface is identified as template interface candidate. In the dataset, biological interfaces are defined according to the NOXclass outputs having a biological score of 80%. Interfaces coming from membrane proteins, synthetic proteins, peptides, DNA, RNA structures and antigen-antibody complexes are also eliminated. Elimination process continued with the hotspot number on the interfaces. A threshold of at least 3 hotspots on each chain of the interface is used. Hotspots are defined as conserved residues having a conservation score equal or larger than 7. Hotspot information is retrieved from HOTSPRINT database. This procedure generates a dataset of biologically relevant, diverse, evolutionarily and structurally nonredundant interfaces. Starting with all available 49,512 interfaces as of February 2006, 8205 distinct interface clusters are generated with their representative interfaces. After elimination of the antigen-antibody complexes, peptides, ligands, synthetic proteins, membrane proteins and interfaces having less than 3 hotspots – evolutionarily conserved residues on the interfaces – at each partner chain and consideration only biologically relevant interfaces, **1738** template interfaces are obtained. The new template dataset is more diverse than the previous version, increased in size from 67 templates to 1738 templates.

This diverse template dataset is used to perform similarity matching in PRISM. The scoring function of the PRISM algorithm considers both the evolutionary similarity (hotspot match ratio) and the structural similarity (residue match ratio, RMSD) to find similar binding regions on the surface of the target proteins.

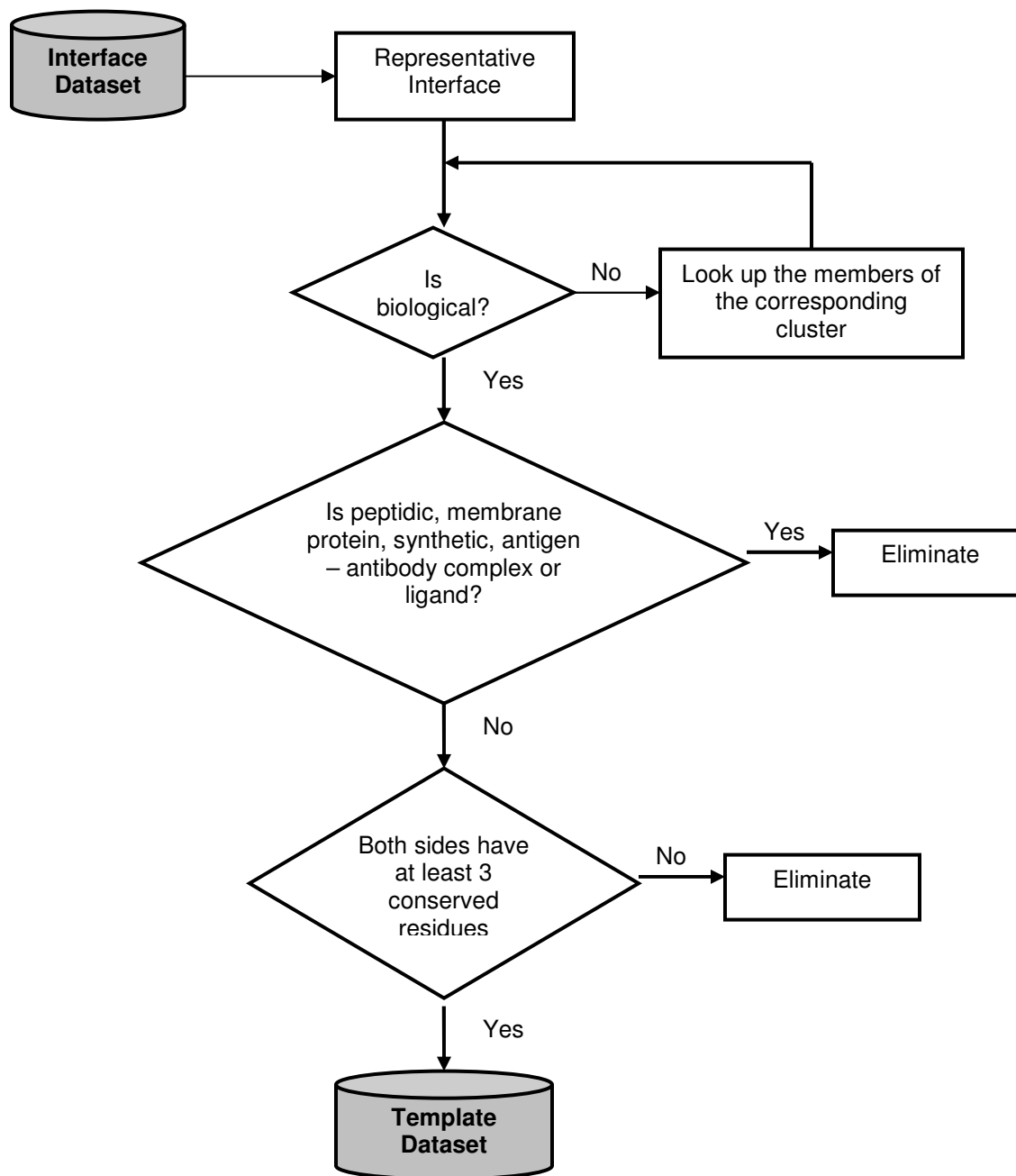


Figure 3.6 Flowchart to generate template dataset.

3.4.2. Target Dataset

Target dataset is the sequentially non-redundant subset of all structures available in PDB which have less than 50 % homology. Monomers in this dataset can be either single chains or polypeptide chains split up from complex structures. Besides the single chains, target dataset contains the complex structures which are also used to detect potential interactions. In the prediction algorithm, surface regions of the target proteins are used to find similarity with the constituent partners of the template interfaces. Surfaces of the target proteins are extracted by invoking NACCESS. If relative surface accessibility of a residue is greater than 5%, it is considered as surface residue.

Target dataset contains 16415 structures, of which 4952 are complex structures, 11463 are monomer structures.

3.4.3. Prediction Algorithm

The prediction algorithm based on that if two proteins contain similar regions to complementary partner of a template interface, it is proposed that these two proteins interact through these similar complementary regions. After the template interfaces are split into their complementary partner chains, these partners are structurally aligned with the surfaces of the target proteins. To measure the similarity, a scoring function is used, which contains two parts; i) evolutionary similarity score and ii) structural similarity score. Evolutionary similarity includes hotspot match ratio; structural similarity part includes RMSD and residue match ratio between target protein and one partner of template interface. Combination of the evolutionary and structural similarity scores with appropriate parameters provides the overall prediction score.

Running the prediction algorithm on the template and target dataset gives 58,817 putative interactions for a similarity score ≥ 0.85 , and 196,012 putative interactions for a similarity score ≥ 0.80 .

Chapter 4

COMPARISON OF THE INTERFACE DATASET

4.1. Presentation of the Dataset with Numbers

All multi (two or more) chain entries have been extracted from the PDB. On February 16, 2006, there were a total of 34817 structures. Atomic distance calculation method (explained in section 3.1.1) is applied to these structures and protein – protein interfaces were generated. Interfaces with less than 10 interface residues were eliminated since they are assumed to be crystal interfaces. As a result, 49512 two-chain interfaces were obtained. Stricter filtering to eliminate crystal structures is further carried out by using the NOXclass algorithm.

By Geometric Hashing technique, all these 49512 interfaces were clustered structurally. **Table 4.1** provides the iterative clustering steps and threshold parameters used to calculate the similarities between interfaces in this study. The first column denotes the six consecutive clustering cycles, from A to F. The second column gives the number of interface clusters at the beginning and at the end of the iteration cycle. At the beginning of the first iteration cycle, the number of interface clusters was the total number of two-chain interfaces in the PDB (49512). Following this cycle, the number decreased to 35744. At each cycle, thresholds of the parameters are relaxed. At the second iteration cycle, number of clusters decreased from 35744 to 20921 clusters. At the end of the entire clustering process, 8205 interface clusters are obtained. Members of each cluster had at least 0.5 chain relative connectivity score, which is defined in detail in **Appendix A**, with no threshold on amino acid identity. The maximal size difference between interfaces was 50 residues.

The data set and the clustering and detailed characterization results are available at <http://prism.cccb.ku.edu.tr/interface>.

Table 4.1 The parameters used during the clustering of the interfaces.

Cycle	Number of Interfaces	Relative Connectivity Score	Minimal % amino acid identity	Maximal amino acid size difference between interfaces
A	49512 → 35744	0.9	90	0
B	35744 → 20921	0.9	80	3
C	20921 → 14132	0.8	50	10
D	14132 → 11297	0.7	25	20
E	11297 → 9533	0.6	10	40
F	9533 → 8205	0.5	0	50

Each of the 8205 cluster is represented by an interface, called “representative interface”. Representative interfaces are the best interfaces signifying their clusters structurally. In the continuing part the current dataset (49512 protein-protein interfaces clustered into 8205 families) presented here and the old data sets extracted in 1994, 2002 are compared. **Table 4.2** shows the change in the number of PDB structures, interfaces, clusters and SCOP families. The number of interface clusters increased from 351 (in 1994) to 3799 (in 2002) to 8205 in 2006. The substantial increase in interfaces and interface families makes the newly generated data set more diverse. The number of PDB entries increased from 2814 to 18687 to 34817. Apparently, the number of interface clusters has increased more rapidly in the last 12 years when compared to PDB entries. Also, the number of SCOP families increased significantly.

Table 4.2 Number of PDB structures, interfaces, interface clusters and protein families for each data set.

Data Set	Number of PDB structures	Two-chain interfaces	Number of interface clusters	SCOP protein families
1994	2814	1629	351	850 (in 1997)
2002	18687	21686	3799	1827 (in 2002)
2006 (current work)	34817	49512	8205	2845 (in 2005)

The generation of new PDB structures illustrates both (i) an increase in the number of interface clusters, i.e., newly discovered interface architectures; and (ii) an increase in the population of clusters, i.e., more interfaces with a given architecture. **Figure 4.1(A)** presents the increases in the number of SCOP families, interfaces, interface clusters and PDB structures over the years. Clearly, there is a rapid increase in the number of interfaces and interface clusters when compared with the increase in the number of SCOP families and PDB structures. Thus, the currently available clusters suggest a broad diversity. In the 12 year period from 1994 to 2006 as shown in **Figure 4.1(B)**, while there was a 13 fold increase in the number of PDB entries, the number of interface clusters increased ~25 folds. Considering the number of PDB entries and the number of interface clusters of the three data sets, we observe an exponential increase; however, the increase in the number of interface clusters is faster than the number of PDB structures. This statement is supported by the representation in **Figure 4.2**. The increase in the number of interfaces and number of PDB structures is plotted annually and the exponential increase is illustrated. This rapid increase in the number of interface families may also be the result of the rapid growth in the determination of the large multi-chain complexes, likely to contain more than one interface. Also, increase in the number of interface families may be a result of the newly determined multi-chain complexes generates new interface architectures different from the previous 3799 families (2002 dataset).

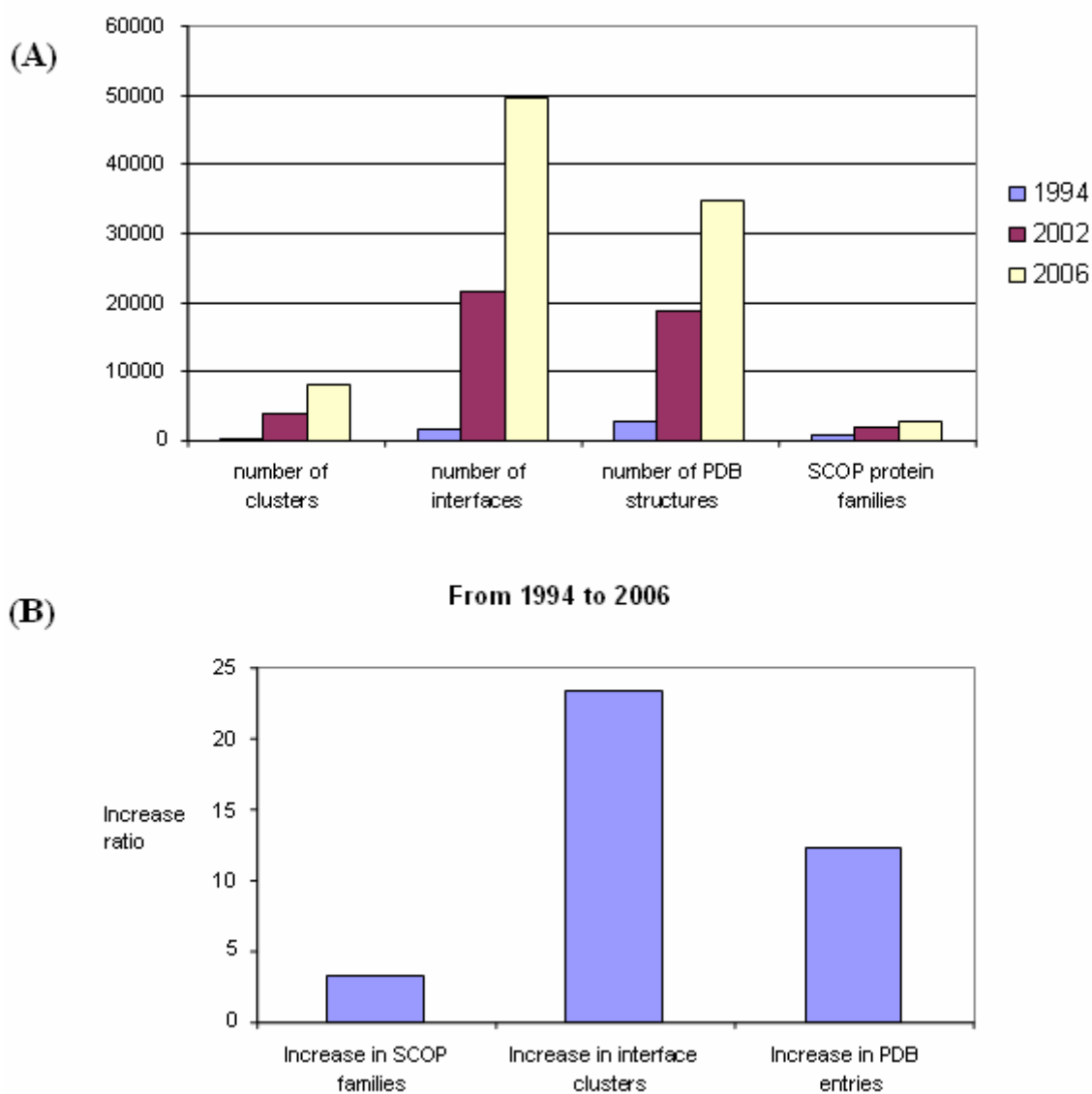


Figure 4.1 (A) Comparison of three dataset from view of numbers of clusters, interfaces, PDB structures and SCOP families. (B) Presentation of the increase ratio in protein families, protein interface clusters and PDB entries. This figure shows the changes in proteins and interfaces in 12 year period.

When only the biological interfaces are considered according to the NOXclass [39] outputs, the cluster number decreases from 8205 to 2279 in 2006, from 3799 to 1190 in

2002 datasets. From 8205 to 3799, there is more than 2 fold increase. However, for only biological clusters, from 2279 to 1190, there is less than 2 fold increase. From these numbers, we can propose that the crystal structures also have an important role in the increase of the interface families. Interfaces coming from crystal structures overamplify the whole dataset.

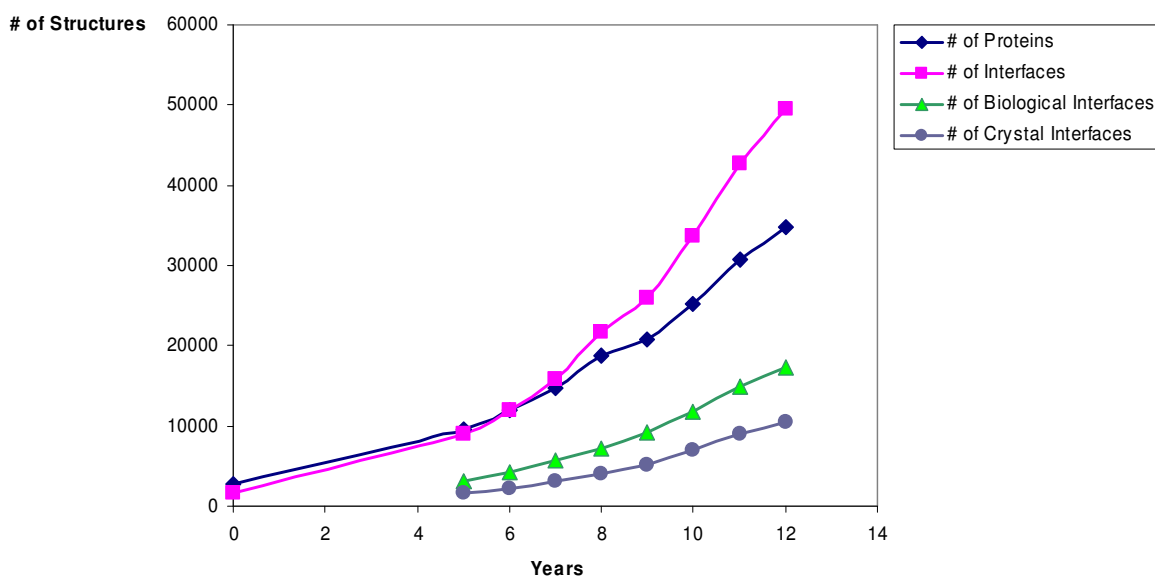


Figure 4.2 Representation of the exponential increase in the interface number and protein number annually.

In addition to the substantial increase in the number of clusters, there is also an increase in the cluster populations (**Figure 4.3**). Currently the cluster populations are larger than in 2002. The largest cluster (in 2002) had 281 members which now increased to 421 in 2006, suggesting frequent usage of the same favorable interface motifs. In 2002, 2946 clusters had less than 5 members; in 2006 we have 6424 clusters have less than 5 members. Further, in the 2006 set there are 6 clusters with more than 300 members. No such cluster was present in 2002.

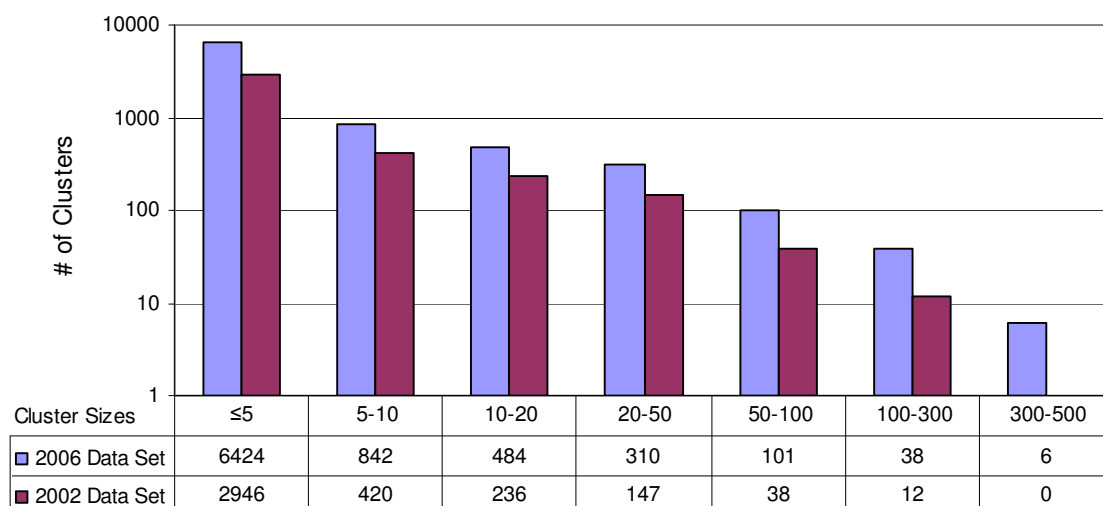


Figure 4.3 Representation of the enlargement in the cluster sizes. Our current data set is colored blue, and the 2002 data set is colored maroon.

4.2. Top 10 Clusters in the Dataset

The representative interfaces of the largest clusters are the most favorable interface architectures in the dataset. In the recent interface dataset, the most populated cluster is # 2171, and its representative is 1o1pAD. Here, we took top ten clusters in size. The ribbon diagrams of representatives of these 10 clusters are shown in **Figure 4.4** to be able to distinguish the secondary structures visually.

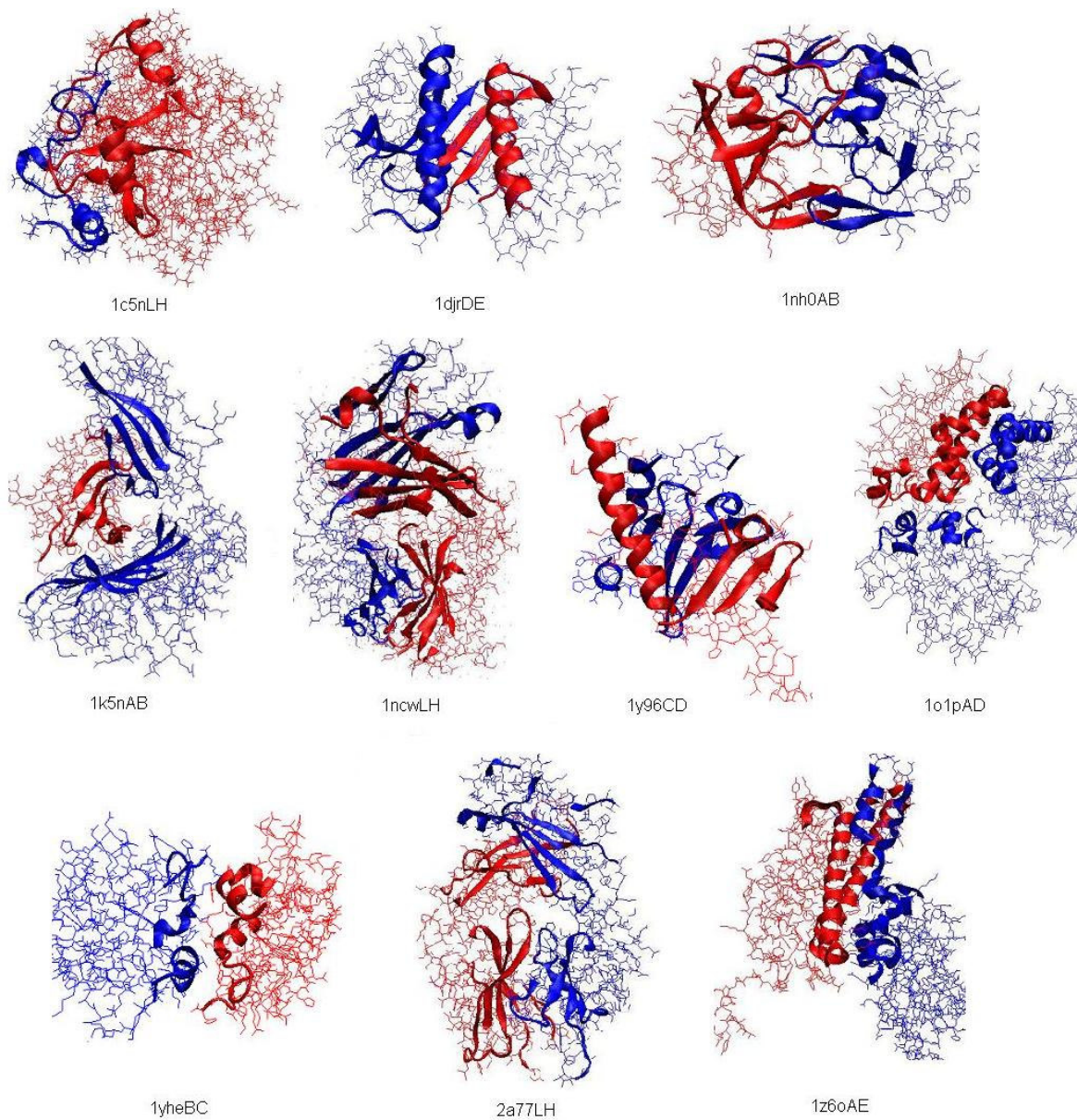


Figure 4.4 Illustration of the representative interfaces of most populated 10 clusters by ribbon diagrams.

The secondary structure information of the interface residues is extracted from DSSP database [53]. When these ten interfaces (twenty regions) in the current dataset are

investigated, we see that seven of them are solely helical; eight are composed of both helices and strands whereas five are formed by strands. When we applied same procedure to 2002 dataset and studied the representatives of the most crowded ten clusters, we investigated that six of the binding regions are helical; eleven are composed of both helices and strands whereas three are formed by strands.

Table 4.3 Details of the top ten representative interfaces in both 2002 and 2006 datasets

2002 DATASET				2006 DATASET			
Interface ID	Cluster ID	Cluster Size	Secondary structure	Interface ID	Cluster ID	Cluster Size	Secondary structure
1djrDE	833	281	Both mixed (beta-helix)	1o1pAD	2171	412	Only helix
1ce1LH	62	243	Left partner only beta, right partner mixed	1djrDE	756	384	Both mixed
1jthBC	358	215	Only helix	1yheBC	6853	350	Only helix
1c7cAB	372	210	Only helix	1ncwLH	2418	334	Both mixed
1g7aCD	1621	189	Left partner only helix, right mixed	1k5nAB	2586	318	Only beta
1qpwBC	1897	174	Only helix	2a77LH	467	312	Left partner mixed, right only beta
1k1tAB	2506	168	Both mixed	1y96CD	5292	255	Both mixed
1c5nLH	551	156	Both mixed	1nh0AB	3488	235	Both mixed
1g7qAB	1622	138	Only beta	1z6oAE	5312	231	Only helix
1cl7LH	63	119	Only beta	1c5nLH	496	219	Both mixed

In **Table 4.3**, the details of top ten clusters representatives for both the dataset 2002 and 2006 are given. So combination of helices and strands are commonly used in the interfaces, and there is a slight preference for helices compared to strands.

Further, these top ten representatives of 2006 dataset are compared functionally with the 2002 dataset. Their functional annotations are tabulated in **Table 4.4**. Comparison of the GO annotations of these ten interfaces in the current dataset and in the 2002 dataset says that a RNA binding protein, a ferritin and an antibody are found to be absent in the 2002 dataset but emerged in the current dataset. From 2002 to 2006, the development of the functions is also obvious.

Table 4.4 Functional Annotations of the top ten representatives of the datasets 2002 and 2006

2006 DATASET		2002 DATASET	
Interface ID	GO Annotation	Interface ID	GO Annotation
1djrDE	Pathogenesis (P) Extracellular region (C)	1djrDE	Pathogenesis (P) Extracellular region (C)
1o1pAD	Heme binding (F) Oxygen binding (F)	1ce1LH	Unknown
1yheBC	Unknown	1jthBC	Intracellular protein transport
1ncwLH	Unknown	1c7cAB	Oxygen transport, oxygen binding, heme binding
1k5nAB	Antigen processing, T cell receptor activity	1g7aCD	Hormone activity
2a77LH	Unknown	1qpwBC	Oxygen transport, oxygen binding, heme binding
1y96CD	RNA processing, regulation of transcription	1k1tAB	Proteolysis
1nh0AB	aspartic type endopeptidase activity, proteolysis, viral reproduction	1c5nLH	Trombin activity, proteolysis
1z6oAE	Iron ion homeostasis, transcription factor activity	1g7qAB	Antigen processing
1c5nLH	Trombin activity, proteolysis	1c17LH	Unknown

4.3 Interface Characterization

Protein interactions can occur between homo- or hetero-dimeric chains, and the interactions can be obligate or non-obligate, depending on whether the monomers are stable on their own. Identification of different types of interactions is crucial to understand the complete organization of the interaction map. To characterize the interfaces, we separate them into homo- and hetero-dimers. Biological versus crystal; and obligatory versus non-obligatory interfaces are distinguished using NOXclass. A NOXclass threshold of 80% is used to define a biological complex. **Table 4.5** summarizes the results for all data sets (current, 2002, and 1994): the first row gives the number of structures available in the PDB on the studied date; the second row provides the number of interface-generating multimers for the 2006 versus previous data sets. The numbers of interfaces, interface clusters and homo-, hetero- dimers are also presented. The number of characterized interfaces is smaller than the total, since some complexes do not have conservation files, gap volume indices; or the ASA of some large proteins can not be calculated by NACCESS, thus cannot be classified by NOXclass. In the redundancy removal step, the non-biological structures are eliminated and the number of interfaces and clusters are given. Antigen-antibody complexes, membrane proteins, synthetic and theoretical proteins are removed. Finally, peptide-protein complexes (chains under 100 residues with undefined domains in the SCOP) were eliminated. 19499 protein interfaces remained out of 49512 and the cluster number decreased from 8205 to 3086. When the same procedure is applied to the 2002 data set, only 1680 clusters remained. More than one third of the interactions are non-relevant biologically. Overall, as the table indicates, the number of new interface architectures (new interaction types) is growing at a rate of 350 per year. New interface architecture generation also continues to increase exponentially (**Figure 4.2**). If we consider these unique representative interface architectures as interaction types in the PDB as proposed by Aloy and Russell, our results are in agreement with theirs. They proposed that eventually

10000 interaction types will be observed. Our results suggest that some time is still needed to reach this limit.

Table 4.5 Content of the current data set and old data sets

	Current Data Set	2002 Data Set	1994 Data Set
Structures (PDB)	34817	18687	2814
PDB structures with interfaces	15268	7243	747
Number of Interfaces	49512	21686	1629
Number of Clusters	8205	3799	351
Number of homo-dimers	31990	13082	766
Number of hetero-dimers	17522	8018	746
Number of Interfaces characterized	27755	11749	1022
Non-Biological	10545	4172	340
Biological	17210	7577	672
Obligate	14501	6333	522
Non-Obligate	2709	1244	150
Redundancy Removal			
Interfaces	19499	9289	N/A
Number of Clusters	3086	1680	N/A
Only biologic interfaces			
Interfaces	12186	5401	N/A
Number of Clusters	2279	1190	N/A

Besides the interaction type identification and comparison with old datasets, each individual interface is also analyzed and all properties like ASA, Gap volume, domain classification, functional annotation, secondary structure information and evolutionary conservation are deposited on the web page.

4.5. Web Page of the Interface Dataset

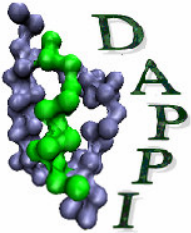
Our new generated non-redundant dataset of protein interfaces is accessible through the <http://prism.cccb.ku.edu.tr/interface>. The dataset is called **DAPPI** which is the abbreviation of “Dataset of Protein Protein Interfaces”. Website is designed for user to download whole dataset and also to access detailed information about each individual protein interface. The main page of the DAPPI is shown in **Figure 4.5**. Here, user can submit an interface ID and see the detailed information about that interface. Further, the clustering results geometric hashing algorithm at different levels and at final step, and the non-redundant clusters list with their types (type I, II, III) are downloadable on the main page. Also, at the bottom of the main page the general statistics about the DAPPI is available (not shown in **Figure 4.5**). DAPPI is an invaluable information source for detailed analysis of an interface. For each interface the interaction type information, interface size, cluster information, domain classification, secondary structure information, conservation scores and computational hotspot information of the interface residues (through a link to HotSprint database), GO annotations, NOXclass outputs, residue propensities are available. In the individual interface pages, users can download the interface residues in PDB format. Further, in the secondary structure files, users can find the secondary structure information residue by residue and their C^α coordinates.

For example, if user queries the interface named 11asAB, the resulting page will be as in the snapshot in **Figure 4.6**. Here, in the general information table, user can obtain the information that 11asAB is the member interface of the Cluster #5, its size is 178 residues, 11asAB is a homodimer – both the chains A and B are identical - , its accessible surface area is 3686 Å², its gap volume is 9056 Å³. Gap volume index – ratio of gap volume to ASA – is 2.46. In this table, also interface residues in PDB format and their secondary structures are available in the “11asAB.pdb” and “11asAB.secondary” files. When users


click on the link “11asAB@HotSprint”, they can access the conservation scores of the interface residues and also the computational hotspots calculated by 3 different methods.

The second table on this page is the SCOP classification table. Here, the domain information of 11asAB is available at all the class, fold, superfamily and family level. Both chains of 11asAB come from d.104.1.1 (Class II aaRS and biotin synthetases) which is an alpha and beta protein.

In the Noxclass table, user can learn that 11asAB is 99.92% biological. If it is biological its interaction is obligate with a percentage of 95.62. In the next table, the residue propensities of 11asAB are tabulated. Here, the high propensities of the amino acids TYR and LYS are noticed. The GO annotations all molecular function, process and localization are in the bottom of the page through a link to Gene Ontology.



DAPPI: DATaset of Protein Protein Interfaces


 KOÇ UNIVERSITY

Search Interface

Interface ID:

Interface ID: e.g. 104IAB

Interface definition and extraction of interfaces from the Protein Data Bank:

An interface can be defined as the set of amino acids which represents a region that links two polypeptide chains in a protein structure by non-covalent interactions. Residues interacting with each other across the binding region form the interface between two chains. Interface residues are selected according to the closeness of two residues, one from each chain. We defined two types of residues in two chain interfaces; interacting and nearby residues. Interacting residues are in direct contact with each other (each from one chain); nearby residues are in the vicinity of the interacting residues in the same chain (Tsai, 1996)

[Structural Clustering Method Information](#)

[Clustering Result at the final Step](#)

[Clustering Result at the different Steps](#)

[TYPE I Clusters](#)

[TYPE II Clusters](#)

[TYPE III Clusters](#)

[Fused Domain Pairs in the Dataset](#)

[HOME](#)

[Koc University, TURKEY](#)

[CCBB](#)

[COSBI](#)

[PRISM](#)

[HOTSPRINT](#)

Figure 4.5 The snapshot of the main page of DAPPI. User can query with Interface ID.

Interface ID: 11asAB

ASPARAGINE SYNTHETASE MUTANT C51A, C315A COMPLEXED WITH L-ASPARAGINE

[Help](#)

General Information	
Cluster ID	#5 (Member)
Interface Size	178 aa.
Interaction Type	homo dimer
ASA {A ^o }	3686.0
Gap Volume {A ^o }	9056.0
Gap Volume Index	2.46
Interface Residues in PDB format	11asAB.pdb
Secondary Structure	11asAB.secondary
Conservation and hot spot information	11asAB@HotSPRINT

NOXclass 6 parameters outputs	
Biological	99.92%
Non Biological	0.08%
Obligate	95.62%
Non-Obligate	4.38%

Details

SCOP Classification					
SCOP Domain	Chain	Family	SuperFamily	Fold	Class
d.104.1.1	A	Class II aminoacyl-tRNA synthetase (aaRS)-like, catalytic domain	Class II aaRS and biotin synthetases	Class II aaRS and biotin synthetases	Alpha and beta proteins (a+b)
d.104.1.1	B	Class II aminoacyl-tRNA synthetase (aaRS)-like, catalytic domain	Class II aaRS and biotin synthetases	Class II aaRS and biotin synthetases	Alpha and beta proteins (a+b)

Residue Propensities	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
	1.39	0.00	0.69	0.80	0.73	0.51	0.85	1.05	1.71	0.96	1.10	0.61	0.92	1.47	0.87	0.83	0.67	1.45	0.69	1.84

Chain	Process	Molecular Function	Localization
A	GO:0006418 , GO:0006529	GO:0004812 , GO:0004071 , GO:0005524	GO:0005737
B	GO:0006418 , GO:0006529	GO:0004812 , GO:0004071 , GO:0005524	GO:0005737

Figure 4.6 The snapshot of the main page of the individual interface page for 11asAB

4.4 Functional Coverage of the Dataset and Comparison with Previous Sets

In **Chapter 3**, the procedure to construct the functional interaction network is described in Figure 3.5. From 49512 interfaces, we have generated 8205 clusters with 8205 representatives. Each partner chain of the representative interfaces is annotated with GO first level functions. GO annotations was not available for some chains.

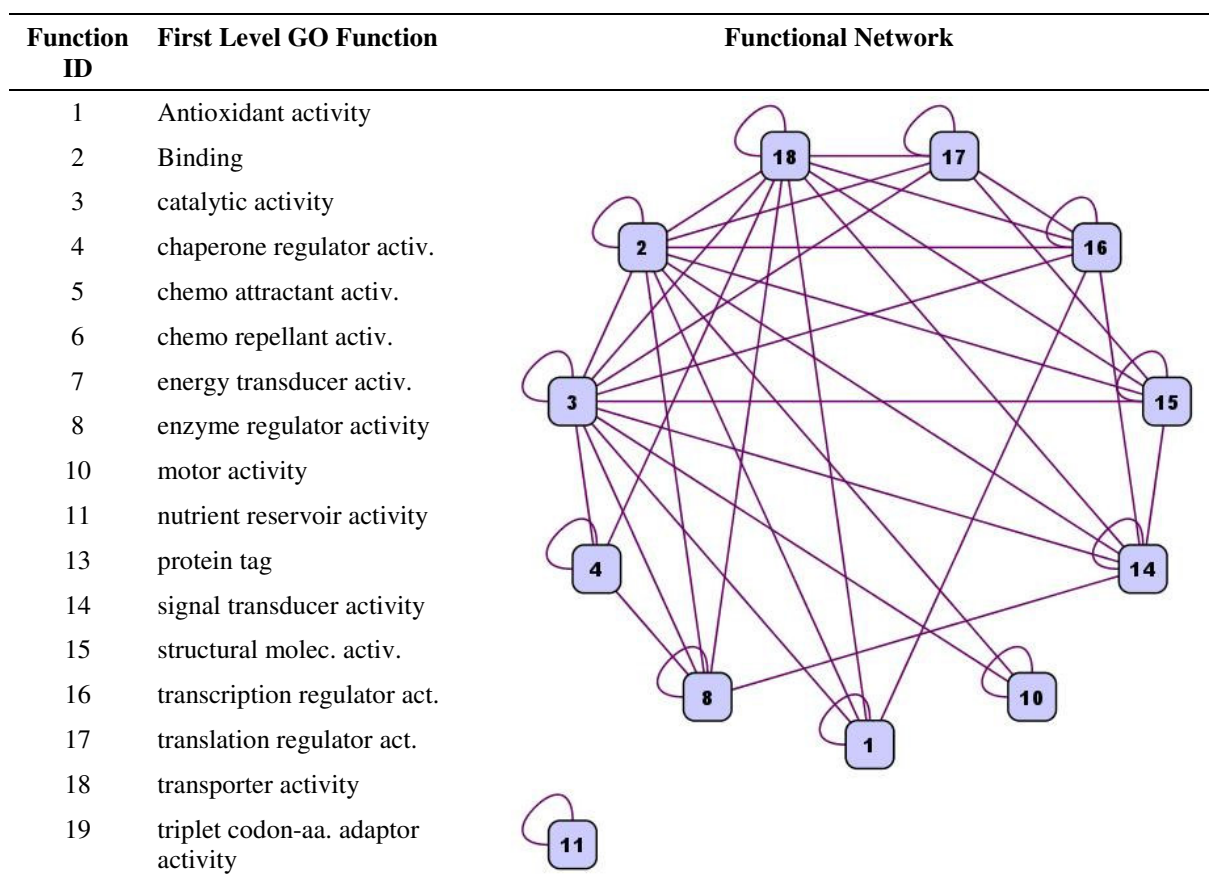


Figure 4.7 Functional interaction network of the proteins coming from the interfaces generated from PDB entries in 2006. Each node identified by a number and the interaction network of the functional classes is plotted. 12 of 17 functional classes are available in PDB.

After application of this procedure, we have constructed the interaction network of the first level functions in GO. This interaction network is illustrated in **Figure 4.7**. In the left part of the figure, the first level functions in GO and their function IDs are shown. These

function IDs are used to name the nodes in the network. The blue edges represent the interaction between two functional classes. There are 44 interactions between 12 functional classes.

We further examined the weight of the interactions: the frequency of a specific type of an interaction between two specified functions. It is observed that pair-wise functional interactions between the catalytic activity and binding; binding and signal transducer activity; binding and transcription regulator activity; binding and translation regulator activity; binding and transporter activity; catalytic and enzyme regulator activity; catalytic and signal transducer activity; and catalytic and transcription regulator activity are more frequently observed, i.e. they are highly connected. On the other hand, antioxidant activity is highly connected only with catalytic activity.

The same procedure has also been applied to the 1994 and 2002 datasets, and the three networks are compared (**Figure 4.8**). The chemo-attractant activity, chemo-repellant, energy transducer, and triplet codon-amino acid adapter activities and protein tagging are not present in none of data sets. On the other hand, as seen from **Figure 4.8**, the PDB-based functional network is getting closer to completion in the 12 years period. In 1994, ten function nodes interacted with each other with a total of 25 interactions. In 2002, the number of nodes (functional classes) and interaction numbers (edges in between functions) increased to 12 and 40, respectively. In 2006, the number of nodes was unchanged; however, the number of edges increased to 44. This indicates that the functional coverage has been constant although there are additional interactions between these functions. In the 2002 data set, the pair-wise interactions between the enzyme regulator and transporter activities; structural molecules and signal transducer activity; transcription regulator and antioxidant activity; and transporter and chaperone regulator activities were not yet formed.

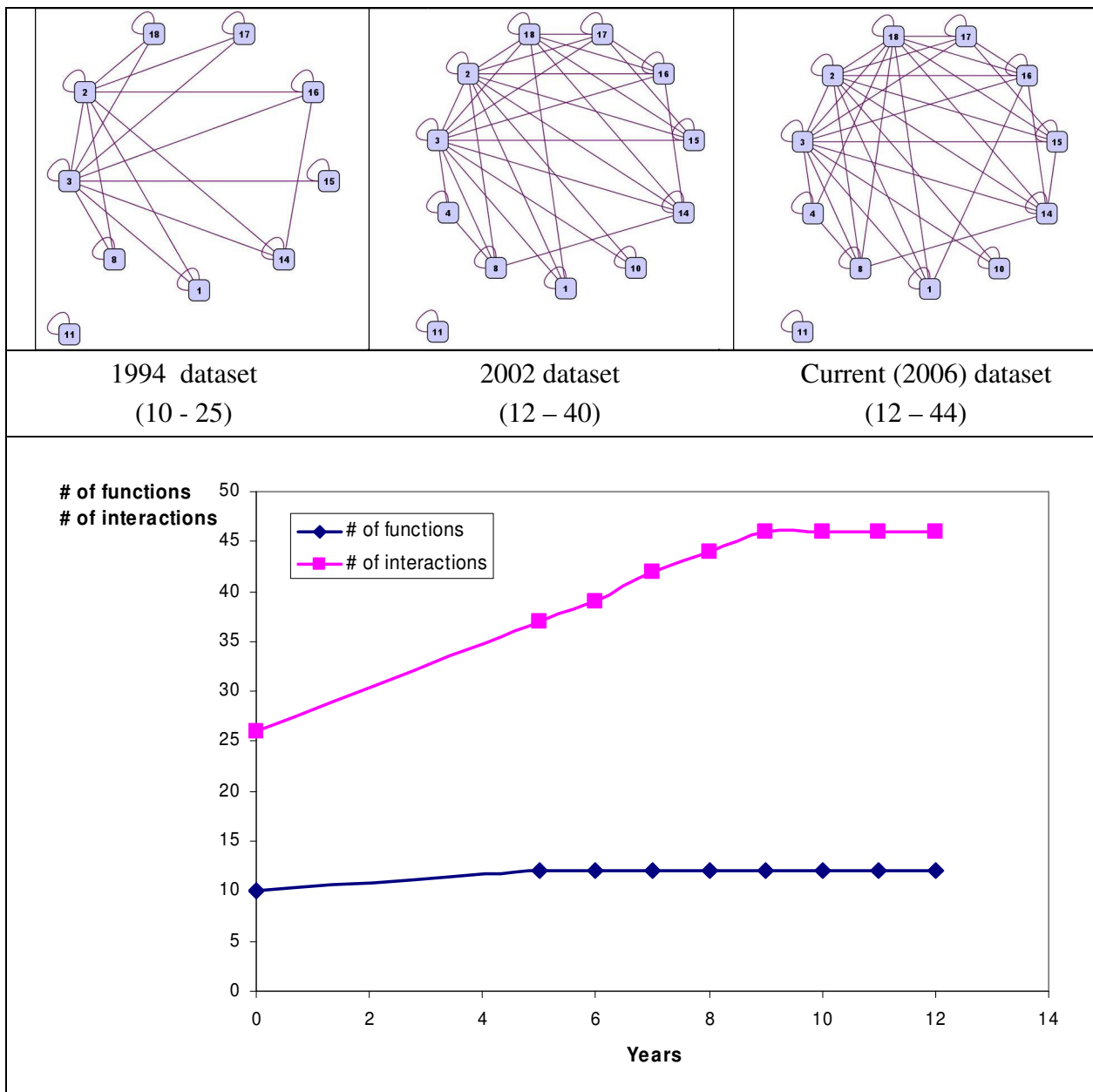


Figure 4.8 The function interaction network of the old and current data sets. The growing pattern of the networks is compared. The numbers in the parentheses represent the number of nodes and edges in the respective datasets.

As seen in **Figure 4.8**, in the functional interaction network of the 1994 data set, most of the functional interactions are not formed yet. With the new PDB structures and new interactions, the functional network has started to grow. The growth rates are shown in the lower panel of **Figure 4.8**.

If we assume that all functions will be covered and each function will have an average connectivity of 5, at the current rate of structural determination we shall still need around 30 years to have the completed network of protein interactions and hence functions. Aloy and Russell estimated that more than 20 years would be needed to obtain the complete representative set of protein interactions. Thus, our results, although based on datasets of protein-protein interfaces, suggest a similar (a little longer) time scale. This same interface analysis leads us to estimate that the total number of interface architectures (i.e. interface clusters) will reach 8,000, a value which is slightly less than what they suggest.

Chapter 5

CLASSIFICATION OF THE PROTEIN INTERFACES

This chapter contains interface cluster types, relationship between protein interaction and protein folding, and also evolution of the interfaces with domain fusion events. Here, these events are supported by some case studies available in the current interface dataset.

5.1. Non-redundant set of clusters, Type I, II and III clusters

As a result of the filtering process expressed in **Section 3.2**, the number of clusters reduced from 8205 to 213 which have at least 5 members and non-homologous to each other. 94 of the 213 clusters are Type I, 57 of them are Type II, and the rest (62 clusters) are Type III clusters.

5.1.1. Type I Clusters

In Type I clusters, the interface architectures are similar; also the overall domain folds of the chains coming from interface partners are same as mentioned in **Section 3.2**. One of the examples of this cluster type is cluster #300, represented by the interface 2aepHL, which has 6 members after elimination of the homologous sequences. In **Figure 5.1**, 4 of the 6 members are visually represented. In this figure, interface residues are colored according to chains types. The residues other than interface residues are colored green. As it is seen in **Figure 5.1**, the interface structures are similar to each other. When these 6 interfaces are aligned structurally by MultiProt, an RMSD value of 1.36 Å is found for 49

residue match between these 6 interfaces. These interfaces are formed by interaction of two immunoglobulin (b.1.1) folds.

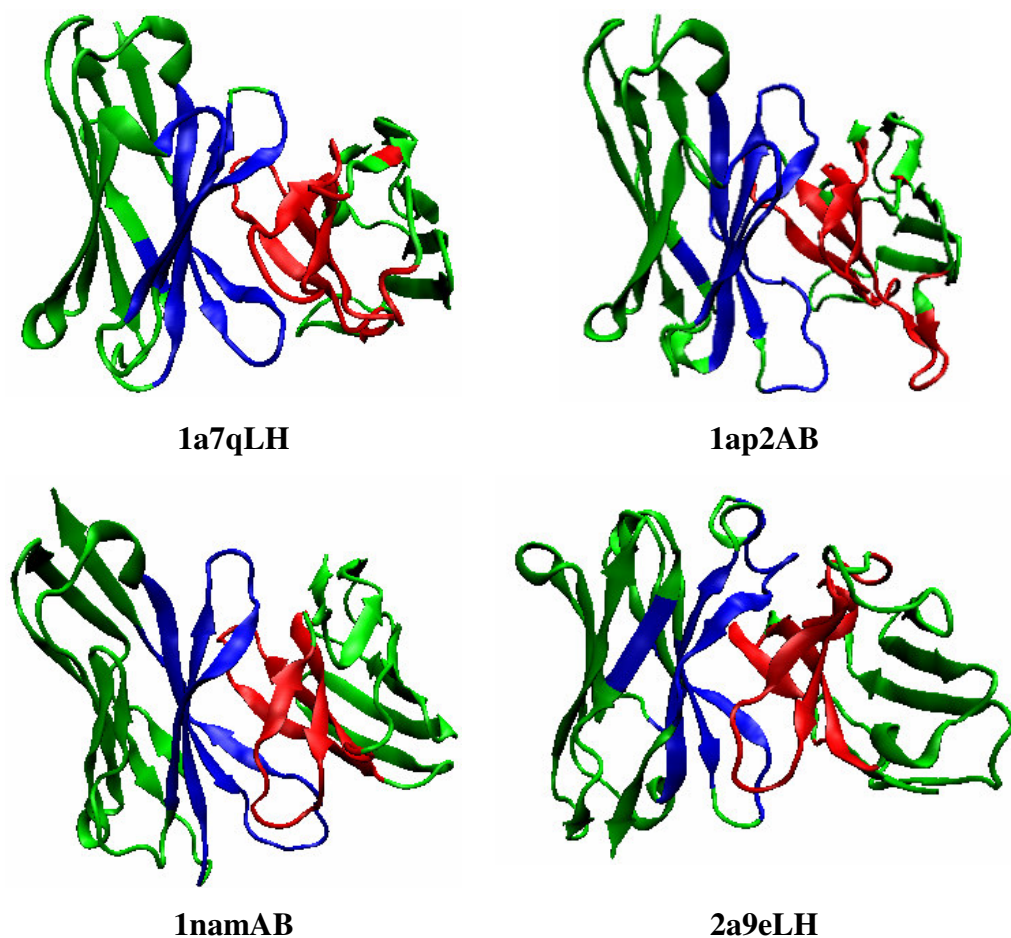


Figure 5.1 Type I cluster examples.

5.1.2. Type II Clusters

These types of clusters contain structurally similar interfaces like in Type I. However, the domain pairs generating the interfaces are different in the clusters. An example of this type is cluster #519, represented by interface 1lvfAB. The members of this cluster are

1lvfAB, 1cbiAB, 1dj8CE, 1ekeAB, 1unkAC, generated from the domain interactions between a.47.2.1 and a.47.2.1, b.60.1.2 and b.60.1.2, a.57.1.1 and a.57.1.1, c.55.3.1 and c.55.3.1. In **Figure 5.2**, these interfaces are shown schematically. In all of the interfaces in this cluster the interaction between two helices are similar.

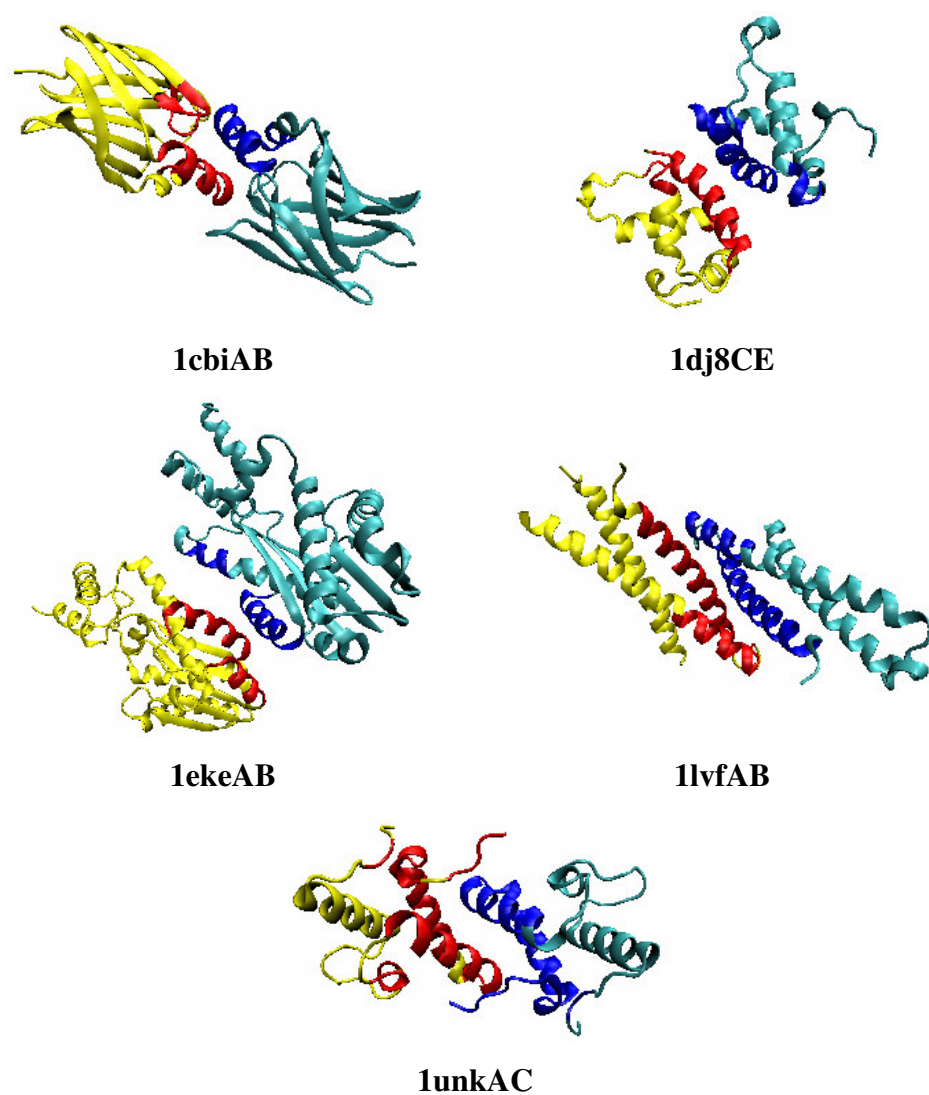
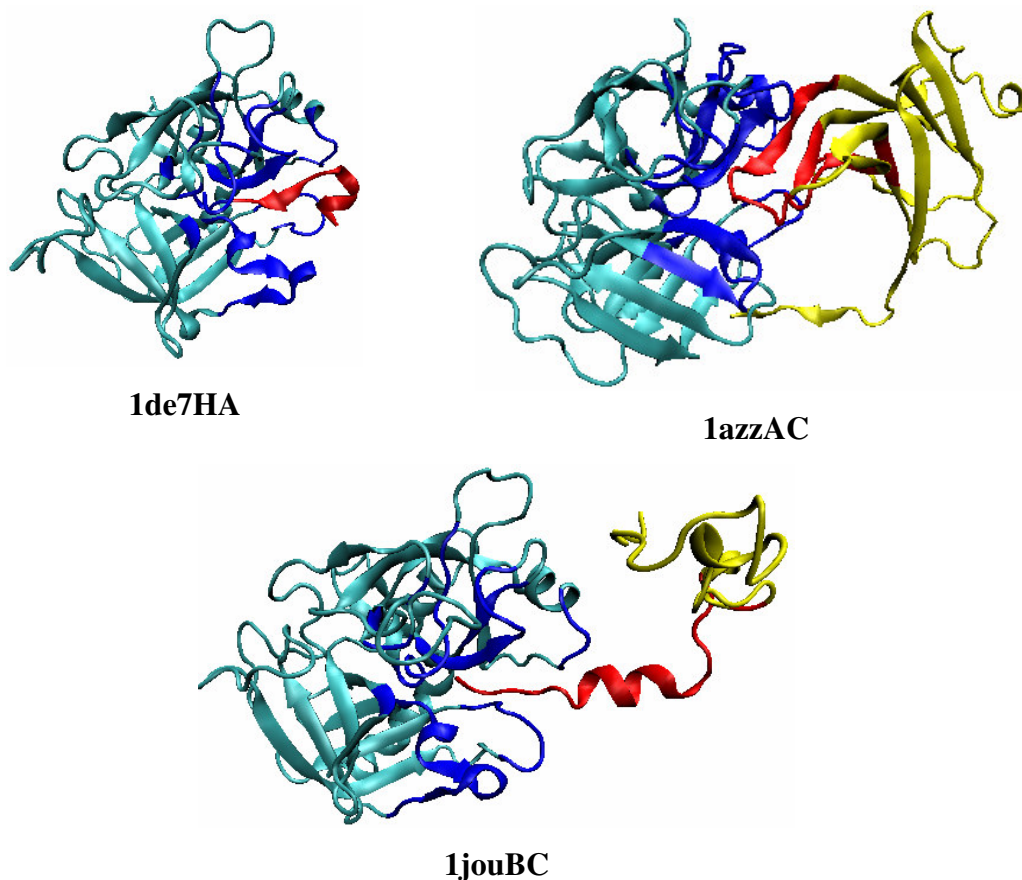


Figure 5.2 Type II cluster examples.

5.1.3. Type III Clusters

In Type III clusters, one side of the interface is conserved structurally, other side is changing. Cluster #425 is an example of Type III clusters which has 13 members after elimination of the sequentially redundant interfaces. Four of these 13 members are shown in **Figure 5.3**. In this figure, blue colored parts are the conserved regions of the cluster, red colored part are the different binding partners of these conserved regions. Here, all of the cyan colored chains and their corresponding interface residues colored blue are classified as Trypsin-like serine proteases in SCOP. In this example, domain of one partner chain is conserved.



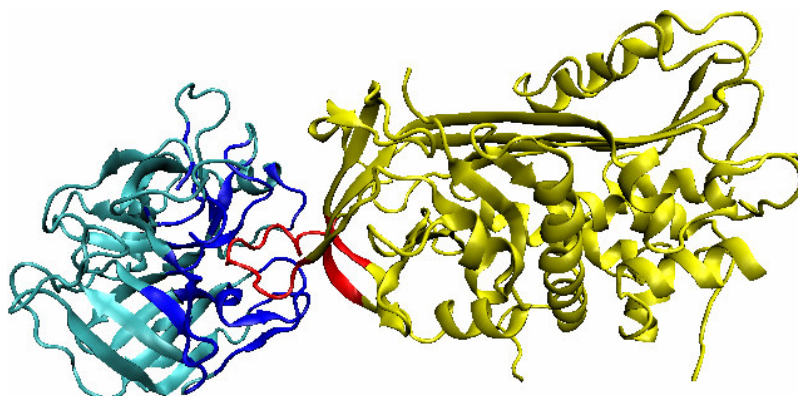
**1sr5AC**

Figure 5.3 Type III cluster examples.

5.2. Relationship between Interfaces and Folds

Non-covalent contacts between residues are crucial for both folding and binding processes [15]. Protein folding and binding are similar processes in principle which means that inter- and intra- recognition of the molecules are related. There are obvious similarities between folds and binding sites. Literature review about this similarity is available in **Chapter 2**.

In this work, the correspondence between binding and folding is analyzed with a different approach. To observe similarities between folds and interfaces, we compared the representative interfaces of the top 10 populated clusters with the most populated folds in SCOP. In SCOP, the most populated folds are the immunoglobulin like (represented by PDB id: 1fna), Rossman (PDB id: 3chy), TIM barrel (PDB id: 1ypi), jelly roll (PDB id: 1sac), α - β plait (PDB id: 1ris), 3 - helix bundle (PDB id: 1enh), globin (PDB id: 1a6n), and β - grasp (PDB id: 1pqb) folds. MULTIPROT is used to compare them structurally with the interfaces. Interestingly, the most popular folds in single chains are those architectures which are most populated in the interfaces.

Table 5.1 Comparison of the most populated folds with the representative interfaces of the most populated clusters. "aa" is the number of amino acids. The interface pdb id nomenclature is first the pdb code, followed by the two chains. Thus, for the first interface entry, 1djr is the pdb code for the complex, and the interface is between chains D and E. The number of matched residues is taken from Multiprot's output.

Fold name	Fold size	Interface Name	Interface Size (aa)	# of Matched residues [MULTIPROT]	RMSD (Å)
Iris	97	1djrDE	119	58	1.92
1enh	54	1o1pAD	156	38	1.86
1a6n	151	1z6oAE	131	62	1.84
1fna	91	1newLH	215	52	1.70
1pgb	56	1nh0AB	130	31	2.00

In **Table 5.1**, the RMSD values and the number of matching residues between the folds and interfaces of the 5 best matches are given. In this table, first two columns give the PDB name of the fold and its size. Third and fourth columns give the name of the compared interface and its amino acid size. Fifth column represents number of matched residues between the compared fold and interface according to the Multiprot results. Last column demonstrates the RMSD value between these matched residues.

To focus on more detailed to the relationship between folding and binding, here, some of the cases tabulated in **Table 5.1** are illustrated schematically. For example, the correspondence between the interface 1o1pAD and the 3-helix bundle fold 1enh is shown in **Figure 5.4**. In **Figure 5.4**, the ribbon diagrams and the matching parts of 1enh and 1o1pAD are shown which highlights this similarity visually. In **(A)**, on the left the ribbon diagram of the interface named 1o1pAD is shown colored according to the chains. Cyan colored part is coming from chain A, orange colored part is coming from chain D. On the right, the ribbon diagram of the 3 – helix bundle fold (1enh) is shown. In **(B)**, the matching parts between 1o1pAD and 1enh are shown (38 residues). The red colored parts are the

matching parts of the interfaces and folds, the yellow colored parts are the unmatched parts of the structures. 1o1pAD is an interface composed of only helical structures, contains 156 residues. When it is compared with the 3-helix bundle fold (1enh) 38 out of 54 residues in 1enh match structurally with the residues in 1o1pAD. The RMSD value is 1.86 Å for this match which is an appropriate value to declare that the structure of the interface 1o1pAD is similar to the 3-helix bundle fold.

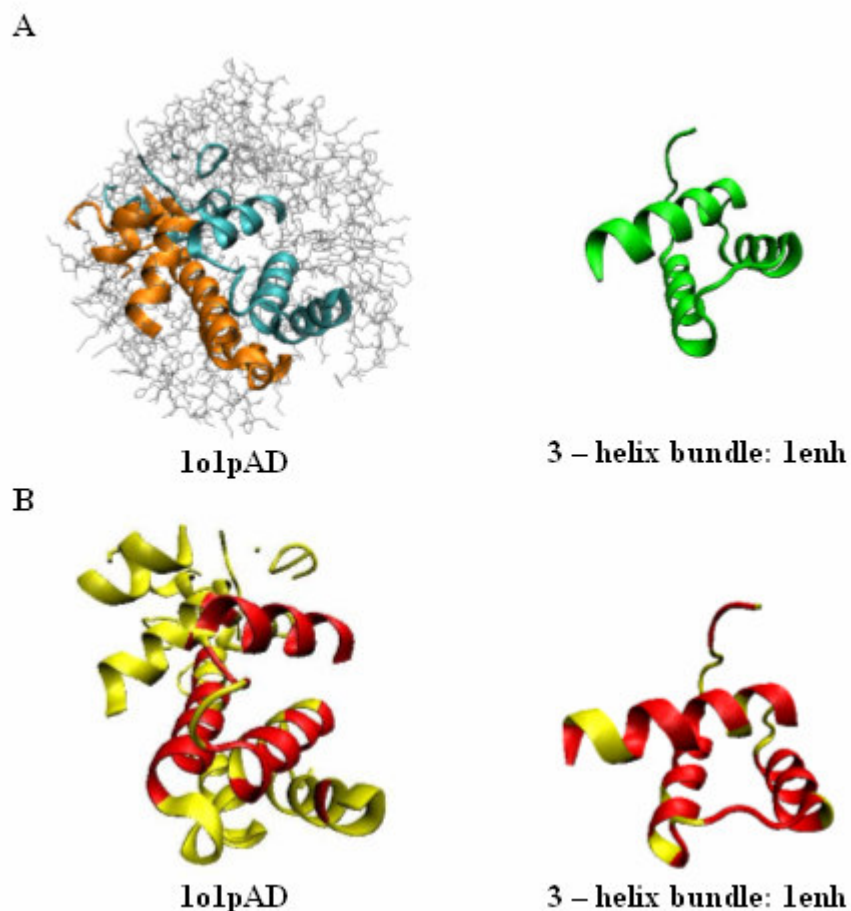


Figure 5.4 Comparison of the 1o1pAD with 1enh (3 – helix bundle fold).

Another example is the similarity between the α – β plait fold (1ris) and the interface 1djrDE as shown visually in **Figure 5.5**. 1djrDE is composed of both helices and strands

like Iris. The ribbon diagram shown in **Figure 5.5 (A)** gives some clues about the similarity between these two structures.

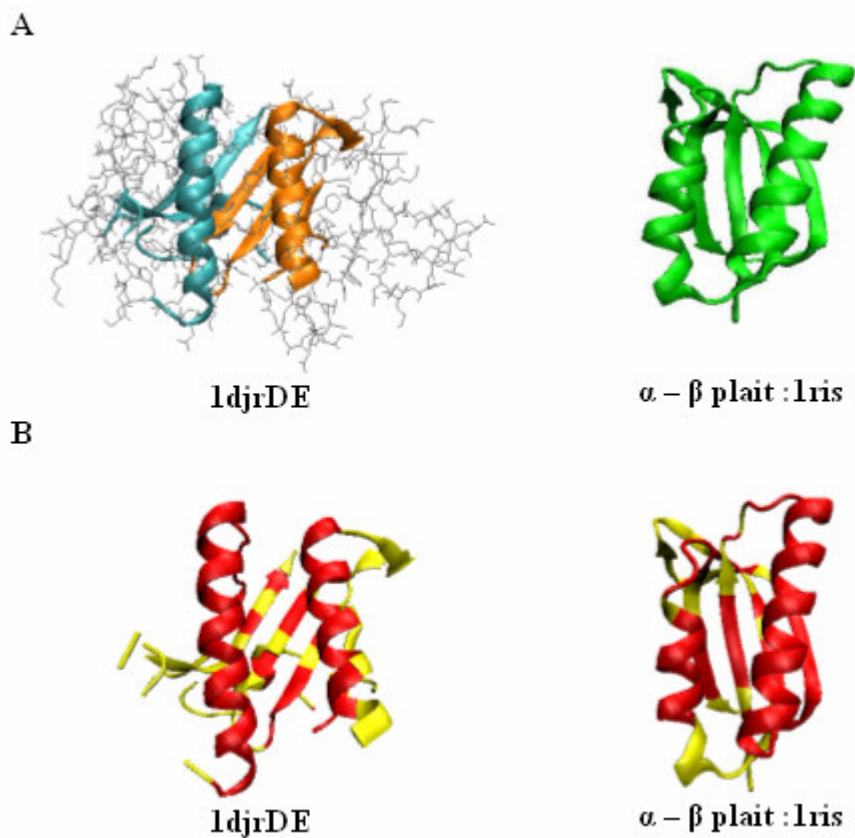


Figure 5.5 Comparison of the 1djrDE with 1iris ($\alpha - \beta$ plait : 1iris).

In (A), on the left the ribbon diagram of the interface named 1djrDE is shown colored according to the chains. Cyan colored part is coming from chain D, orange colored part is coming from chain E. On the right, the ribbon diagram of $\alpha - \beta$ plait fold (1iris) is shown. When they are compared structurally with each other by Multiprot, a high matching ratio is obtained. 58 out of 97 residues in 1iris are matching structurally with the residues on the 1djrDE. The RMSD value between these 58 matching residues is 1.92 Å. These matching

residues are also illustrated in **Figure 5.5 (B)**; the red colored parts are the matching parts of the 1djrDE and 1ris, the yellow colored parts are the unmatched parts of the structures.

Figure 5.6 also illustrates two more examples of the correspondence between folding and binding. In these figures, same coloring methods are used as in previous examples: the red parts are matched residues; the yellow parts are unmatched residues. The RMSD value for structural matching between 1z6oAE and 1a6n, shown in **Figure 5.6 (A)**, is 1.84 Å between 62 residues. In (B), the similarity between the upper part of the 1ncwLH and immunoglobulin fold 1fna can be distinguished visually. Also the result of the structural matching between 1ncwLH and 1fna supports this similarity. 52 of the 91 residues of 1fna are matching with the residues 1ncwLH. RMSD between these 52 matched residues is 1.70 Å.

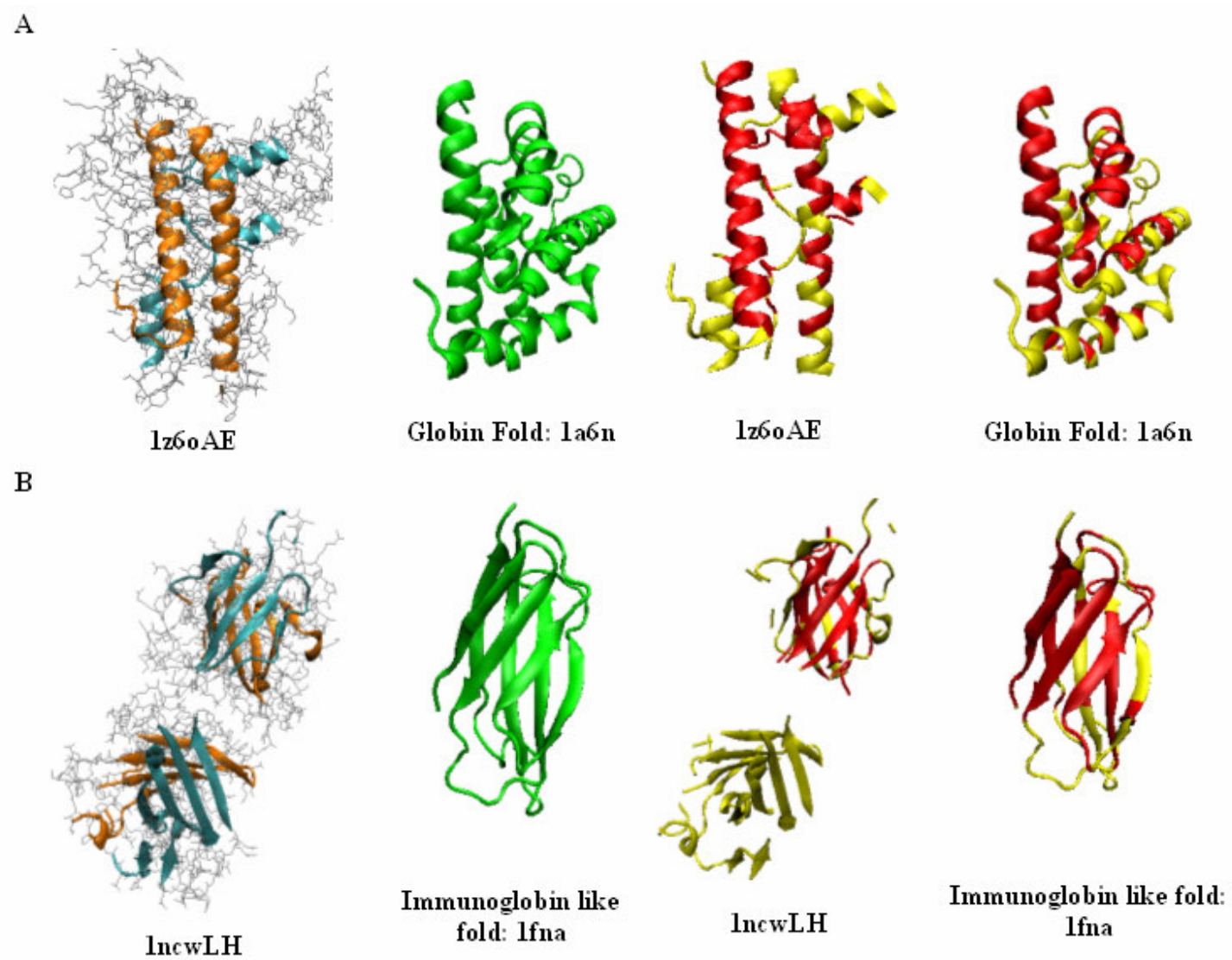


Figure 5.6 Two more examples for similarity between folding and binding.

5.2. Domain Fusion Events in the New Extracted Data Set

This section involves overall domain fusion events available in the new generated dataset. First, the domain fusion maps are presented. In the continuing, part domain fusion is supported by some case studies. General information about domain fusion is available in **Section 2.6**.

5.2.1 Overall Domain Fusions in the Dataset

According to the domain fusion definition in **Section 2.6**, by using our new generated dataset of protein interfaces, we generated a domain fusion map of structures in PDB. To generate this map, all chains in PDB are also used. If the interaction of two domains is intra- in one species, inter- in another species, we proposed domain fusion between these two domains is acceptable. In the construction step of the domain fusion map, only one domain containing partner chains of the interfaces are considered for inter domain interactions and two domain containing single chains of whole PDB structures are considered for intra domain interactions. As a result of these strict interactions, we obtained 106 unique domain fusions between 114 domains at super family level of SCOP. This domain fusion map is illustrated in **Figure 5.7** which does not contain domain fusion events occurring between same super families.

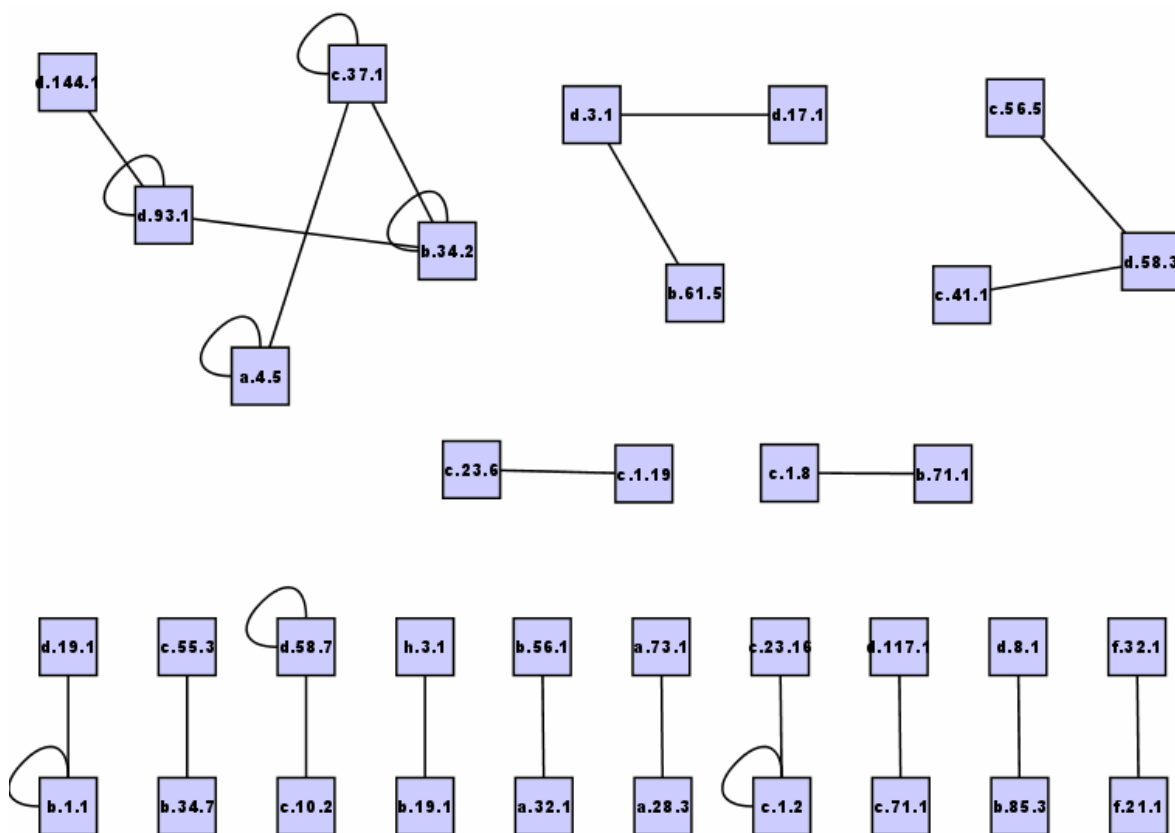


Figure 5.7 Domain fusion map of the PDB derived from new generated interface dataset. The fusion events occurring between same superfamilies are not shown in this map. Full list is available in <http://prism.cccb.ku.edu.tr/interface>

A more general map is constructed without the two domain pair limitation. This map contains 224 fusions between 223 domains. In **Figure 5.8**, intersection of this general domain fusion map of the structures in PDB (as of February 2006) with the domain fusions available in literature is illustrated, where nodes represent the domains, edges represent domain fusion between two domains. In **Figure 5.8**, there are 28 domain fusion events between 25 domains. However, in the dataset there are much more fusion events available. Our interface dataset picks up 224 domain fusion events between 223 domains. The full list of fused domains is accessible through the web site <http://prism.cccb.ku.edu.tr/interface>.

This map elucidates how domain fusion events organized on a large scale in the super family level.

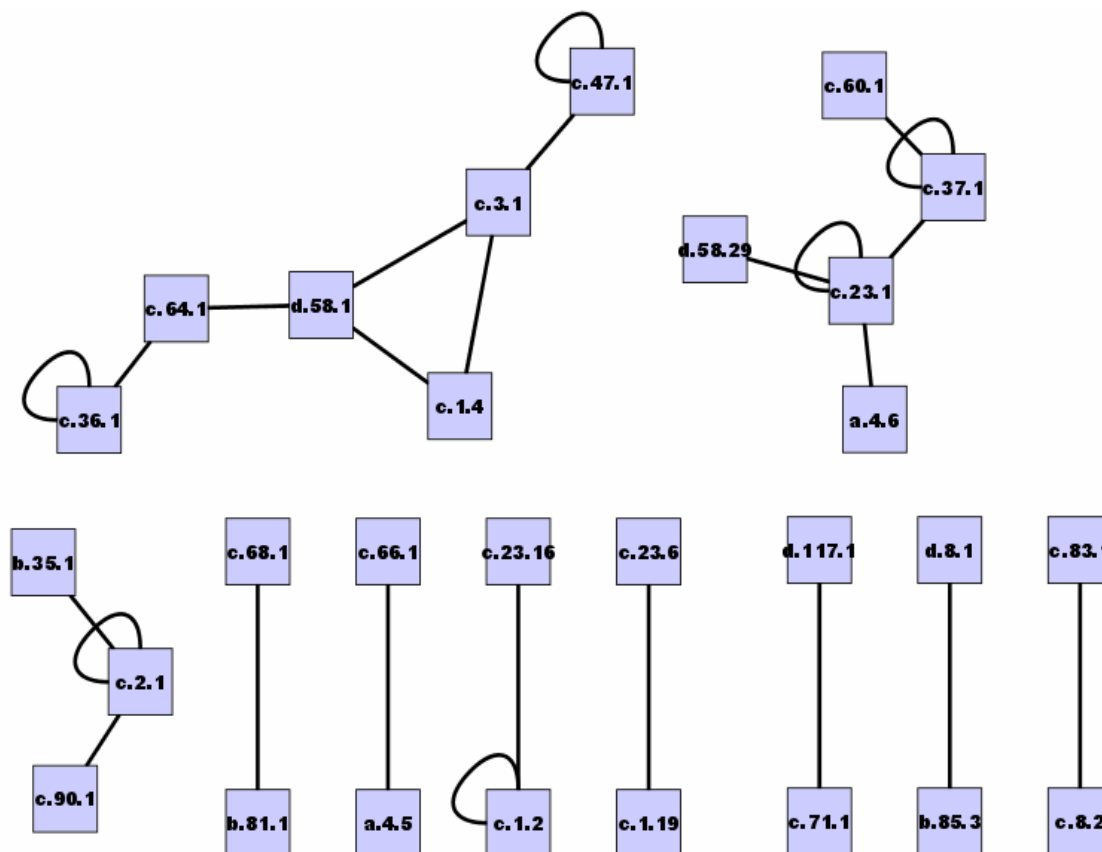


Figure 5.8 Intersection of the general domain fusion map of the structures in PDB (as of February 2006) with the domain fusions available in literature [48].

5.2.2 Case Studies in the Dataset

SCOP version 1.71 [52] is used to identify the domain classification. An example of gene fusion is seen in the enzyme urease as shown in **Figure 5.9**. The enzyme is composed of four domains, a urease, gamma-subunit (d.8.1.1), a urease, beta-subunit (b.85.3.1), an alpha-subunit of urease, catalytic domain (c.1.9.2), and an alpha-subunit of urease

(b.92.1.1). The numbers in the parentheses are the SCOP Domain IDs. The beta and gamma subunit domains are located on separate chains in *Klebsiella aerogenes* (pdb ID: 1a5k) and *Bacillus pasteurii* (pdb ID: 1ubp) but in *Helicobacter pylori* (pdb ID: 1e9y) are fused, located in the same chain. As a result of this possible fusion event, 1e9y produces only one interface (1e9yAB), 1a5k produces two interfaces (1a5kAC, 1a5kBC) and 1ubp produces three interfaces (1ubpAB, 1ubpBC, 1ubpAC). Although the interaction between the gamma- (d.8.1.1) and beta- subunits (b.85.3.1) are conserved in *Helicobacter pylori* and *Bacillus pasteurii*, there is no interaction between these domains in *Klebsiella aerogenes*. However, the absence of this interaction does not affect the function of urease in the organism.

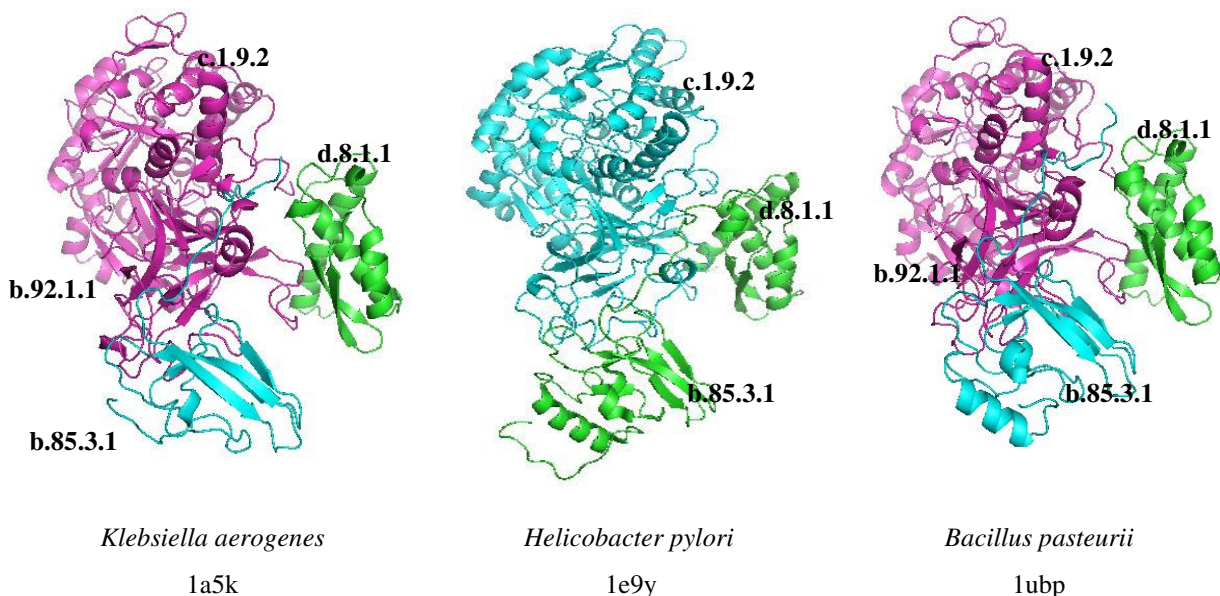
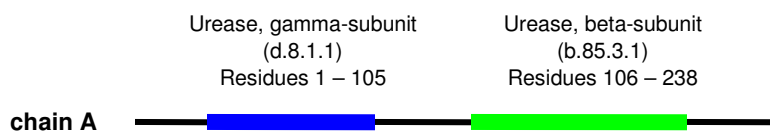


Figure 5.9 Gene fusion examples in urease enzyme. Gamma and beta subunits of the urease from *Klebsiella aerogenes* (1a5k A and B), *Bacillus pasteurii* (1ubp A and B) and *Helicobacter pylori* (1e9y A) are interacting in similar way and functioning in urease activity. Chains are colored differently to lead the eye.

As expected, the surface area is small between the two mentioned domains, for 1ubpAB it is 419 Å², and according to the NOXclass output this interaction is a biological interaction with a probability of 66.85%. Our structure-based clustering allows straightforward detection of such cases. Interfaces 1a5kAC and 1ubpAC are in the same cluster; 1e9yAB is a member of another cluster, although 1e9yAB contains a similar interaction between domains c.1.9.1, b.92.1.1 and d.8.1.1. In **Figure 5.10**, the schematic representation of this domain fusion event is illustrated.

Helicobacter pylori PDB:1E9Y



Bacillus pasteurii PDB:1UBP

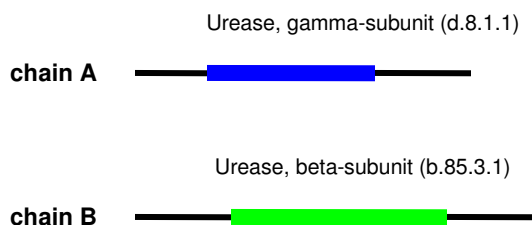


Figure 5.10 Schematic representation of the domain fusion.

Another example for gene fusion in the current data set is imidazole glycerophosphate synthase enzyme (**Figure 5.11**); observed also in other studies. This enzyme consists of two domains, the histidine biosynthesis (c.1.2.1) and class I glutamine amidotransferase (c.23.16.1).

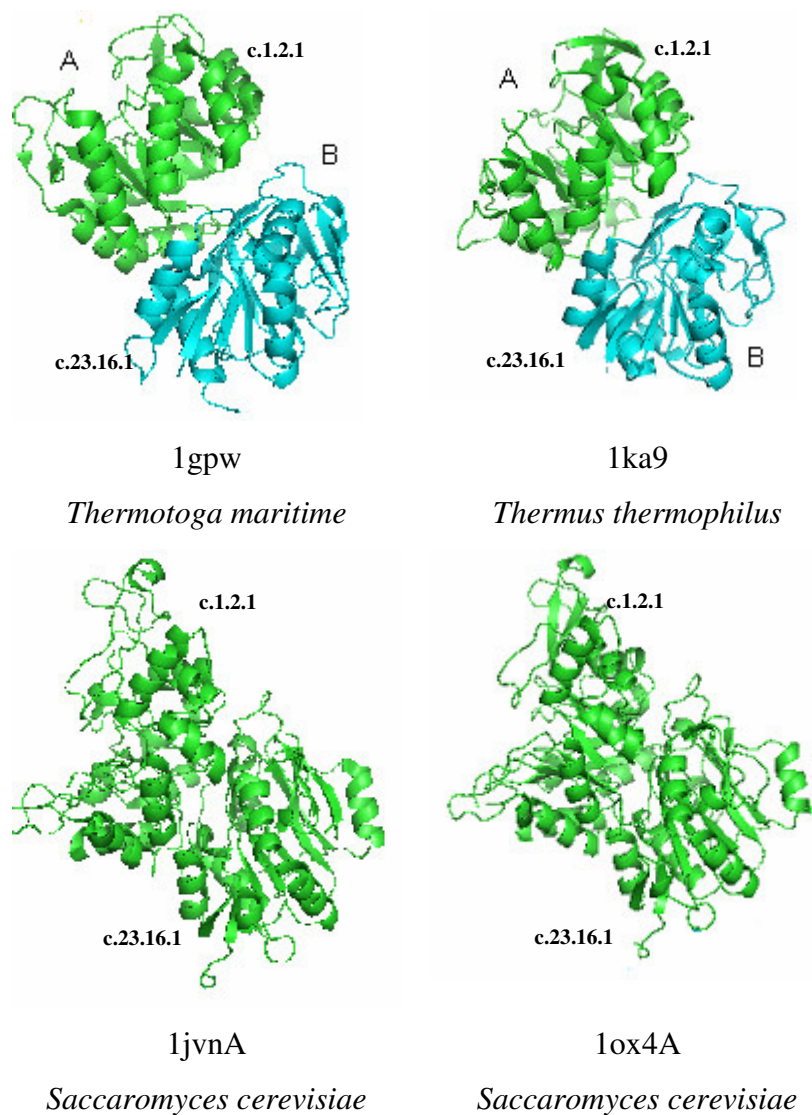


Figure 5.11 Gene fusion example in imidazole glycerophosphate synthase enzyme. The histidine biosynthesis (c.1.2.1) and class I glutamine amidotransferase (c.23.16.1) domains from *Thermotoga maritima* (1gpw A and B), *Thermus thermophilus* (1ka9 A and B) and *Saccharomyces cerevisiae* (1jvn A, 1ox4 A) are interacting in similar way and functioning in imidazole glycerophosphate synthase enzyme activity.

In *Thermotoga maritima* (1gpwAB, 1gpwAD, 1gpwBE, 1gpwCD, 1gpwEF) and *Thermus thermophilus* (1ka9FH), these two domains interact while belonging to separate

chains; however, in *Saccharomyces cerevisiae* (1jvnA, 1jvnB, 1ox4A, 1ox4B, 1ox5A, and 1ox5B) these two domains are fused and interact in the same chain. In **Figure 5.11**, the proteins are colored by their chains. Despite the *Saccharomyces cerevisiae* gene fusion, these domains interact in a way similar to the domains in other species (*Thermotoga maritime*, *Thermus thermophilus*). 1gpw has six chains which form 5 interfaces whereas 1ka9 has 2 chains forming a single interface.

To support the domain fusion events in our dataset, one more case is presented here. The domain-domain interaction between f.21.1.2 (Cytochrome b of cytochrome bc1 complex) and f.32.1.1 (a domain/subunit of cytochrome bc1 complex) is available in some species as intra-, in some others as inter- contact. In *Gallus gallus*, *Saccharomyces cerevisiae* and *Bos taurus* f.21.1.2 and f.32.1.1 contact is intra- and occurs through the chain 1bccC in *Gallus gallus* and 1be3C in *Bos taurus*. However, in *Chlamydomonas reinhardtii* and *Mastigocladus laminosus* (bacteria), it is inter-chain contact. These domains interact with each other through the interfaces 1q90BD in *Chlamydomonas reinhardtii* (green algae, protistae) and 1vf5AB *Mastigocladus laminosus* (bacteria) in the interface dataset. In **Figure 5.12**, proteins having domain fusion between f.21.1.2 and f.32.1.1 are illustrated. Proteins are colored according to the domains; cyan represents f.21.1.2 (Cytochrome b of cytochrome bc1 complex) and green represents f.32.1.1 (a domain/subunit of cytochrome bc1 complex). The PDB names of the proteins and their corresponding species are also shown in **Figure 5.12**.

In this example, we see that prokaryotes prefer to interact through inter- domains. On the other hand, the eukaryote species such as *Gallus gallus*, *Bos taurus*, also *Saccharomyces cerevisiae* choose intra-domain interactions in the evolutionary constraints. When we look up other cases from this perspective, we also see such a trend that the more primitive species prefer inter domain interaction, but evolutionarily advanced species prefer

the interactions having better than random chance in their cells, in other words intra domain interactions. This trend may be the result of the crowded traffic in the cells of eukaryotes.

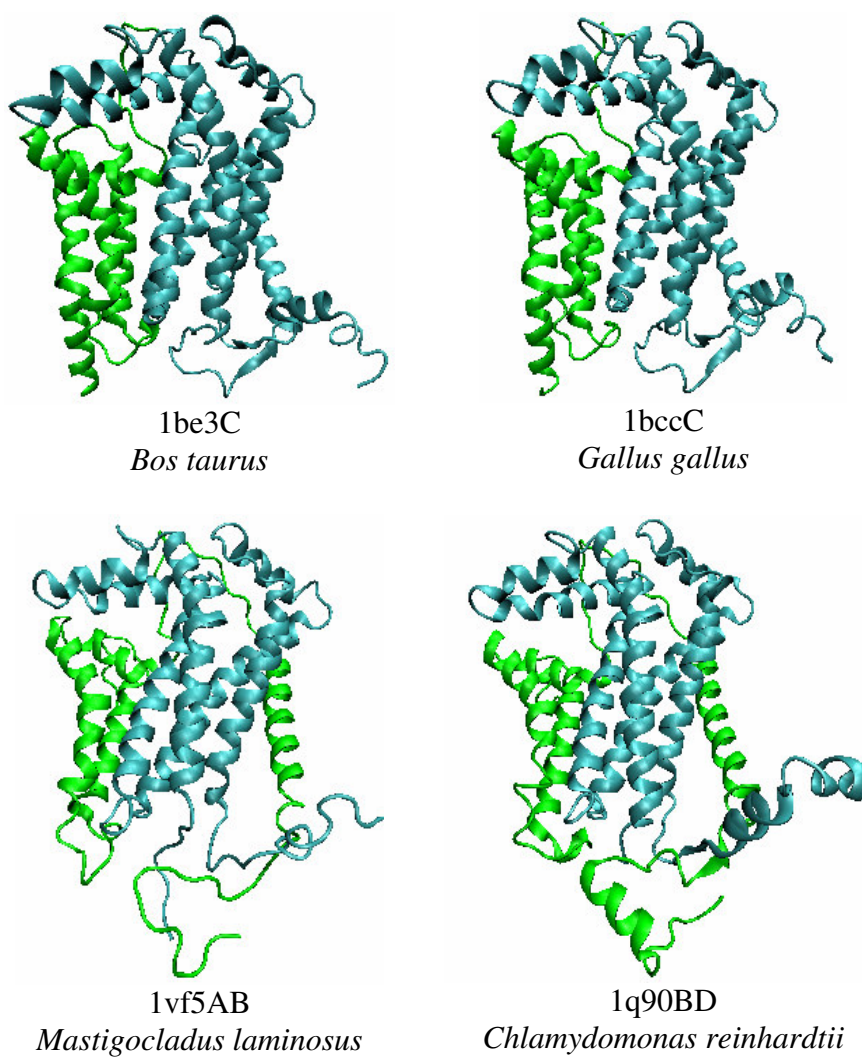


Figure 5.12 Domain fusion events in cytochrome BC1 complex.

Chapter 6

CONCLUSION

Organization of the large scale protein-protein interaction datasets allows analyses of interface properties, such as sequential and structural conservation, residue propensities, interface size, shape and complementarities. These provide insight into the types and evolutionary history of protein interactions. Interfaces are used as a framework for prediction of interactions; examination of interface architectures and provides clues to protein recognition mechanisms.

In this study, we started with all structures available in PDB. After extraction of the protein – protein interfaces, they are clustered structurally. As a result, 8205 structurally non-redundant interface clusters are generated. Comparison of the new 8205 interface-cluster dataset with the older (1994 and 2002) datasets indicates that the number of clusters has increased through the 12 years period much more rapidly than the available PDB structures and SCOP domains. This growth largely stems from the larger numbers of multi-chain and high molecular weight proteins in the PDB. The increased cluster diversity implies discovery of new interface architectures; this is in addition to the observed increase in similar interface structures, as indicated by more populated clusters. We observe that new interface architectures have emerged as well as existing interface clusters have become more populated. The number of unique interfaces continues to grow exponentially and has still not reached its upper limit. After elimination of the peptides, membrane proteins, antigen-antibody complexes and ligands, we currently end with 3086 unique

interface clusters. Plotted against the older data sets, we detect an increase in the unique type of interface architectures at a rate of approximately 350 new architectures per year.

The current number of distinct clusters allows analysis of the functional divergence and evolution of the interfaces. We further studied the coverage of protein function- function interaction maps. We observe that the functional coverage is still not complete; only 12 functions out of 18 are found in the PDB. The total number of different interfaces is predicted to be around 8000. The time needed to discover all these interaction types is found to be almost 30 years at the current pace of experimental structure determination.

We applied our dataset a filtering process to eliminate homologous chains. As a result, interface clusters reduced from 8205 to 213. These clusters are separated into 3 types according to their interface similarities and global fold similarities. Type I clusters include similar interfaces and same partner domain pairs. Type II clusters contain similar interface architectures; however, these architectures form as a result of interaction of different domain pairs. Type III clusters have interfaces whose one side is conserved structurally, and other partner side is changing. Also, complementary chains have different domain folds. These cluster types elucidate the fact that some architecture can be preferred by different domain pairs, and also in some interactions one side is conserved and the type of interaction changes with the other partner chain.

Binding and folding are similar processes. Based on this statement, we detected similarities between domain folds and our interfaces. We observed in our dataset that some interface architectures are more favorable and frequently used in protein-protein associations. Not surprisingly, when these interface architectures are compared with most populated domain folds, we observe a high degree of similarity. Therefore, nature uses similar preferred fold templates for single chains and for interfaces.

As an application of our newly generated interface dataset, a domain fusion map is generated and we noticed a trend that eukaryotes prefer intra chain interactions; on the

other hand primitive species choose inter chain interaction for some domain-domain interactions to perform a cellular function. The presented case studies support also this trend.

In conclusion, this nonredundant dataset of protein – protein interfaces is a rich source for studies about protein – protein interactions such as detection of the binding region patterns, identification of the critical residues for protein interactions, protein function prediction, drug design etc. This dataset has been template for PRISM web server to predict putative protein – protein interactions and HotSprint database to identify hot spots on the protein interfaces.

Appendix A

APPENDIX

A.1. Interface Extraction Methods

An interface can be defined as the set of amino acids which represents a region that links two polypeptide chains in a protein structure by non-covalent interactions. Protein-protein interactions take place through interfaces. To identify protein interfaces there are three methods in the literature. First, the atomic distance calculation method; considers the distance between two atoms belonging different chains. Second method is accessible surface area (ASA) calculation method which considers the difference between monomer and complex ASA values of the chains. In other words, the buried and inaccessible to solvent residues are identified as interface residues. Third method is Voronoi Diagram, a geometric approach to interface identification. This geometrical method is also atomic distance based. Because the mostly preferred methods are the atomic distance calculation and ASA calculation methods, in the continuing part only these two methods are explained.

A.1.1. Atomic Distance Calculation Method

Atomic distance calculation method depends on the distance between any two atoms each from one chain. If the distance is smaller than a cutoff value, these two atoms are identified as interface atoms. There are several applications of this method. Davis et al. (2005) used a cutoff value of 6.05 Å to allow water mediated contacts also. In the studies of Tsai et al., Keskin et al., and in the current work two types of residues are defined; interacting and nearby residues. If distance between two atoms each from one chain is

smaller than the sum of van der Waals radii of these two atoms plus 0.5 Å then these atoms are defined as interacting atoms. Nearby residues are defined as if distance between Ca of a residue and Ca of an interacting residue is less than 6 Å. Gong et al defined the interface residue as if its atoms are within 5 Å cutoff (PSIMAP method). This cutoff value considers van der Waals radii of the interacting atoms and a solvent molecule like water.

2.1.2. Accessible Surface Area Calculation Method

ASA method is based on the detection of the buried parts of the proteins after complexation. When two partner chains form a complex they lose some of their solvent accessible parts. ASA method considers the buried regions of the proteins. If a residue loses more than 1 Å² after complexation, then this residue is identified as interface residue. As a methodology, first the ASA of the residues are calculated at their monomer state. Then, the complex ASAs of the residues are calculated. The difference between monomer ASA and complex ASA is compared with the threshold of 1 Å². If it is larger than the threshold then, this residue is identified as interface residue.

These 3 approaches cover each other with a high percentage. Gong et al. have used all these three methods in their dataset and compared these three methods. As shown in Figure 2.1, according to data of Gong et al, there are small differences between these three methods [41]. Also, overlap of the most used two methods ASA calculation and atomic distance calculation method is high.

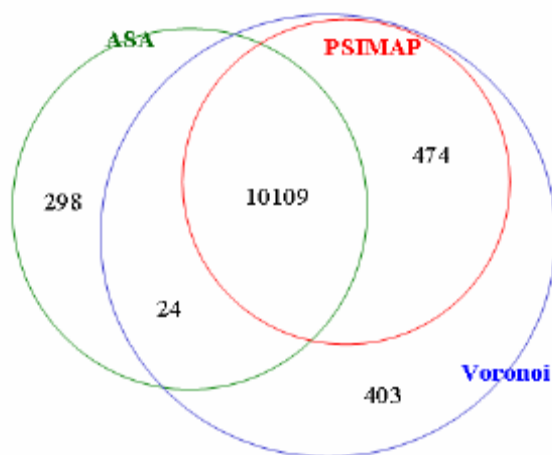


Figure A.1 Overlap of three interface identification methods [41].

A.2. Geometric Hashing Algorithm

In this algorithm each residue is considered as a point in 3-D space. The algorithm uses only the C^α coordinates and does not consider connectivity of the C^α points in the matching. There are three consecutive steps: (i) hash table construction, (ii) voting and (iii) extension step.

i. Hash Table Construction

To find local similarity between two points, hash table is constructed. For every consecutive three C^α atoms, an orthogonal reference frame is identified. Orthogonal frames are calculated by following the equations below.

$$R_x = C_{i-1} - C_i$$

$$V_t = C_{i+1} - C_i$$

$$R_z = R_x \times V_t$$

$$R_y = R_x \times R_z$$

Here, (R_x, R_y, R_z) is the x, y, z axes of the new orthogonal frame. All residues within a threshold of 15 Å around the center of the frame are projected onto the orthogonal frame of i. In the Hash Table, each entry is identified with its orthogonal frame and the new coordinates of the atoms.

ii. Voting

In this step, same procedure explained in part(i) is applied also to the second protein. After construction of the new frames for both proteins, The voting step contains the comparison of two proteins and calculation of the RMSD value between these two structures. According to the local similarities between them the transformation vector and rotation matrix are calculated and the Ca atom pairs are superimposed. After superimposition of the Ca's, the RMSD values between the matching parts are calculated.

iii. Extension

In this step, top local alignments obtained in voting step are extended iteratively to uncover best global alignments.

The similarity between two interfaces is calculated by considering the size difference of these two interfaces and also the percentage of the identical residues matches. Geometric Hashing algorithm generates a record of the matched atom pairs. An individual score is assigned for each matched pairs. If both of the terminal atoms, as shown in **Figure A.2**, matches with each other, then the score of the A_i, B_j match is 1.0. If only one terminal atom is matched, then the score is 0.75. A score of 0.5 is given when besides one matching pair of terminal residues; other terminal residue is matched with a noncontiguous residue. If the terminal residues are not matched with each other, then the score of this match is zero. The connectivity score is the sum of the individual scores of each matched pairs. The term

“relative connectivity score”, which is used as a parameter in the clustering step, is the division of the connectivity score to the size of the interface.

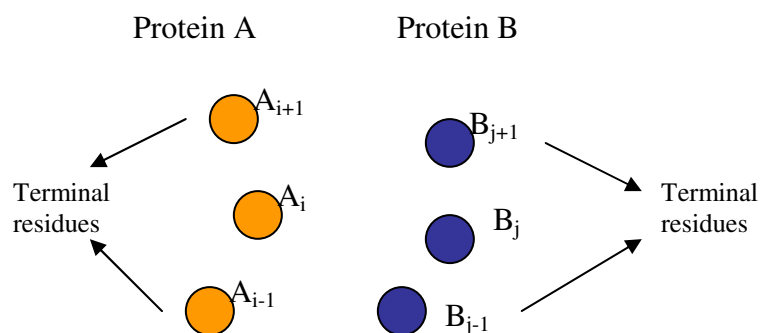


Figure A.2 Representation of the terminal residues.

A.3. Webservers, Softwares, Tools, Databases

A.3.1. NACCESS

Naccess is a program used for the calculation of the accessible surface areas of the molecules. It basically rolls a solvent probe on the desired molecule. The radius of the solvent can be chosen by the user, but the default value is 1.4 Å. The path gained by the center of the probe gives the accessible surface area. Naccess takes files in PDB format as input. Besides the accessible surface area, the output file of the Naccess gives also relative accessible area for each individual residue. Relative accessibility can be described as the percent accessibility of a residue relative to the accessibility of it in the tripeptide ALA-X-ALA. Generally if this value is larger than 5% then, this residue is identified as surface residue [9, 41]. In this work, we used Naccess with default values to calculate ASA [54].

Usage with default values: `naccess protein1.pdb protein2.pdb`

A.3.2. SURFNET

SURFNET is a program which uses PDB file format to take 3D coordinates of the molecules and generates molecular surfaces and measures gaps between molecules. This program is useful especially to define protein active sites. Gap regions on the molecules are the empty spaces between two or more than two molecules. Also, gap region can be available in a single molecule which corresponds to internal cavities and grooves. SURFNET outputs can also be visualized and rendered by some molecular modeling and graphics packages. In this study, only gap volume calculation skill of the SURFNET is used. Gap volumes between two partner chains of the protein interfaces are calculated by SURFNET [55].

A.3.3. MULTIPROT

Multiprot is fully automated software which identifies multiple structural alignments of a given set of protein structures. Structural alignment method is based on the Geometric Hashing Algorithm which detects common parts of the given structures in all possible ways. This is a sequence-order and directionality independent algorithm. Multiprot considers only C^α atoms. In the output file, the matched residue pairs, number of them and the RMSD value between these residues are present. The algorithm does not force all residues to participate in the alignment; on the contrary, it searches the best scored partial alignment for the given structures. In parameters file, by changing parameter user can change the alignment conditions. Its sequence order independent feature makes Multiprot appropriate for protein interfaces analysis. Multiprot is used both in clustering part and in cluster type separation part [51].

A.3.4. CLUSTALW

ClustalW is a multiple sequence alignment program for protein or DNA sequences. As input sequence information of the desired proteins or DNAs are given and in the output the multiple sequence alignment of these structures are produced by the program. It calculates the best match and shows the similarities, differences and identities. In global alignments overall sequences are aligned by using gaps. In local alignments, only particular regions are aligned to each other. ClustalW uses global alignment for multiple sequence alignment. It has some options like input file format, substitution matrix preference, etc... In the output, besides the multiple sequence alignment, pairwise alignments of the sequences and their scores are also provided. Phylogenetic trees are also produced by the multiple sequence alignment [50].

A.3.5. CytoScape (network visualization and analysis)

Cytoscape is molecular interaction network visualization software which also integrates biological information such as gene expression profiles, GO annotations etc... Additional features like network analyzer, functional enrichment generator, and additional file format support can be installed as plugins. Cytoscape user can visualize the protein – protein interaction network or other networks by loading .sif file which contains pairwise interaction information. Network visualization properties such as node shape, color, edge shape, color etc. can be defined by the user. It has also various filtering and selection tools. The more, the resulting graph can be organized several layouts such as hierarchical layout, spring embedded layout, circular layout etc [56]. Here, we used Cytoscape for visualization of functional interaction network of PDB. Cytoscape is downloadable through the web page <http://www.cytoscape.org/>.

A.3.6. VMD (molecule visualization)

VMD is a molecule visualization and analysis tool. Biological systems such as proteins, nucleic acids, lipid bilayer assemblies, etc. can be visualized by the help of VMD. VMD can read standard Protein Data Bank (PDB) files and display the contained structure. It has various molecular representation methods and an advanced coloring and rendering properties. VMD can be used also to animate and analyze the trajectory of molecular dynamics (MD) simulations, and can interactively manipulate molecules being simulated on remote computers (Interactive MD) [57].

A.3.7. Other

SCOP, the Structural Classification of Proteins, contains the detailed description all protein structures available. SCOP has a hierarchical structure. The most specialized levels are family and super family levels; represent the near and far evolutionary relationships. In the upper levels, fold and class information is available [52]. SCOP is accessible through the web site <http://scop.mrc-lmb.cam.ac.uk/scop/>

DSSP, Dictionary of Protein Secondary Structure, is the secondary structure assignment database for all structures available in PDB. It assigns each residue one of the eight secondary structure types (B, E, G, H, I, S, T, U). In this work we group these eight types into helix (H), beta (B) and loop (T). G, H and I are assigned as helix; B and E are beta and S, T and U are loop [53].

GO, Gene Ontology, is a project to describe gene products in several databases. Numerical identifiers are used for each annotation in GO, such as GO:nnnnnnnn. GO is categorized into 3 classes; molecular function, cell localization and biological process which describes the molecule according to its cellular role [58, 59]. GO is accessible through <http://geneontology.org>.

BIBLIOGRAPHY

- [1] von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417**(6887):399-403.
- [2] Chothia C, Janin J: **Principles of protein-protein recognition.** *Nature* 1975, **256**(5520):705-708.
- [3] Jones S, Thornton JM: **Principles of protein-protein interactions.** *Proc Natl Acad Sci U S A* 1996, **93**(1):13-20.
- [4] Tsai CJ, Nussinov R: **Hydrophobic folding units derived from dissimilar monomer structures and their interactions.** *Protein Sci* 1997, **6**(1):24-42.
- [5] Chothia C: **Proteins. One thousand families for the molecular biologist.** *Nature* 1992, **357**(6379):543-544.
- [6] Aloy P, Russell RB: **Ten thousand interactions for the molecular biologist.** *Nat Biotechnol* 2004, **22**(10):1317-1321.
- [7] Tsai CJ, Lin SL, Wolfson HJ, Nussinov R: **A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique.** *J Mol Biol* 1996, **260**(4):604-620.
- [8] Keskin O, Tsai CJ, Wolfson H, Nussinov R: **A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications.** *Protein Sci* 2004, **13**(4):1043-1055.
- [9] Aytuna AS, Gursoy A, Keskin O: **Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces.** *Bioinformatics* 2005, **21**(12):2850-2855.

-
- [10] Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A: **PRISM: protein interactions by structural matching**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W331-336.
- [11] Valdar WS, Thornton JM: **Protein-protein interfaces: analysis of amino acid conservation in homodimers**. *Proteins* 2001, **42**(1):108-124.
- [12] Nooren IM, Thornton JM: **Structural characterisation and functional significance of transient protein-protein interactions**. *J Mol Biol* 2003, **325**(5):991-1018.
- [13] Jones S, Thornton JM: **Analysis of protein-protein interaction sites using surface patches**. *J Mol Biol* 1997, **272**(1):121-132.
- [14] Larsen TA, Olson AJ, Goodsell DS: **Morphology of protein-protein interfaces**. *Structure* 1998, **6**(4):421-427.
- [15] Ofran Y, Rost B: **Analysing six types of protein-protein interfaces**. *J Mol Biol* 2003, **325**(2):377-387.
- [16] Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES: **Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?** *Protein Sci* 2004, **13**(1):190-202.
- [17] Chakrabarti P, Janin J: **Dissecting protein-protein recognition sites**. *Proteins* 2002, **47**(3):334-343.
- [18] Bahadur RP, Chakrabarti P, Rodier F, Janin J: **Dissecting subunit interfaces in homodimeric proteins**. *Proteins* 2003, **53**(3):708-719.
- [19] Thorn KS, Bogan AA: **ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions**. *Bioinformatics* 2001, **17**(3):284-285.
- [20] Fischer TB, Arunachalam KV, Bailey D, Mangual V, Bakhru S, Russo R, Huang D, Paczkowski M, Lalchandani V, Ramachandra C *et al*: **The binding interface**

- database (BID): a compilation of amino acid hot spots in protein interfaces.** *Bioinformatics* 2003, **19**(11):1453-1454.
- [21] DeLano WL: **Unraveling hot spots in binding interfaces: progress and challenges.** *Curr Opin Struct Biol* 2002, **12**(1):14-20.
- [22] Gao Y, Wang R, Lai L: **Structure-based method for analyzing protein-protein interfaces.** *J Mol Model* 2004, **10**(1):44-54.
- [23] Guerois R, Nielsen JE, Serrano L: **Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations.** *J Mol Biol* 2002, **320**(2):369-387.
- [24] Kortemme T, Baker D: **A simple physical model for binding energy hot spots in protein-protein complexes.** *Proc Natl Acad Sci U S A* 2002, **99**(22):14116-14121.
- [25] Gonzalez-Ruiz D, Gohlke H: **Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding.** *Curr Med Chem* 2006, **13**(22):2607-2625.
- [26] Huo S, Massova I, Kollman PA: **Computational alanine scanning of the 1:1 human growth hormone-receptor complex.** *J Comput Chem* 2002, **23**(1):15-27.
- [27] Rajamani D, Thiel S, Vajda S, Camacho CJ: **Anchor residues in protein-protein interactions.** *Proc Natl Acad Sci U S A* 2004, **101**(31):11287-11292.
- [28] Hu Z, Ma B, Wolfson H, Nussinov R: **Conservation of polar residues as hot spots at protein interfaces.** *Proteins* 2000, **39**(4):331-342.
- [29] Keskin O, Ma B, Nussinov R: **Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues.** *J Mol Biol* 2005, **345**(5):1281-1294.
- [30] Ma B, Elkayam T, Wolfson H, Nussinov R: **Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces.** *Proc Natl Acad Sci U S A* 2003, **100**(10):5772-5777.

-
- [31] Bogan AA, Thorn KS: **Anatomy of hot spots in protein interfaces.** *J Mol Biol* 1998, **280**(1):1-9.
- [32] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**(1):235-242.
- [33] Henrick K, Thornton JM: **PQS: a protein quaternary structure file server.** *Trends Biochem Sci* 1998, **23**(9):358-361.
- [34] Janin J, Rodier F: **Protein-protein interaction at crystal contacts.** *Proteins* 1995, **23**(4):580-587.
- [35] Janin J: **Specific versus non-specific contacts in protein crystals.** *Nat Struct Biol* 1997, **4**(12):973-974.
- [36] Valdar WS, Thornton JM: **Conservation helps to identify biologically relevant crystal contacts.** *J Mol Biol* 2001, **313**(2):399-416.
- [37] Carugo O, Argos P: **Protein-protein crystal-packing contacts.** *Protein Sci* 1997, **6**(10):2261-2263.
- [38] Shoemaker BA, Panchenko AR, Bryant SH: **Finding biologically relevant protein domain interactions: conserved binding mode analysis.** *Protein Sci* 2006, **15**(2):352-361.
- [39] Zhu H, Domingues FS, Sommer I, Lengauer T: **NOXclass: prediction of protein-protein interaction types.** *BMC Bioinformatics* 2006, **7**:27.
- [40] Davis FP, Sali A: **PIBASE: a comprehensive database of structurally defined protein interfaces.** *Bioinformatics* 2005, **21**(9):1901-1907.
- [41] Gong S, Park C, Choi H, Ko J, Jang I, Lee J, Bolser DM, Oh D, Kim DS, Bhak J: **A protein domain interaction interface database: InterPare.** *BMC Bioinformatics* 2005, **6**:207.

-
- [42] Teyra J, Doms A, Schroeder M, Pisabarro MT: **SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces.** *BMC Bioinformatics* 2006, **7**:104.
- [43] Stein A, Russell RB, Aloy P: **3did: interacting protein domains of known three-dimensional structure.** *Nucleic Acids Res* 2005, **33**(Database issue):D413-417.
- [44] Winter C, Henschel A, Kim WK, Schroeder M: **SCOPPI: a structural classification of protein-protein interfaces.** *Nucleic Acids Res* 2006, **34**(Database issue):D310-314.
- [45] Tsai CJ, Nussinov R: **Hydrophobic folding units at protein-protein interfaces: implications to protein folding and to protein-protein association.** *Protein Sci* 1997, **6**(7):1426-1437.
- [46] Haliloglu T, Keskin O, Ma B, Nussinov R: **How similar are protein folding and protein binding nuclei? Examination of vibrational motions of energy hot spots and conserved residues.** *Biophys J* 2005, **88**(3):1552-1559.
- [47] Chia JM, Kolatkar PR: **Implications for domain fusion protein-protein interactions based on structural information.** *BMC Bioinformatics* 2004, **5**:161.
- [48] Hua S, Guo T, Gough J, Sun Z: **Proteins with class alpha/beta fold have high-level participation in fusion events.** *J Mol Biol* 2002, **320**(4):713-719.
- [49] Yanai I, Derti A, DeLisi C: **Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes.** *Proc Natl Acad Sci U S A* 2001, **98**(14):7940-7945.
- [50] Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**(22):4673-4680.

-
- [51] Shatsky M, Nussinov R, Wolfson HJ: **A method for simultaneous alignment of multiple protein structures.** *Proteins* 2004, **56**(1):143-156.
- [52] Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247**(4):536-540.
- [53] Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers* 1983, **22**(12):2577-2637.
- [54] Hubbard SJ, Thornton JM: **NACCESS.** In. Department of Biochemistry and Molecular Biology, University College, London; 1993.
- [55] Laskowski RA: **SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions.** *J Mol Graph* 1995, **13**(5):323-330, 307-328.
- [56] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13**(11):2498-2504.
- [57] Humphrey W, Dalke A, Schulten K: **VMD - Visual Molecular Dynamics.** *Journal of Molecular Graphics* 1996, **14**:33-38.
- [58] **The Gene Ontology (GO) project in 2006.** *Nucleic Acids Res* 2006, **34**(Database issue):D322-326.
- [59] Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R: **The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.** *Nucleic Acids Res* 2004, **32**(Database issue):D262-266.

VITA

Nurcan Tunçbağ was born in Istanbul, Turkey, on December 9, 1982. She received her B.Sc. Degree in Chemical Engineering from Istanbul Technical University, Istanbul, in 2005. From September 2005 to September 2007 she worked as teaching and research assistant at Koç University, Istanbul, Turkey. She has worked on “Protein – Protein Interfaces and their Applications” and “p53 Pathway Analysis” during her M.S. study. At the time of press, she had a submitted paper, titled “Architectures and Functional Coverage of Protein-Protein Interfaces”. She has attended ISMB’07 (Vienna, Austria) where she has presented a poster about “Protein Interactions by Structural Matching (PRISM)”.

She currently lives in Istanbul, Turkey, and will continue her education with Ph.D. in Computational Science and Engineering at Koç University, Turkey.