DETERMINATION OF PROTEIN-PROTEIN BINDING SITES USING MACHINE LEARNING TOOLS

by

Fidan Sümbül

B.S. in Chemical Engineering, Boğaziçi University, 2006

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Chemical Engineering
Boğaziçi University
2008

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my thesis supervisor, Prof. Türkan Haliloğlu, for her guidance, friendly encouragement and patience throughout the research and writing of this thesis.

I am grateful to Prof. Ethem Alpaydın and Assist. Prof. Elif Özkırımlı for their kind interest, for the time they devoted to reading and commenting on my thesis. Special thanks to Prof. Pemra Doruker for her contribution to enrichment of my knowledge, moral support.

It was great opportunity for me that this work has been supported by the Bogazici University B.A.P. (08A505D).

I gratefully thank to Mehmet Gönen for his kindness, guidance and advices on my work.

Special thanks are due to Seda Aktaş for her friendship, help and constructive suggestions in my study.

I thank to all my friends in Polymer Research Center including Canan Dedeoğlu for their help and attitudes. I also want to thank Esin Yiğit, Sevinç Şimşek and Tülin Keçeli for making sure that they will always be there for me through good and bad days.

I am grateful to my family for their endless love and encouragement. This thesis is dedicated to them.

ABSTRACT

DETERMINATION OF PROTEIN-PROTEIN BINDING SITES USING MACHINE LEARNING TOOLS

Protein-protein interactions are involved in almost all biological processes. Thus, the understanding of the principles underlying these interactions is of great significance. This is mainly to identify the functional sites in proteins and study how proteins function. The whole surface of the protein is not available for interaction with other proteins. There are some distinctive properties that differentiate binding residues from the rest of surface residues. To explore and further to predict the binding interfaces, the present work is composed of two sections. The first part is the identification of differentiating properties for three main groups of residues in a protein, namely, core, binding and non-binding surface residues on a database of 263 proteins. These properties are sequence and structure related characteristics, and as well dynamic peculiarities, of residues such as; the residue propensity, hydrophobicity, side chain polarity and charge, conservation, accessible surface area, and the fluctuations. Some residues prefer being at interface or core rather than the non-interface surface. The hydrophobic residues are favored at interface or in core of the protein. Positively charged polar residues are abundant at interface while the non-polar or polar but neutral ones are mostly found in the core. The interface and core residues have also higher conservation scores. The residues that have higher fluctuations with rest of the residues in the fastest and in the slowest modes by Gaussian Network Model (GNM) are mainly located at interface of proteins. These aforementioned properties are also analyzed in terms of the type of interactions, namely, homogeneous versus heterogeneous complexes and transient versus permanent complexes for a further understanding of the interaction sites. In the second part, these properties are used to predict the binding residues of proteins using support vector machines (SVM) and multiple kernels learning (MKL). Both of these methods are supervised classifier. The maximum accuracy obtained by SVM is 81.3 %, which is the highest observed accuracy in binding site prediction over the literature. The contributions of the grouped properties to the final results are determined by MKL. The type of amino acid, conservation score, accessible surface area

and state of the amino acid (core or surface), relative correlations between fluctuations in both fast and slow modes, and the packing of the residue have the most contribution.

ÖZET

PROTEİN-PROTEİN BAĞLANMA BÖLGELERİNİN MAKİNE ÖĞRENMESİ KULLANILARAK TAHMİNİ

Protein-protein etkileşimi bir çok biyolojik işlemlerde önemli rol oynamaktadır. Bu nedenle etkileşi belirleyen özelliklerin anlaşılması gerek proteinin foksiyonunun belirlenmesi gerekse o proteindeki önemli amino asitlerin belirlenmesi açısından oldukça önemlidir. Proteinler yüzeydeki aminoa citleri aracılığı ile etkileşime geçerler ancak proteini bütün yüzeyi bağlanmaya elverişli değildir. Yüzeyde bulunan bazı bölgeler, yüzeyin geri kalan kısmından farklı bir takip özelliklee sahip olduğu için, protein sadece bu bölgesi aracılığı ile etkileşime girebilmektedir. Bu çalışmada, öncelikle bağlanma yüzeyindeki amino asitleri yüzeyin geri kalan kısmından ayıran özellikler araştırılıp, daha sonra bu özellikler muhtemel bağlanma amino asitlerinin makine öğrenmesi ile tahmininde kullanıldı. Söz konusu özellikler; aminoa sitlerin bulunma sıklıkları, hidrofobisitileri, yan zincirlerinin yüklülük durumu ve yüklü ise yükünün ne olduğu, evrim boyunca korunması, yüzey alanı, hareketliliği ve amino asitlerin salınımlarının birbirleri ile olan korelasyonu. Bu özellikler proteinin üç bölgesi; etkileşim yüzeyi, yüzeyin geri kalan kısmı ve proteinin çekirdeği, açısından incelendiğinde görüldü ki, bazı amino asitler yüzeyde veya çekirdekte olmayı tercih ederken bazıları ise bağlanma bölgesinin dışında kalan yüzeyi tercih etmekteler. Öte yandan etkileşim bölgesindeki amino asitler, yüzeyin geri kalan kısmına göre daha hidrofobik ve evrim boyunca daha çok korunmuş amino asitlerden oluşmakta. Protein kompleksleri 4 gruba ayrılarak homojen ve heterojen kompleksler birbirleri ile ve geçici ve zorunlu kompleksler de kendi aralarında karşılaştırıldı. Daha sonra incelenen bu özellikler kullanılarak birer makine öğrenmesi metodu olan destek vektör makinesi ve çoklu kernel öğrenmesi metodları ile muhtemel bağlanma amino asitleri tahmin edilmeye çalışıldı. Destek vektör makinesi ile mevcut koşullarda ulaşılan maximum doğruluk %81.3 olarak gerçekleşirken çoklu kernel öğrenmesi ile görüldü ki, nihayi sonuca en çok etki eden özellikler, amino asit tipi, korunumu, yüzey alanı ve protein içerisinde sözkonusu amino asitin bulunduğu yer, gerek hızlı gerekse yavaş modlarda amino asitlerin salınımları

arasındaki korelasyon ve sözkonusu amino asitin yakın çevresindeki amino asitlerin ne oldugu bilgisi olduğu görüldü

TABLE OF CONTENTS

ACKNOWLEDGEMENT iii
ABSTRACT iv
ÖZET vi
LIST OF FIGURES x
LIST OF TABLES xiii
LIST OF SYMBOLS/ABBREVIATIONS xv
1. INTRODUCTION 1
2. LITERATURE SURVEY 4
3. METHOD 9
3.1. Definition of Protein-Protein interfaces 9
3.2. Determination of Exposed versus Buried Amino Acids 9
3.3. Calculation of Accessible Surface Area 11
3.4. Calculation of Hydrophobicity 11
3.5. Calculation of Side Chain Polarity and Charge 13
3.6. Calculation of Conservation Scores 14
3.7. In Silico Alanine-Scanning Mutagenesis 15
3.8. The Dynamic Characteristics of the Amino Acids 16
3.8.1. Calculation of the Temperature factor 16
3.8.2. Relative correlations Between Fluctuations 17
3.9. Support Vector Machines Learning Algorithm 18
3.10. Multiple Kernel Learning (MKL) 21
3.11. Dataset 22
4. RESULTS AND DISCUSSION 24
4.1. Organization of Protein-Protein Interfaces 24
4.1.2. Hydrophobicity 26
4.1.3. Side Chain Polarity and Charge 27
4.1.4. Evolutionary Conservation Distribution 28
4.1.5. Temperature Factors (B-factors) 29
4.1.6. Accessible Surface Area Differences between Binding and Nonbinding Residues 30

4.1.7. Alanine Scanning Mutations 31
4.1.8. Relative Correlations between Fluctuations in Fastest Modes 33
4.2. Heterogeneous versus Homogeneous Interactions 35
4.2.1. Residue Propensity of the Homo and Hetero Interface 37
4.2.2. Temperature Factor (B-factor) 38
4.2.3. Accessible Surface Area 39
4.2.4. Properties That Do Not Show Differences between Homo and Hetero Complexes 40
4.3. Transient versus obligatory interactions 41
4.3.1. Residue propensity at obligate and non-obligate interfaces 42
4.3.2. Side chain polarity and charge 43
4.3.3. Hydrophobicity 44
4.3.4. Conservation, temperature factor and accessible surface area 45
4.4. Machine learning results 46
4.4.1. Support vector machines (SVM) results 47
4.4.2. Multiple kernel learning (MKL) 49
4.5. Prediction Case Studies 52
4.5.1. Pyrimidine operon regulatory protein PYRR 53
4.5.2. Glial cell-derived neurotrophic factor 54
4.5.3. Protein (alcohol dehydrogenase) 55
4.5.4. Dengue virus NS3 protease in complex with a Bowman-Birk inhibitor 56
4.5.5. A chimera of beta-catenin and alpha-catenin 57
4.5.6. Rubredoxin 58
4.5.7. Inorganic pyrophosphatase 59
4.5.8. Internalin E-catalin complex 60
4.5.9. Dimeric hemoglobin 61
5. CONCLUSIONS AND FUTURE WORK 62
5.1. Conclusions 62
5.2. Recommendations 64
APPENDIX 66
REFERENCES 69

LIST OF FIGURES

Figure 3.1. Accessible surface area calculations..... 11

Figure 3.2. The linear classification hyperplane and the two different classes 19

Figure 3.3. Linear separation of feature space..... 20

Figure 4.1. The amino acid distribution of the unbound proteins..... 25

Figure 4.2. The difference between core, binding and non-binding surface of the protein in their hydrophobicity values. 26

Figure 4.3. The distribution of polar and charged residues in core, binding and non-binding surface of the protein. Side chain polarity is shown in (a) and the charge is in (b)..... 27

Figure 4.4. Conservation scores of the amino acids. The non-binding surface residues are almost variable although the binding and the core residues are conserved. 28

Figure 4.5. The distribution of B-factors in core, binding and non-binding surface residues of unbound structures..... 29

Figure 4.6. Accessible surface area distribution of binding and non-binding surface residues 30

Figure 4.7. (a) The complex structure of protein 1QA9.pdb (b) Difference in the total energy of the protein with single point mutation of each residue with Alanine. (b) Chain A and the energetic residues. 32

Figure 4.8. (a) Relative fluctuations of the residues, $< \Delta Rij^2 >$, in the fastest modes of the dynamics. (b) The residues that have high correlations in fluctuations

in fastest modes. 34

Figure 4.9. (a) Relative fluctuations of the residues, $< \Delta R_{i1}^2 > < \Delta R_{in}^2 >$ in slowest modes. (b) The complex structure of 1QA9.pdb and anchor or anchoring groove residues. 35

Figure 4.10. The distribution of residue types in homogeneous and heterogeneous interactions 37

Figure 4.11. B factor in homogeneous and heterogeneous interactions 38

Figure 4.12. The distribution of accessible surface area values in homogeneous and heterogeneous interactions 39

Figure 4.13. Indifferent properties between homogeneous and heterogeneous complexes 40

Figure 4.14. Residues in transient versus obligatory interactions 42

Figure 4.15. (a) Side chain polarity (b) Side chain charge in homogeneous and heterogeneous interactions 43

Figure 4.16. Hydrophobicity in transient versus obligatory interactions 44

Figure 4.17. (a) Conservation (b) Mobility (c) Accessible surface area values in transient versus obligatory interactions 45

Figure 4.18. Complex structure and prediction results for 1A4XA.pdb 53

Figure 4.19. Complex structure and prediction results for 1AGQA.pdb 54

Figure 4.20. Complex structure and prediction results of Protein 1B16A.pdb 55

Figure 4.21. Complex structure and prediction results for 1DF9C.pdb 56

Figure 4.22. Complex structure and prediction results for 1DOWA.pdb 57

Figure 4.23. Complex structure and prediction results of 1E5DA.pdb 58

Figure 4.24. Complex structure and prediction results for 1E9GA.pdb 59

Figure 4.25. Complex structure and predicted results of both 1O6SA.pdb and 1O6SB.pdb 60

Figure 4.26. Complex structure and predicted binding residues of 3SDHA.pdb 61

LIST OF TABLES

Table 3.1. Amino acid codes and their maximum accessible (Rost and Sander, 1994). 10

Table 3.2. Hydrophobicity index of amino acids (Kessel *et al*, 2003) 12

Table 3.3. Amino acid side chain polarity and charge 14

Table 4.1. The SVM results obtained by using 20 % of the dataset for training and validation. 47

Table 4.2. The SVM results obtained by using 40 % of the dataset for training and validation. 48

Table 4.3. SVM with polynomial kernel results 48

Table 4.4. SVM with linear kernel results without dynamics of protein 49

Table 4.5. MKL results on residue based grouping 50

Table 4.6. Contribution of residue groups 51

Table 4.7. MKL results on property-based grouping 51

Table 4.8. Contribution of properties 52

Table A.1. Dataset 66

Table A.1. Hetero complexes 67

Table A.1. Homo complexes 67

Table A.1. Transient complexes 68

Table A.1. Permanent complexes 68

LIST OF SYMBOLS/ABBREVIATIONS

R_i	Position vector of i^{th} residues
ΔR_i	Displacement vector for R_i
ΔR_{ij}^2	Correlations between ΔR_i and ΔR_j
u_i	Displacement vector
γ	Force constant
Γ	Kirchoff- connectivity matrix
λ_i	i^{th} eigenvector
K	Kernel
d_k	Coefficient of kernels
3D	Three dimensional
GNM	Gaussian Network Model
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
SVM	Support Vector Machines
MKL	Multiple Kernel Learning

1. INTRODUCTION

Proteins are one of the building stones of the cell. They composed of different combinations of 20 amino acids. Each has its own amino acid sequence and a specific function. The main entity that makes the proteins interesting is their role; they are responsible for many processes in the cell. They transport the information within the cell, catalyze reactions, come together to form new molecules that have a different function, and etc. Although some proteins work individually, many of them need one or more partner to work together for a specific task. To this end, protein-protein interactions are of importance for the survival of the cell. Not all protein-protein interactions are crucial and desirable; some of these interactions are the start instruction of undesirable processes in the cell, such as diseases. This makes the determination of protein-protein interacting residues important also for the drug design.

Proteins interact through their surface residues. These residues are not any of the surface residues but those with some physical and chemical, and also some topological (structural) peculiarities; they are referred to as interface residues. Nevertheless, when the residues at the interfaces are analyzed, it could be seen that they do not contribute the binding interaction in the same way. Some of these residues are more critical for the interaction across the interface; these are named as “binding hotspots”. The rest of them are at the interface as a result of the 3D arrangement of the monomeric chains for the complex structure. There should be a distinguishing property that differentiates the binding residues from the rest of the surface. The identification of binding residues is also of importance in docking processes, which build structural models for protein-protein complexes.

The next question is that, are the interface residues interaction partner specific or not? Some experiments have been performed to answer this question. For example, DeLano et al. (DeLano *et al*., 2000) generated some random peptides and observed the binding of these random peptides to the human immunoglobulin G and further consistently to the same Fc fragment. This implies that the interacting surface can be extracted even

without knowing the binding partner because of the imprinted location of the binding sites of the protein on its structure.

The interface residues can be determined precisely via experimental methods such as X-ray crystallography or NMR. However, these methods are too expensive, hard and require long time. Because of these reasons, the number of currently deposited complex structures on PDB (Bernstein *et al*., 1977) is much lower than the number of monomers. Both the difficulties about the laboratory experimentation and the existence of structural and chemical differences between the binding residues and the rest have given birth to the computer experimentation on trying to predict them.

Although it is known that the identification of protein-protein binding residues can be done computationally, the computational successes have been far from being satisfactory. The prediction accuracy needs to be improved. The aim of this study is to try to predict the binding residues of proteins without knowledge of the binding partner by using machine learning tools.

This thesis mainly consists of two parts. First part is the identification of the chemical or structural properties of amino acids that differentiate the interface residues and the rest of the surface residues in three main groups of interfaces: interfaces including all types of interactions means that; interfaces between the monomers of the same protein called homo complexes, of the different proteins called hetero complexes, interfaces between transiently interacting proteins, and interfaces of obligatory complexes. After that, only the interface residues of homo complexes and the hetero complexes and then the interface residues of transient and obligatory complexes are compared in order to see if there are any differences. After the examination of characteristics of amino acids at interfaces, the research is continued by predicting the binding residues of proteins via machine learning tools using these differentiating properties as features, and then determining the contribution of these differentiating features on binding.

To this aim, the dataset that Porollo and Meller (Porollo and Meller, 2007) compiled from PDB is selected because it covers homogeneous, heterogeneous, transient and obligatory complexes. Interface, core and non interface surface residues are determined

based on the accessible surface areas (ASA) of residues using DSSP (Kabsch and Sander, 1983).

Following, the binding site residues are analyzed based on some chemical properties (hydrophobicity, side chain polarity and charge) and some structural properties (including ASA, mobility, residue preferences based on size). The evolutionary conservation scores of the residues are calculated using Consurf (Landau *et al*., 2005). The mean-square fluctuations as an indicator of the mobility is calculated using Gaussian network model (GNM) (Bahar *et al*., 1997; Haliloğlu *et al*., 1997). The correlation between the fluctuation of residues in fast modes, and also in the slow modes is calculated again using GNM. In silico alanine mutations are carried out to identify energetically hot binding residues (Haliloglu and Ben-Tal, 2008).

The prediction of the binding residues has been done by two different machine learning methods; support vector machines (SVM) and multiple kernel learning (MKL) because this is a classification problem. There are two classes that each residue may choose: being an interface residue or non-interface residue. MKL is an important tool because it also gives the contribution of each separated feature on classification. Knowing the contribution of the separated parts of the dataset to the prediction of binding sites allowed us to determine the importance of the features that tried separately.

2. LITERATURE SURVEY

The information about the function of a protein may very often be written on a small number of residues that are dispersed in primary sequence but cluster in spatial region (Zhang and Grigorov, 2006). One of the most important functions of proteins is to bind to other proteins. Nevertheless, it is not a trivial task to localize the functional interfaces and to elucidate the contribution of each interface residue to binding affinity even if the structure of the protein is available (Lichtarge *et al*., 1996). Thus, the identification of these functionally important regions that may contribute to the understanding of function is of great significance from both fundamental understanding of sequence-structure-function paradigm as well as the practical applications in pharmaceutical designs.

Recently, several studies have attempted to identify protein-protein interaction sites or interfaces. Some studies (Neuvirth *et al*, 2004; Dong *et al,* 2007) have tried to understand the characteristics of protein interfaces that may offer an opportunity to predict the binding sites. Several others have used the structure or/and sequence features of known protein-protein interaction sites. It has been obvious that the prediction requires identification of unique features of the binding sites with respect to evolution, structural and dynamics. Then to rectify good methods that are made use of these unique features will get the success in this area. These studies are briefly described below:

In terms of chemical features, the hydrophobic residues are more abundant at the interface region than the non-binding surface (Jones and Thornton, 1997; Neuvirth *et al*, 2004). Additionally, the interfaces of obligatory complexes were found to be more hydrophobic compared to the transient complexes because a large exposed hydrophobic patch is energetically unfavourable (Jones and Thornton, 1996). In addition to hydrophobic ones, it was found that the polar and aromatic residues are preferred at interface regions (Zhang and Grigorov, 2006; Bradford and Westhead, 2005) especially when the interaction is less permanent (Nooren and Thornton, 2003). The residues containing charged side chains are depleted except arginine (Zhou and Shan, 2001; Dong *et al,* 2007) which is one of the most abundant residues found in the interface regardless of the type of interaction (Dong *et al*, 2007). The reason behind the enrichment of arginine is due to the cation-π

interactions (Crowley and Golovin, 2005). In light of foregoing, in agreement with Thornton (Jones and Thornton, 1996) and Lo Conte et al. (Lo Conte *et al*., 1999), Neuvirth et al. (Neuvirth *et al*, 2004), found that Tyr, Met, Cys and His are the most favoured amino acids and Thr, Pro, Lys, Glu and Ala are the least favoured ones at the interfaces. Additionally, Lo Conte et al. (Lo Conte *et al*., 1999) observed that Arg also prefer to be at interface.

The evolutionary conservation of residues has also been used as a distinctive feature when the structure is known for the prediction of protein-protein interfaces (Neuvirth *et al*, 2004; Bradford and Westhead, 2005; Fariselli *et al,* 2002; Burgoyne and Jackson, 2006; Liang *et al,* 2006; Panchenko *et al,* 2004). Interface residues especially hot spots are more conserved over evolution relative to non-interface surface residues and clusters in space (Zhou and Shan, 2001; Keskin *et al,* 2005). Sequence conservation may arise either for functional meaning that the evolutionary selection at binding sites to maintain the protein function, or for structural that is the selection to hold the stability of the folded state (Lichtarge *et al,* 1996; Chung *et al,* 2006). Many researchers try to overcome this limitation by distinguishing conservation due to the structural and due to the functional reasons (Panchenko *et al,* 2004; Cheng *et al,* 2005). Some studies claims that the conservation of polar residues is more important for identifying hot spots (Hu *et al,* 2000). Keskin et al (Keskin *et al,* 2005) have also showed that there is a good correlation between the structurally conserved residues and the experimental alanine scanning hot spots.

The three-dimensional structure of protein complexes is very important because the features extracted from this structure provide useful information for the understanding of the mechanism of the interaction (Dong *et al,* 2007). One of the most widely used feature extracted from the three dimensional structure is the accessible surface area (ASA) (Zhou and Shan, 2001). Interface residues have higher ASA than non-interface surface residues (Jones and Thornton, 1997; Chen and Zhou, 2005). The interface residues became buried upon complexation whereas the ASA of non-interface surface residues stays unchanged so the latter residues try to maximize the intramolecular interactions that reduce their solvent accessibilities (Zhou and Qin, 2007).

Another feature that shows a different behaviour for the interface and non-interface surface residues is the mobility of the residue. Temperature factor is an indication of the mobility of residues. It has been shown that, the interface residues have lower temperature factor (TF) than the rest of surface residues of the protein in unbound form (Neuvirth *et al*, 2004; Jones and Thornton, 1995). Actually, it is expected that a highly solvent accessible residue has high temperature factor.

The dynamic characteristics of the residues sustain important features although most of the studies have not taken this into account. However, proteins are not rigid bodies. Haliloglu et al. (Haliloglu *et al,* 2005) analyzed the dynamics of the surface residues using Gaussian Network Model (GNM) and observed that, the binding hot spots densely packed at the interface and show high frequency fluctuations with respect to the rest of the surface residues. The protein-protein interfaces especially enzyme-binding sites contain cavities and the residues forming these cavities exhibit high fluctuations in the fast modes, called high frequency vibrating residues (Ertekin *et al,* 2006). Even without any reference to the amino acid preferences or geometrical features, Haliloglu et al (Haliloglu *et al,* 2008), showed that, using only the GNM the binding residues can be predicted in terms of anchor residues and anchoring groove residues by the slowest modes and the fastest modes, respectively. The high frequency vibrating residues in the unbound structure and in complex structure could be referred as hot spot residues (Haliloglu *et al,* 2005; Ertekin *et al*, 2006; Haliloglu *et al*, 2008).

The contribution of the residues at the interface to the binding energy has been studied by alanine scanning experiments (DeLano *et al*., 2000; Clackson and Wells, 1995; Bogan and Thorn, 1998; Thorn and Bogan, 2001). The participation of hot spot residues to the binding free energy in a complex structure is higher than any other residues at the interface. The alanine scanning can also be performed computationally by mimicking the experiments; that is to substitute each residue with alanine in unbound monomer one by one and calculates the change in the energy function that is designed to score the interaction of the residue with others (Haliloglu and Ben-Tal, 2008). Here the premise that the hotspots could be detectable also in the unbound structures. These so called hot spots residues are expected to give a dramatic increase in an energy function.

The properties that have been mentioned so far are distinctive characteristics of interface versus noninterface surface residues but most of them not adequate to precisely determine the interface residues alone. For the accurate determination of interface residues different methods are used to combine distinctive features. Neurvith et al (Neuvirth *et al*, 2004) tried to predict binding residues via a linear combination of (Neuvirth *et al*, 2004) solvent accessibility, evolutionary conservation and hydrophobicity. The advantage of this method is its simplicity but the resulting performance is low. Some researchers modelled the contribution of features via a scoring function (Landau *et al*., 2005; Burgoyne and Jackson, 2006; Liang *et al,* 2006; Van Dijk *et al,* 2004; Murakami and Jones, 2006; Hoskins *et al,* 2006). This is more complicated than the linear combinations but the accuracy is higher. However, this requires physical insight. The last and the most widely used method is the machine learning tools such as support vector machines (SVM) (Bradford and Westhead, 2005; Chung *et al,* 2006; Yan et al, 2004a; Brodner and Abagyan, 2005) and neural networks (Porollo and Meller, 2007; Zhou and Shan, 2001; Fariselli *et al,* 2002; Chen and Zhou, 2005; Ofran and Rost, 2003). The accuracy in machine learning is higher compared to prediction via a linear function or a scoring function, but the transparency in the prediction method is lost. Some researchers introduce a two stage classifier to further improve the performance (Yan *et al,* 2004b). Another new approach to the prediction is the conditional random field (CRF) proposed by Li et al. (Li *et al,* 2007), which assigns the protein sequence a state label. The combination of the results of the different classifier is a recently used method (Qin and Zhou, 2007). The studies performed in this area not only differ in their methods but also in their starting points. Ofran and Rost (2003) used only the information comes from the sequence of a protein, while Zhou and Shan (2001) used the nearest structure neighbour and the information obtained from both sequence and structure of a protein as the features of learning.

The protein-protein interface prediction has to compromise two constraints: coverage, sometimes called recall or sensitivity, and accuracy. The predicted regions should cover as many of the real interface residues as possible, but at the same time as few false positives, that are assigned as interface but in reality they are not, as possible. Since there is an enormous difference between the number of interface residues and the non interface surface residues, a small number of positive predictions can be obtained. There is

no common denominator for the determination of accuracy and coverage in the studies that try to predict the binding residues. In order to overcome this uncertainty, Zhou et al. (Zhou and Qin, 2007) compared the six most widely used web servers; cons-PPISP (Chen and Zhou, 2005), Promate (Neuvirth *et al*, 2004), PINUP (Liang *et al,* 2006), PPI-Pred (Bradford and Westhead, 2005), SPPIDER (Porollo and Meller, 2007), and Meta-PPISP (Qin and Zhou, 2007) on two different datasets. For the 25 CAPRI target proteins the prediction accuracy of the servers ranges from 25% to 31% at coverage of 30% where the interface residues are the 14% of the total surface residues. There are some complications about these servers. For example they only use the surface residues for training and testing their methods. However, in reality a protein consists of core, non-interface surface and interface residues, thus any method should distinguish these three types of residues on their own. Another deficiency of these servers is that, they use small or consisting of only one type of interaction such as transient complexes datasets for training which will affect the prediction results very much.

3. METHOD

3.1. Definition of Protein-Protein interfaces

The protein-protein interface residues have been defined based on the accessible surface area (ASA) change of residues upon complexation in given complex structures; that is the ASA difference between the unbound and bound form of an individual chain (Porollo and Meller, 2007; Jones and Thornton, 1996). In order to determine the accessible surface area of each residue DSSP (Kabsch and Sander, 1983) program is used. First, for each chain, the coordinates are extracted from its complex structure. The main assumption is here that the three dimensional structure of the chains so the coordinates of the atoms do not change upon complex formation or relaxation although this may not be the case. Then the solvent exposure for each amino acid residue of this single unbound chain is computed via DSSP. Again using DSSP the ASA of each residue in the complex structure is calculated. Since the interface residues become buried upon complexation, the ASA of them will change and the solvent exposure of the rest of the residues will stay unchanged. An amino acid is classified as an interface residue if the change in its accessible surface area is greater than 5 $Å^2$ (Porollo and Meller, 2007). In this work, since the coarse grained approach is used, this threshold is reasonable. DSSP program defines secondary structure, geometrical features and solvent exposure of proteins, given atomic coordinates in PDB format Accessible surface area or so called solvent exposure is the number of water molecules possible in contact with the residue. The detailed information about the calculations is given in the corresponding section.

The chains may be involved in multiple interactions via their distinct residues and one can determine all the interface residues regardless of the number of interacting chains by looking at the ASA differences. This is the advantage of this method.

3.2. Determination of Exposed versus Buried Amino Acids

When the three dimensional structure of a protein is considered, there are two states that one residue can stand; core or surface, sometimes called as buried or exposed,

respectively. The determination of whether the residue is exposed or buried, the relative accessible surface area (RelAcc) of the residues are used (Rost and Sander, 1994). The relative solvent accessibility is defined as the ratio of the accessible surface area of the residues, estimated by the DSSP program for the first hydration shell, over the maximum solvent exposure proposed by Rost and Sander, 1994. There is a used for distinguishing the exposed and buried residues in two states as: if the relative accessibility (RelAcc) is less than 16% the residue classified as buried, otherwise the residue is classified as exposed (Rost and Sander, 1994).

Table 3.1. Amino acid codes and their maximum accessible (Rost and Sander, 1994).

Amino Acid	**3-letter code**	**1-letter code**	**MaxAcc***	**Amino Acid**	**3-letter code**	**1-letter code**	**MaxAcc***
Alanine	Ala	A	106	Leucine	Leu	L	164
Arginine	Arg	R	248	Lysine	Lys	K	205
Asparagine	Asn	N	157	Methionine	Met	M	188
Aspartic Acid	Asp	D	163	Phenylalanine	Phe	F	197
Cystine	Cys	C	135	Proline	Pro	P	136
Glutamic Acid	Glu	E	194	Serine	Ser	S	130
Glutamine	Gln	Q	198	Threonine	Thr	T	142
Glycine	Gly	G	84	Tryptophan	Trp	W	227
Histidine	His	H	184	Tyrosine	Tyr	Y	222
Isoleucine	Ile	I	169	Valine	Val	V	142

* Maximum solvent accessibility in $Å^2$

In the present work, the determination of interface residues is carried out before separating the surface residues and core residues. This is as some binding residues, particularly may be referred as especially hot spots. They are found in the cavities at the interface where the ASA of them may be very small and easily be considered as buried. If the separation of interface residues is performed using only the surface residues, these cavities will be missed. In this work these two phenomena is thought as two different features. One feature vector represents the state of a residue that is surface or core, and

another feature vector corresponds to the class of those residues as being an interface or non-interface amino acid.

3.3. Calculation of Accessible Surface Area

The accessible surface area (ASA) is the atomic surface area of a molecule that is accessible to a solvent, and is usually expressed in $Å^2$ (square Ångstroms). ASA is calculated using the 'rolling ball' algorithm, which uses a sphere (representing the solvent) of a particular radius to 'probe' the surface of the molecule made up atomic spheres with their radii of van der Waals radius (Kabsch and Sander, 1983). A typical value of a 'probe radius' is 1.4 Å, which approximates the radius of a water molecule. The accessible surface area is traced out by the centre of the rolling 'solvent' sphere. This surface area different from the one calculated by just using the van der Waals radii of the atoms.

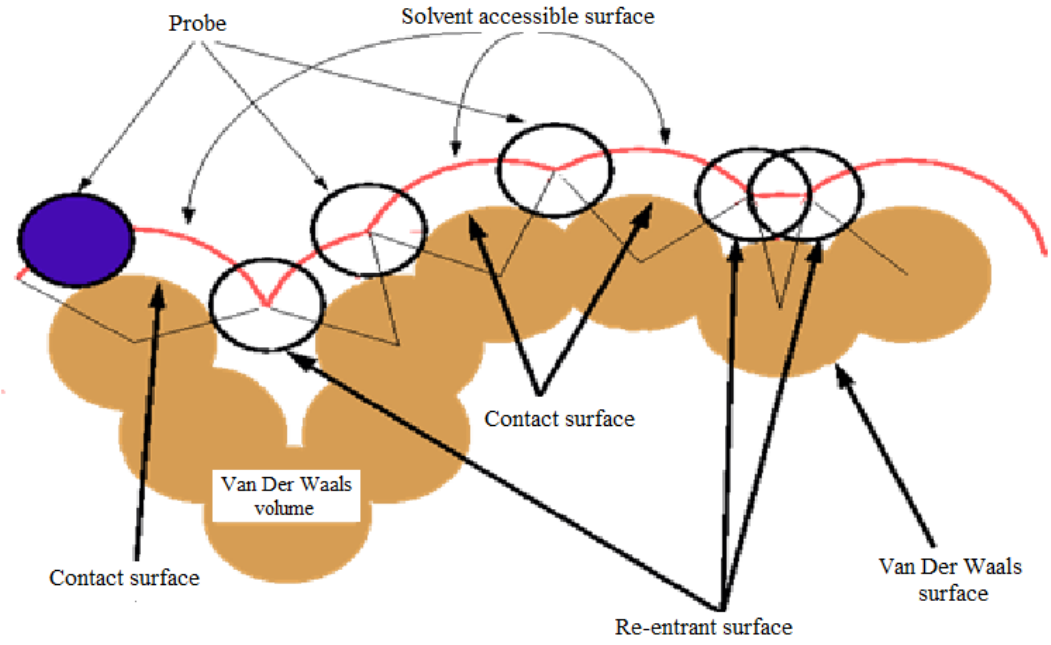


Figure 3.1. Accessible surface area calculations

3.4. Calculation of Hydrophobicity

The hydrophobic effect is believed to have a major role in determining the protein structure. During protein folding, residues with non-polar side chains that are driven from

water are gathered in the protein interior (Rose *et al,* 2004). Thus, the hydrophobic amino acids are likely to be found in the interior, whereas the hydrophilic amino acids are likely to be in contact with the aqueous environment. The hydrophobic forces are also important for protein-protein interactions. In order to quantify this effect, hydrophobicity scales were designed for the amino acids (Wolfenden *et al,* 1981; Janin, 1979).

Hydrophobic amino acids are incapable of forming hydrogen bonds with water as they have no, or very small, electrical charges in their structure. In aqueous solution they disrupt the hydrogen bonding structure which is formed between water molecules themselves. A more stable situation is obtained if the hydrophobic molecules join together and leave the water molecules to their own devices. Effectively this results in a bonding or linkage between hydrophobic materials in aqueous solution.

The majority of hydrophobic amino acids have a side chain which is purely hydrocarbon. They vary in size and, other things being equal, a larger hydrophobic side chain will be more strongly hydrophobic than a smaller one.

The hydrophobicity index is a measure of the relative hydrophobicity, or how soluble an amino acid is in water. In this work, the hydrophobicity index given below is used. The values represent the loss of free energy due to the transfer of the backbone hydrogen bond from water into the membrane (Kessel *et al*, 2003). As positive as the index value, the amino acid is hydrophilic, and as negative as the index value, the amino acid is hydrophobic.

Table 3.2. Hydrophobicity index of amino acids (Kessel *et al*, 2003)

1-Letter amino acid code	A	R	N	D	C	E	Q	G	H	I
Hydrophobicity index	-0,2	19,8	7,7	11,5	0,4	9,5	5,4	0	6,8	-2,6
1-Letter amino acid code	L	K	M	F	P	S	T	W	Y	V
Hydrophobicity index	-2,6	7,4	1,3	-1,5	2,8	0,8	1,1	1,3	4,3	-1,2

3.5. Calculation of Side Chain Polarity and Charge

Each amino acid has at least one amine and one acid functional group as the name implies. The different properties result from variations in the structures of different R groups. The R group is often referred to as the amino acid side chain.

The polar amino acids all contain, in their side chain, an electronegative atom (oxygen, nitrogen or sulphur) which takes on a partial negative charge causing an attached hydrogen atom to be partially positive. This enables the formation of hydrogen bonds with other molecules. On the other hand, amino acids which have side chains with pure hydrocarbon alkyl groups or aromatic are non-polar (Branden and Tooze, 1991).

The polar amino acid group can be further subdivided into those with acidic side chains, which will carry a negative charge, and those with basic side chains, which will be positively charged and neutral side chains.

The side chain polarity and the acidic or basic character is given as two different feature vector in this study. One vector represents the polarity, which is 0 (zero) for non-polar amino acids and 1 (one) for polar amino acids. The second vector consists of -1 for the amino acids that have acidic side chains means that the amino acid structure contains two acid groups and one amine group, +1 for the amino acids contains basic side chains which have an extra amine functional group that produces a basic solution, and 0 for the neutral amino acids that have either equal amount of amine and acid group that neutralize themselves or contains neither of them. The list of these two properties is given below.

Table 3.3. Amino acid side chain polarity and charge

1-letter code	Side chain polarity	Side chain acidity or basicity	1-letter code	Side chain polarity	Side chain acidity or basicity
A	0	0	**L**	0	0
R	1	1	**K**	1	1
N	1	0	**M**	0	0
D	1	-1	**F**	0	0
C	1	0	**P**	0	0
E	1	-1	**S**	1	0
Q	1	0	**T**	1	0
G	0	0	**W**	0	0
H	1	1	**Y**	1	0
I	0	0	**V**	0	0

3.6. Calculation of Conservation Scores

Some amino acids evolve slowly referred as conserved while some others rapidly called variable. This variation of rate corresponds to the selection of amino acids. The conserved amino acids are important either for the folding of the protein to its special 3D structure which are called structurally important residues, or for the interactions of the protein which are called functionally important amino acids. The functionally important residues on the protein surface usually are the active sites of the protein and represent potential binding sites. These functionally important residues are preserved over evolution. If an amino acid is conserved, most probably it plays an important role. The conservation score of the amino acids can be a distinctive feature for the determination of binding sites.

The determination of conservation an amino acid is consists of several steps. First, the close homologous sequences that have known structure are searched for. This is the first and the most important step. If the query sequence does not have enough number of homologues, then the conservation score of it cannot be calculated. PSI-BLAST (Altschul *et al*., 1997) is a tool that used to perform this search. Then the homologous sequences are aligned using an algorithm such as MUSCLE (Edgar, 2004) or CLUSTALW (Thompson *et al*., 1994). The multiple sequence alignment (MSA) file is obtained and then the next step

is to build a phylogenetic tree consistent with this MSA file. The conservation scores can be calculated based on this MSA file.

A web-based tool, Consurf (Glaser *et al*., 2003), which is used for this study, is designed to perform all above calculations and also it calculates the conservation scores. The server leaves the choice of sequence alignment tool and the calculation method of conservation scores. For this research the default algorithms MUSCLE for homologous sequence alignment and the empirical Bayesian to calculate the conservation scores are used. It gives scores to each of the amino acids from 1 to 9 where '1' corresponds to a variable and '9' corresponds to a conserved residue.

3.7. *In Silico* Alanine-Scanning Mutagenesis

The energetic characteristic of each amino acid residue is very important to understand its significance for the protein function or structure. The most accurate way to determine the individual contribution of amino acids to the protein's total free energy is alanine-scanning mutagenesis (DeLano et al., 2000). Alanine-scanning is simply mutating a single residue with Ala and looking for the change in its free energy $\Delta\Delta G$ (Bromberg and Rost, 2008). In silico mutagenesis is very cheap compared to the experimental mutagenesis, yet less accurate.

In the computational approach, each amino acid residue is replaced by alanine and the energy difference in wild type and mutant structure absolute terms is calculated (Haliloglu and Ben-Tal, 2008). While the residue is mutated, the backbone and side chain conformations are assumed to be unchanged. The difference in the energy function reflects the extent of the contribution of this residue to the stability of the structure. To this end, it should be noted that the energy function is low resolution and knowledge based potentials, the details are given in Haliloglu and Ben-Tal (2008).

3.8. The Dynamic Characteristics of the Amino Acids

3.8.1. Calculation of the Temperature factor

The temperature factor sometimes referred as the B-factor is the standard deviation from the mean value in the coordinates of atoms of the amino acids. It is obtained during the refinement of X-Ray crystal structure of proteins. The uncertainty in the position of the atoms in protein crystal increases with the disorder that may be static or dynamic (Parthasarathy and Murthy, 1997). Although the molecule's conformation is stable, some region of it may have different conformations in different copies of the molecule. This is the static disorder. The dynamic disorder comes from the thermal motion, meaning vibration of the molecule from the rest position, of the molecule. Typically, the ends of chains have higher B-factor values, and hence their positions are less certain than the residues in the core of the protein, where disorder is less.

B-factor is calculated from (Kuriyan and Weis, 1991);

$$B = 8 \times \pi^2 \times \langle \Delta r^2 \rangle / 3 \tag{3.1}$$

Noting that;

$$\langle \Delta r^2 \rangle = \langle u_x^2 \rangle + \langle u_y^2 \rangle + \langle u_z^2 \rangle \tag{3.2}$$

Where $u_x,\ u_y,\ u_z$ are the displacements along cartesian coordinates.

The B-factors of the proteins are presented in their PDB files. However, since in this work, the unbound forms of the chains are used and the temperature factors given in the PDB files belong to the complex structures, new mobility values of each amino acids must be calculated. In order to calculate the mobility of the amino acids, the only parameter needed is the mean replacement vector. The replacement vector can be obtained from Gaussian Network Model (GNM), which has been show to be successful in describing the dynamic characteristic of proteins (Bahar *et al*., 1997; Haliloğlu *et al*., 1997) The temperature factor predicted by GNM calculations are in excellent agreement with the values collected from X-ray crystallography (Bahar *et al*., 1997; Haliloğlu *et al*., 1997;

Demirel and Keskin, 2005). GNM has a coarse grained approach based on only alpha carbons (Cα) coordinates. For the GNM calculations,Cα coordinates are extracted from the unbound structure of all chains studied here.

3.8.2. Relative correlations Between Fluctuations

Proteins are not rigid bodies; they have dynamic character especially the surface residues. Fluctuations of the amino acids are calculated using Gaussian Network Model (Bahar et al., 1997; Haliloğlu et al., 1997). GNM is a low resolution network model, which takes Cα coordinates as nodes to represent the residues. The fluctuation of each amino acid from its mean position is modelled by Gaussian distribution. It is assumed that each node is connected to all other nodes within a cut-off distance by elastic springs, and forms an elastic network between contacting residues with harmonic potentials. The fluctuation of a residue is denoted by the vector ΔR and the correlation between the fluctuations of residue i and j is represented by $\langle \Delta R_i \,.\, \Delta R_j \rangle$ and it is obtained from the following expression

$$\langle \Delta R_i \,.\, \Delta R_j \rangle = \frac{3}{2} (\Gamma^{-1})_{ij} = \frac{3}{2} \sum_k \lambda_k^{-1} \, (u_k)_i (u_k)_j \tag{3.3}$$

Noting that, λ_k is the k^{th} nonzero eigenvalue and (u_k) is the corresponding normalized eigenvector, and Γ is the Kirchhhoff connectivity matrix, which describes the dynamic characteristics of the molecule. It is defined as a $n \times n$ matrix for a protein consisting n residues. The Kirchhhoff matrix is constructed using below formula,

$$\Gamma_{ij} = \begin{cases} -\gamma & i \neq j \;\; and \; R_{ij} \;\leq\; r_{cut} \\ 0 & i \neq j \;\; and \; R_{ij} \;>\; r_{cut} \\ -\sum_k \gamma & i \;=\; j \;\neq k \end{cases} \tag{3.4}$$

Here R_{ij} is the distance between residues i and j, r_{cut} is the cut-off distance that is usually taken between 6.5 Å and 7 Å, and γ is the scaling parameter.

The correlation between the fluctuations $\langle \Delta R_{ij}^2 \rangle$ of residues is used to determine the relative fluctuations of the residues between them. The relative fluctuations of residues i and j mean that,

$$\langle \Delta R_{ij}^2 \rangle = \langle \left(\Delta R_i - \Delta R_j \right)^2 \rangle \tag{3.5}$$

$$\langle \Delta R_{ij}^2 \rangle = \langle \Delta R_i^2 \rangle - 2\langle \Delta R_i \, . \Delta R_j \rangle + \langle \Delta R_j^2 \rangle \tag{3.6}$$

The first and the last term in Eq. (3.6) are the fluctuations of the residues i and j respectively, and second term is the correlation between these two fluctuations. If the fluctuations are correlated, the second term will be high and it will decrease the relative fluctuation, and if it is low, means the fluctuations are anticorrelated, the relative fluctuations will increase according to the Eq. (3.6).

The dynamic fluctuations in two ends of dynamic spectrum are of interest in GNM calculations. Fluctuations in fast modes give the high frequency local fluctuations, where fluctuations in slow modes describe global and most cooperative motions (Bahar *et al*., 1998). It was recently been shown that the relative fluctuations in fastest and slowest modes carry information about the binding sites (Haliloglu *et al,* 2008) and these fluctuations can be used to determine the residues involved in binding (Ertekin *et al.*, 2006; Haliloglu *et al*., 2008).

3.9. Support Vector Machines Learning Algorithm

Support vector machines (SVM) is a kernel-based technique that represents a major development in machine learning algorithms. It has received much attention recently, and has been successfully applied to different areas. Support vector machines are a group of supervised learning methods applied to classification and regression (Alpaydin, 2004; Ivanciuc, 2007). This type of learning requires a training data which the class labels are known, so that the rules and other parameters can be optimized (Vapnik, 1998).

The main idea of the SVM in classification is that, it constructs a hyperplane in n-dimensional space that maximizes the margin between the two datasets that belong to two different classes given labels $+1$ and -1. Suppose a training data compose of l samples, $\{x_i , y_i\}$ where $i = 1, \ldots, l, x_i \in R^d$ and $y_i \in \{-1, +1\}$. All the training data will satisfy the following constraints:

$$w \,.\, x_i + b \geq +1 \text{ for } y_i = +1 \tag{3.7}$$

$$w \,.\, x_i + b \leq -1 \text{ for } y_i = -1 \tag{3.8}$$

The set of points x which lie on the hyperplane satisfy,

$$w \,.\, x + b = 0 \tag{3.9}$$

The vector w is the normal vector and it is perpendicular to the hyperplane (Burges, 1998). If the Eq. (3.7) is satisfied then the point belongs to the first class, and otherwise to the second class.

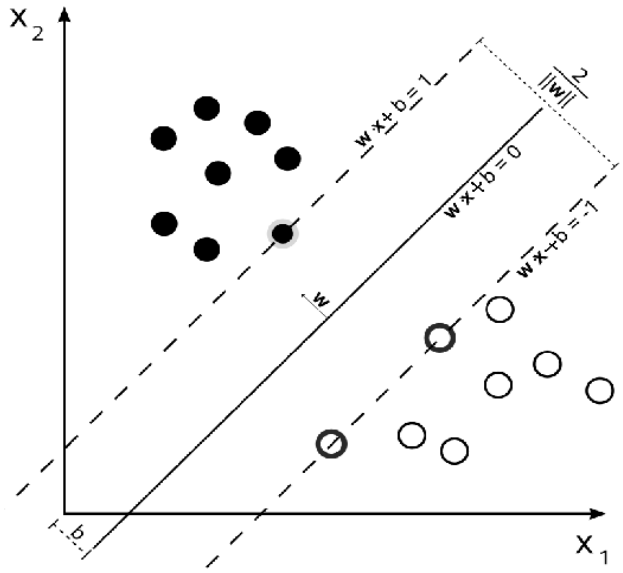


Figure 3.2. The linear classification hyperplane and the two different classes

These samples represented inside circles in Figure 3.2 are called support vectors. The special characteristic of SVM, which makes it attractive, is that the solution to a

classification problem is represented by these support vectors that determine the maximum margin hyperplane. Actually learning with SVM means learning the parameters that define the hyperplane which is w vector and the coefficient b.

SVM can also separate two classes that could not be separated linearly. In such cases the linearly nonseparable data points are projected to a high dimensional feature space that is linearly separable, using nonlinear feature functions ϕ as shown in Fig.3.3 (Ivanciuc, 2007; Burges, 1998).

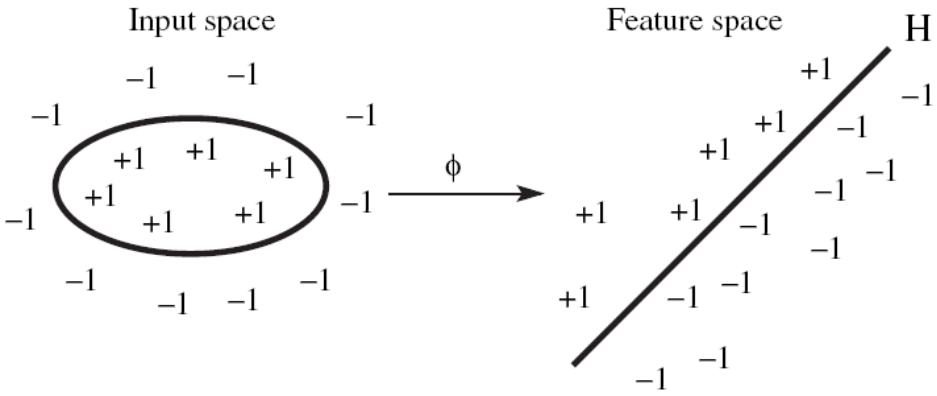


Figure 3.3. Linear separation of feature space

In computing the classification hyperplane it is not practical to use feature vectors, as the feature space is high dimensional. It is computed using special nonlinear functions called kernels. The most commonly used kernels are; linear kernel, polynomial kernel, radial basis functions such as Gaussian and sigmoid (Schölkopf and Smola, 2002). Since the only data that must be kept in memory for classifying new samples are the kernel functions, SVM can handle thousands of features.

Using kernels provides SVM to model complicated data with a small memory usage. However, because there is no theoretical tool to predict which kernel will give the best results for a given dataset, experimenting with different kernels is the only way to identify the best kernel function. There is one parameter that needs to be optimized during validation of the SVM algorithms; the complexity parameter c. In order to determine the classification accuracy of an SVM algorithm, 10-fold cross validation is used in this study. Then the best combination is tested on test data. The accuracy of the method is defined as

the proportion of the truly classified examples to the whole number of examples in the test set.

3.10. Multiple Kernel Learning (MKL)

Multiple kernel learning has recently been a topic of interest. SVM is an efficient tool for learning problems (Bach *et al*., 2004). The performance of the learning has strongly depends on the data representation (Sonnenburg *et al*., 2006). Recent applications and development on SVMs has shown that, using multiple kernels instead of a single kernel improve the performance of the classifier and interpretability of the decision function (Lanckriet *et al*., 2004).

In MKL, the common approach is to consider that the kernel $K(x, x')$ is convex linear combinations of basis kernels:

$$K(x, x') = \sum_{k=1}^{M} d_k \, K_k \, (x, x') \tag{3.10}$$

With $d_k \geq 0$, $\sum_k d_k = 1$.

M is the total number of kernels. The kernels K_k which are combined to form a new kernel can be classical kernels (e.g., linear, polynomial, and Gaussian kernels), the same kernel with different hyperparameters (e.g., degree in polynomial kernel) or kernels over different data representations or different feature subsets (Bach *et al*., 2004). Learning both the coefficients of the kernels (d_k) and the support vector coefficients in a single optimization problem is known as the multiple kernel learning (MKL) problem.

MKL has been used in the field of computational biology. Lanckriet et al (Lanckriet *et al*., 2004) tried to predict the ribosomal and membrane proteins, Sonnenburget al. (Sonnenburg et al., 2006b), use the MKL to find the biologically relevant sequences etc.

Accuracy of both learning methods is based on the confusion matrix as shown in Table 3.1. [Alpaydin, 2004]

Table 3.4. Confusion matrix

	Predicted class	
True class	Yes	No
Yes	TP: True Positive	FN: False Negative
No	FP: False Positive	TN: True Negative

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100 \quad (3.11)$$

3.11. Dataset

The proteins that are used to construct the dataset used for testing and training are that Porollo and Meller (Porollo and Meller, 2007) have used. They selected the proteins that satisfy the following conditions;

- The structures deposited in PDB must be containing at least two chains
- Each chain must contain at least 30 residues.
- Complexes must not contain DNA or RNA sequences
- Any complex must not be in more than 50% sequence identity with the other chains in the set
- Complexes must be real and determined by X-ray crystallography
- The interactions coming from just the crystal packing are excluded.

The resulting number of proteins is 435, consisting of 262 heterocomplexes and 173 homocomplexes.

Some complexes are excluded in the calculations. For example; during conservation score calculations, the number of homologous sequence found were less than 5, which is not an enough number of sequences for the sequence analysis, many complexes have variable side chain conformations and this makes alanine-scanning mutagenesis impossible, some complexes contain undetermined amino acid coordinates. After these

complexes are removed from the list, the remaining dataset consists of 263 complexes for training and testing, where 107 of them are homo complexes and 89 are hetero complexes. In order to analyse transient and obligatory interaction interfaces, NOXclass [Zhu et al. (2006)] a web server is used. The server classifies the interactions according to their interface area, the ratio of the interface area to protein surface area, conservation of the interface, amino acid composition of the interface, the shape of the interface and the correlation between the amino acid composition of the interface and the protein surface by using SVM. Based on the NOXclass's classification, 48 of the complexes are transiently interacting and 88 of them are obligate. The complexes that contain both types of interactions are excluded. 10 complexes used for the prediction case study. The list of these proteins is given in the appendix.

For the learning, a matrix is constructed using all characteristics that are mentioned above. The features defining each instance, where the instances are the residues that are consist of four main groups. One is the residues' characteristics such as ASA, hydrophobicity, side chain polarity and charge, conservation score, B-factor, the maximum relative fluctuations in fast modes, the relative fluctuation with the tails of the protein in slow modes, the difference in the total free energy when mutated with the alanine, a 1-by-20 vector representing the type of amino acid, and the residues' PAM score. The second group of features is consisting of the same properties of the residues included in the sequence sliding window of length 7 (centered at the residue of interest). The third group is the properties of the structure neighbours that are in 10 $\dot{A}$ distances in space that are not sequence neighbour. The fourth and the last group represent the packing of the residue, which means that a 1-by-20 vector each column of it represent a specific amino acid, counts every amino acids found in a 10 $\dot{A}$ shell. For example, the first column of the vector represents Ala, and if there are 3 Ala around the residue of interest, the first element of the vector becomes 3.

4. RESULTS AND DISCUSSION

In the prediction of the protein-protein interaction sites, it is important to understand the mechanism of the interaction. Which features differentiate the amino acids that are at interface from the other amino acids of the protein? The work presented here is divided into two sections. In the first part, the quantitative differences between the protein surfaces involved in protein-protein interactions, the non-interface surface residues and the core residues are analyzed. In the second section, the information obtained from the detailed analysis of protein interfaces is used to predict the binding residues of an unbound protein.

The dataset that is used in this work consist of 263 proteins that contain 59748 amino acids. 12604 of them are interface, 25450 are non-interface surface and 21694 are core residues.

4.1. Organization of Protein-Protein Interfaces

The distinguishing features of protein interface amino acids from the rest of the amino acids are represented in the following figures. The number of amino acids in the three classes, namely, core, binding and non-binding surface, are too different from each other and in order to make a comparison between them, all the frequencies counted for each of the property are normalized by the number of sample in each of the class. The frequencies of the occurrence for the continuous properties such as, temperature factor and accessible surface area, are counted by dividing the values into small intervals.

4.1.1. Residue Preferences of Interfaces

Proteins are different arrangements of 20 amino acids. Each amino acid has its own characteristics, and these characteristics determine the structure of proteins. Some amino acids may be favored in protein interfaces while some others disfavored. In order to understand the amino acid propensities in protein-protein interface the occurrence of each amino acid at protein core, non-interface surface and interface regions are evaluated. The counted occurrences are normalized by the total number of amino acids that found in each

of three regions; this is as the numbers of examples coming from these three classes are not equal. As shown in Figure 4.1 the most abundant amino acids at the interface are Arg, Cys, His, Ile, Leu, Met, Phe, Trp, Tyr and Val. Phe, Ile, Leu, Met and Val are all hydrophobic, and Trp, Tyr and His are polar aromatic residues. These amino acids are also favorable at the core of the protein. Our results for the four amino acids, namely Cys, His, Met and Tyr agree well with the results stated in Jones and Thornton (1996) and Neurvith et al (2004), while our results for all of the amino acids agree with Dong et al (2007) except the amino acid Cys. The difference between these aforementioned studies spotlights the size of the dataset. The datasets that Jones and Thornton (1996) and Neurvith et al (2004), used is small compared to the dataset of Dong et al (2007) and the dataset used for this work.

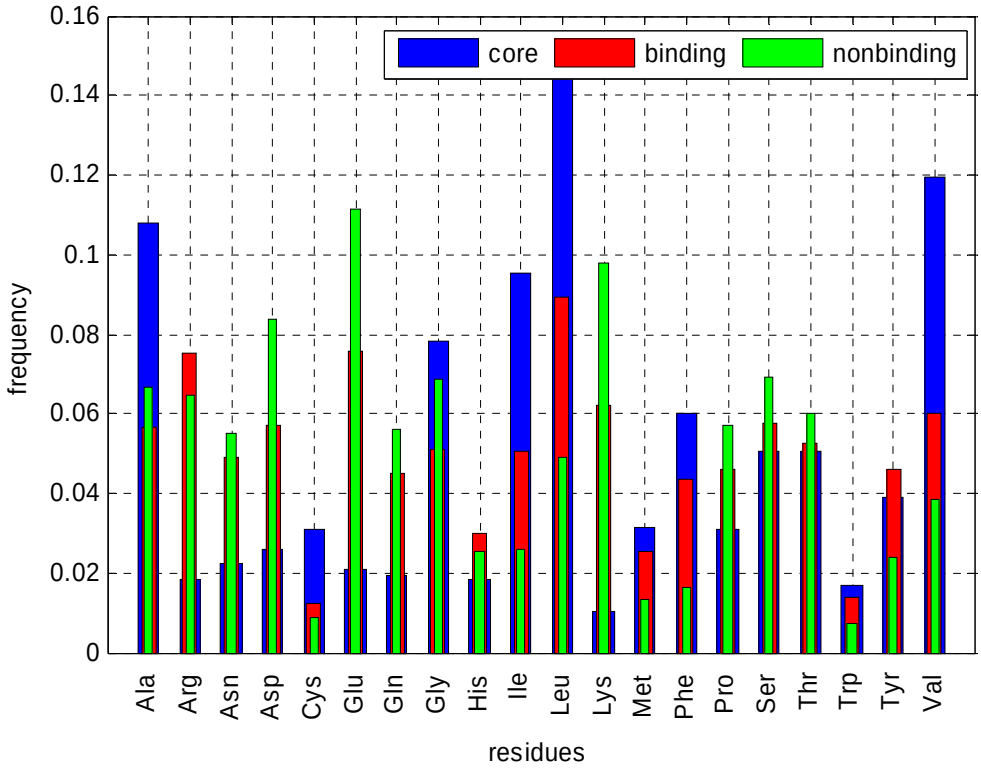


Figure 4.1. The amino acid distribution of the unbound proteins.

Ala, Asn, Asp, Glu, Gln, Lys, Pro, Ser, Thr prefer being at the non-interface surface rather than core or the interface regions. The abundance of Ala, Pro, Thr and Glu is common both Dong et al (2007) and Neurvith et al (2004). Ofran and Rost (2003) also reported the plenty of Lys and Ser at non-interface surface.

Bio-physically similar residues, such as, Ile and Leu or Asp and Glu have the same trend. This is an indication of the reliability of the work presented here

4.1.2. Hydrophobicity

The hydrophobicity values are compared for the core, interface, and the non-interface surface residues of proteins in Figure 4.2 It is assumed that proteins associate with others though hydrophobic patches at their surfaces. As seen, the results obtained by counting the occurrence of the each hydrophobic value in three types of the regions proves this assumption. The core and the interface of the protein are rich in hydrophobic residues while the non-interface surface is not. The hydrophilic ones are most favored at the surface of the protein that is not involved in interaction.

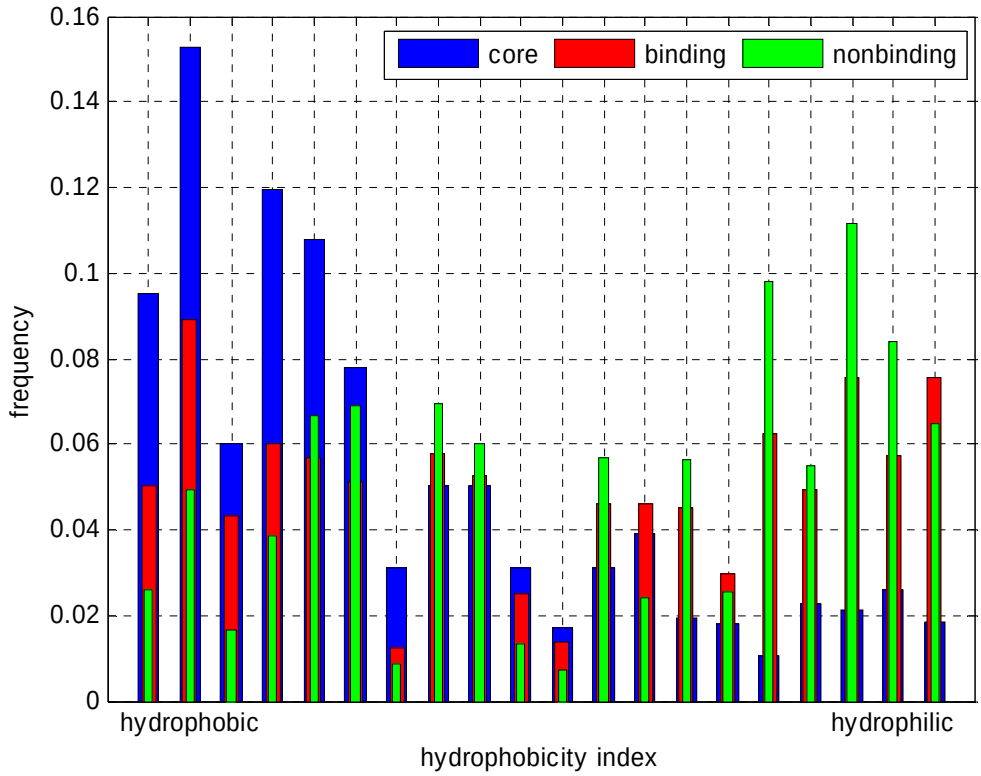


Figure 4.2. The difference between core, binding and non-binding surface of the protein in their hydrophobicity values.

4.1.3. Side Chain Polarity and Charge

The polarity and the charge of the side chains are chemical properties of the amino acids. When Figure 4.3 (a) is analyzed horizontally, the core of the proteins is composed of non-polar residues rather than polar ones, and the opposite situation is true for interface and the non-interface surface residues. Figure 4.3 (b) shows the charge of the polar residues. Polar but neutral residues are more abundant in the core of the protein. The negatively charged polar residues are most favorable at the non-binding surface while the positively charged polar ones are at the binding region of the proteins.

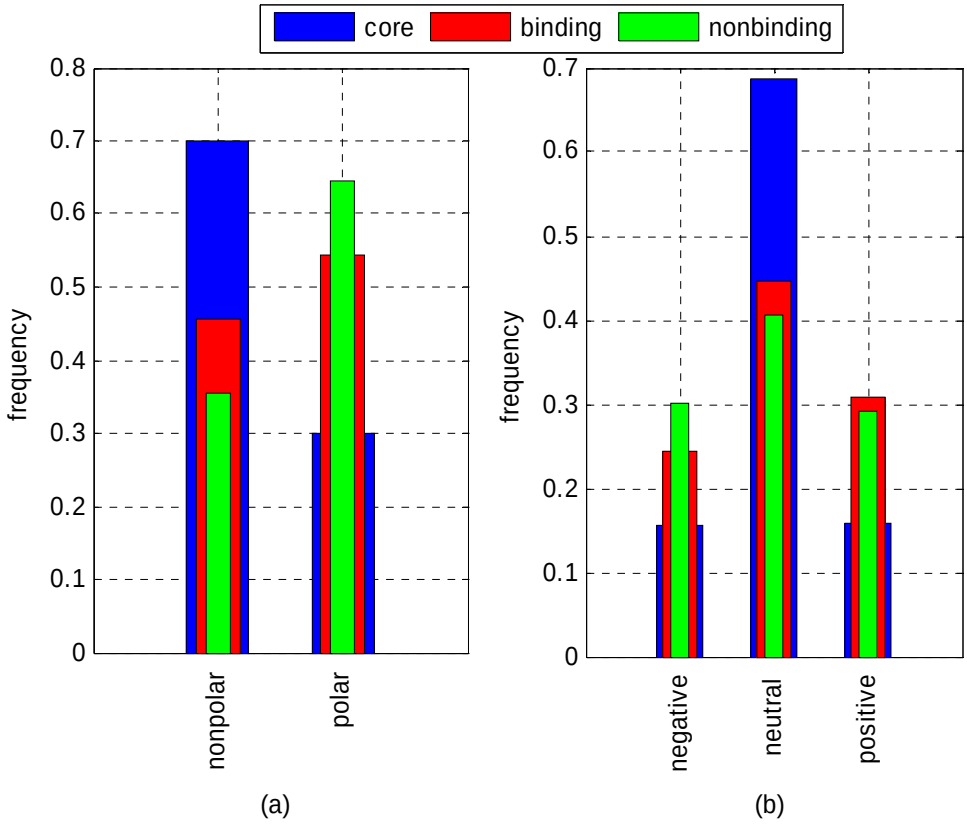


Figure 4.3. The distribution of polar and charged residues in core, binding and non-binding surface of the protein. Side chain polarity is shown in (a) and the charge is in (b).

4.1.4. Evolutionary Conservation Distribution

The conservation of an amino acid over the evolution is an indication of its importance for the protein structure and function. The structurally important residues form the folding core. Since binding is a very significant function of proteins, some of the functionally important ones may form the binding core of the protein. Analyzing the data according to the degrees of conservation of the amino acids displays significant differences for the three group of residues (Figure 4.4). As seen both the core and the binding residues are more conserved compared to the non-interface surface residues.

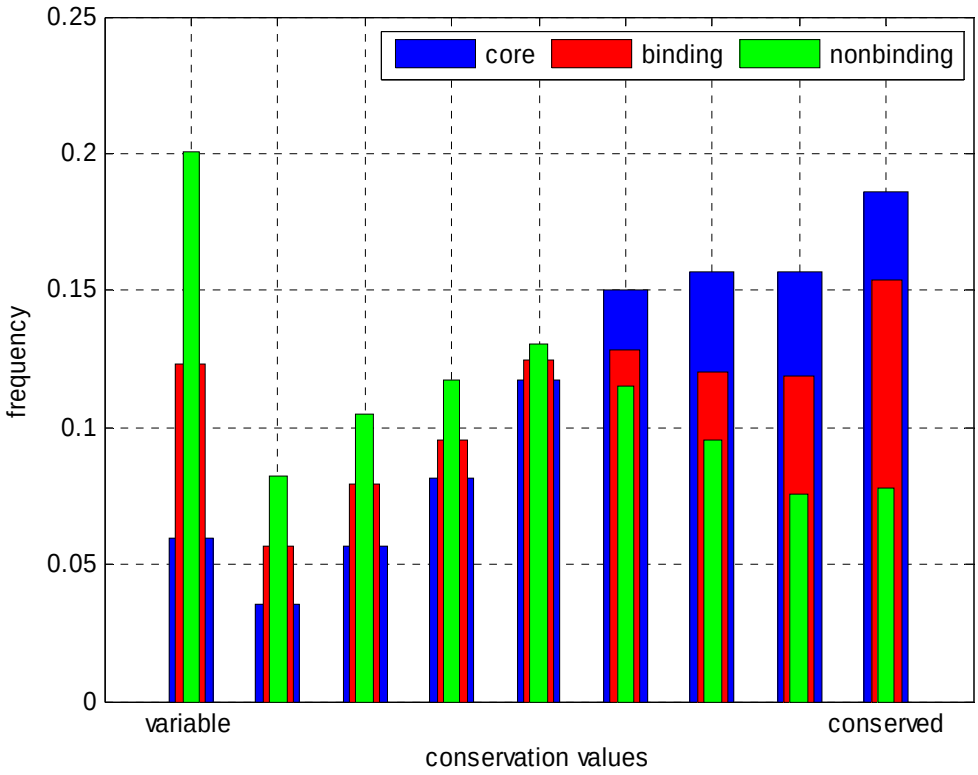


Figure 4.4. Conservation scores of the amino acids. The non-binding surface residues are almost variable although the binding and the core residues are conserved.

4.1.5. Temperature Factors (B-factors)

It is known that the interface residues in general have lower temperature factors (B-factors) compared to the exterior of the protein in a complex structure. The question could be whether this behavior holds also for the unbound state. When the B-factors calculated for unbound proteins by Gaussian Network Model (GNM) is analyzed over the whole dataset, it is observed that the interface residues have lower B-factors already in unbound state although they are exposed forming this state. The core residues again have the lowest scores, as the fluctuations of these residues are restricted by their close neighbors.

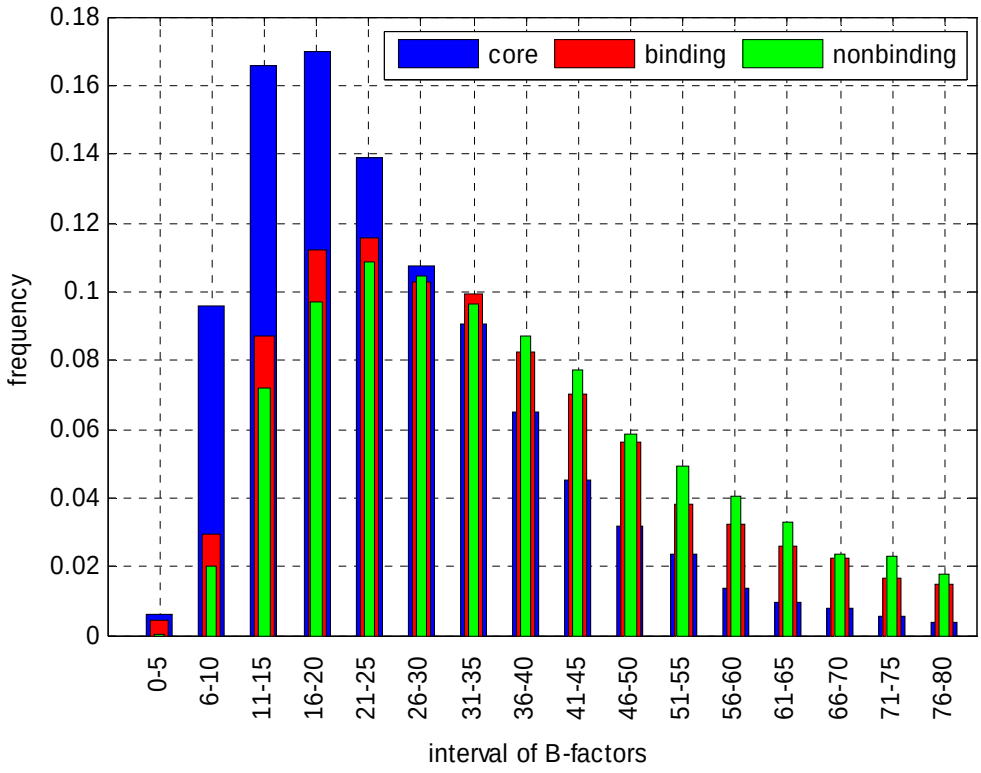


Figure 4.5. The distribution of B-factors in core, binding and non-binding surface residues of unbound structures.

4.1.6. Accessible Surface Area Differences between Binding and Nonbinding Residues

The accessible surface areas of residues are calculated using DSSP program for their unbound states. In Figure 4.6. the distribution of the probabilities of solvent exposure of amino acids at binding interface and non-binding surface is presented in Figure 4.6. It is seen that the binding residues have too low such as less than 20 $Å^2$ or too high such as 110 $Å^2$ surface areas compared to the non-binding surface residues. The low ASA values indicate the residues at the cavities that may be seen as buried according the relative ASA analysis while the high values indicate eaves. This suggests the duality in the structural arrangement of the binding residues. The most interesting outcome of this analysis is the presence of the binding residues at the cavities on the surface.

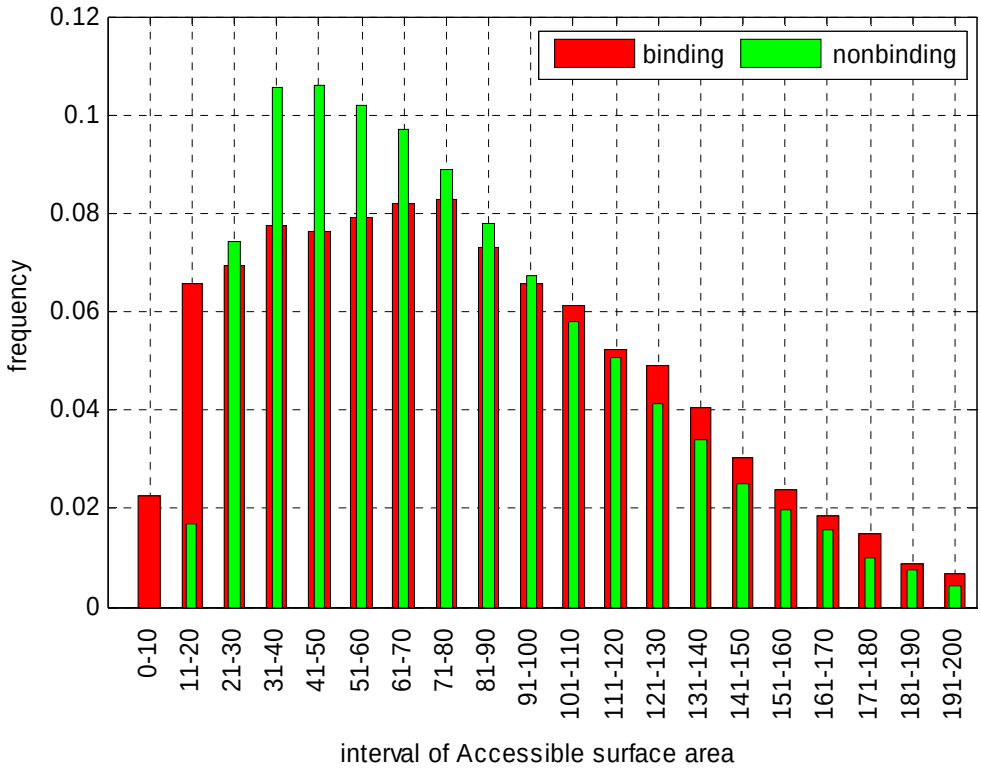
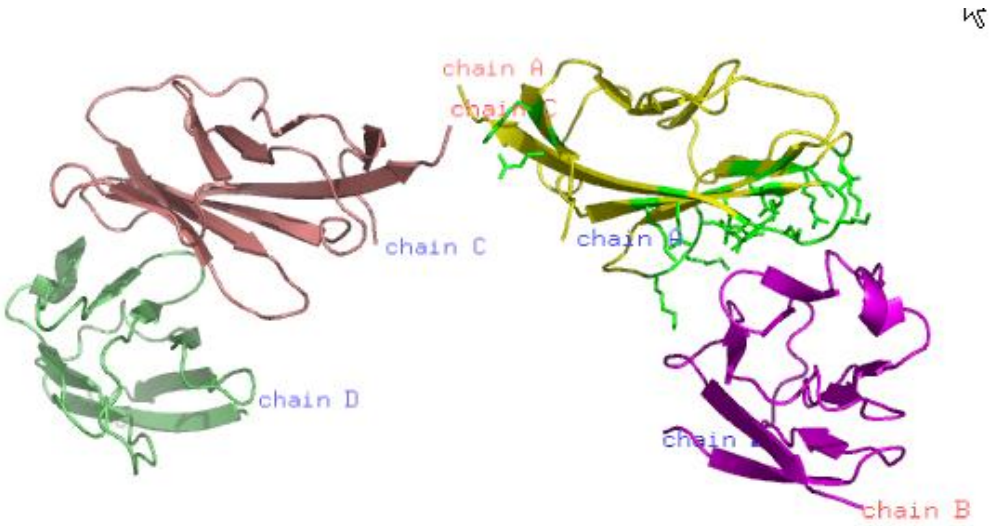


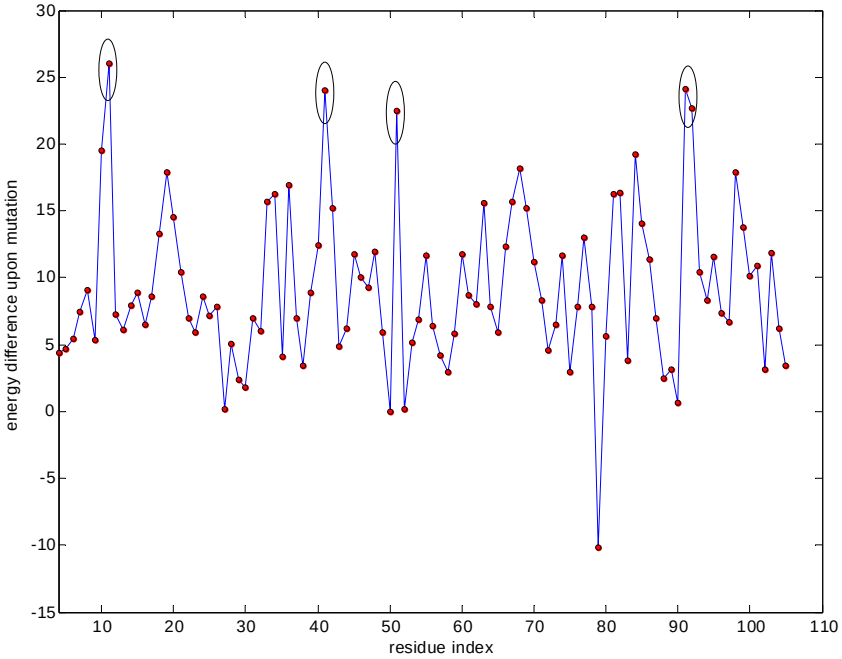
Figure 4.6. Accessible surface area distribution of binding and non-binding surface residues

4.1.7. Alanine Scanning Mutations

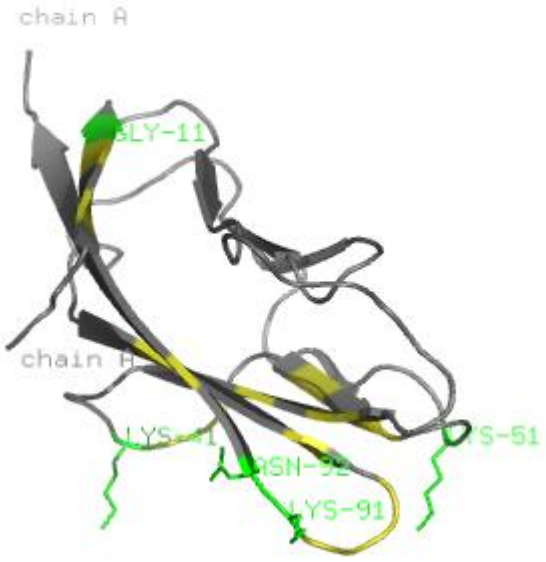
Alanine scanning mutations determine the contribution of each amino acid to the total free energy of the protein. As an illustrative example, the results are presented for protein 1QA9.pdb. The complex structure of this protein is shown in Figure 4.7. (a). Figure 4.7 (b) displays the energy difference obtained with alanine mutations to each residue of the structure (Chain A of 1QA9.pdb). The premise here is as follows: The mutations that lead to a significant change in the protein's energy may point to the structurally and/or functionally important residues. For this particular case, these residues are located at the surface and they are functionally important. Figure 4.7. (c) is the structure of the chin on which the latter residues are highlighted in red. These residues are in general referred as hot spot residues.



(a)



(b)



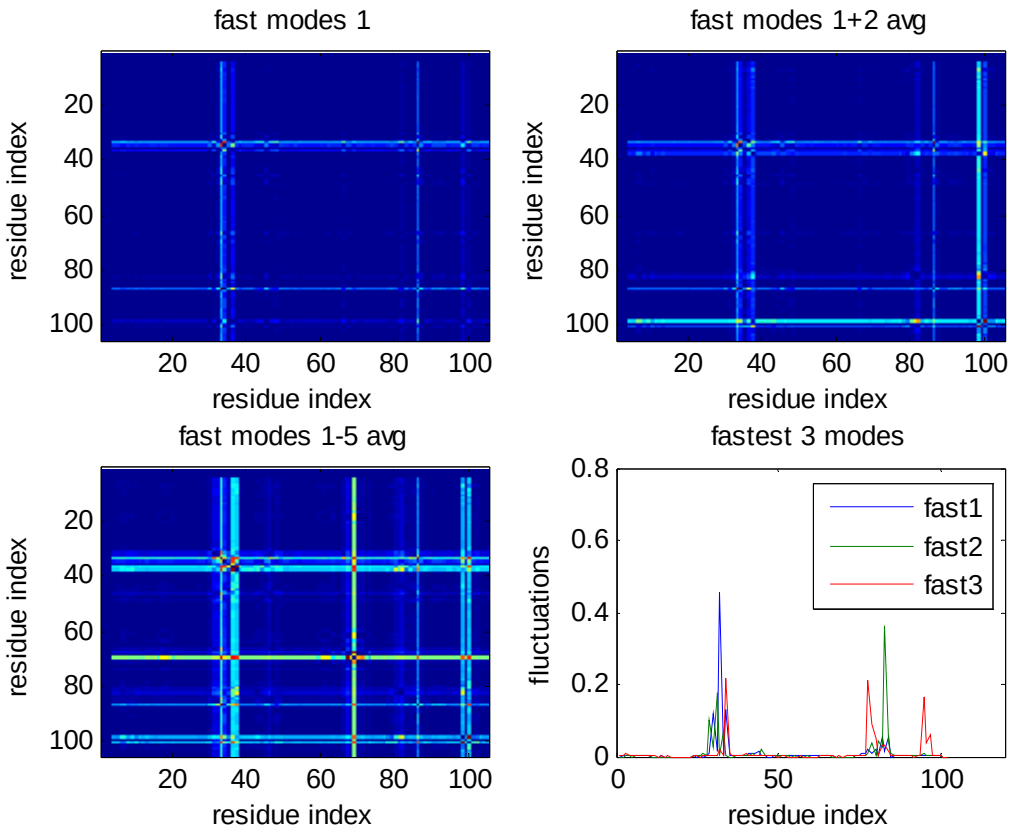
(c)

Figure 4.7. (a) The complex structure of protein 1QA9.pdb (b) Difference in the total energy of the protein with single point mutation of each residue with Alanine. (b) Chain A and the energetic residues.

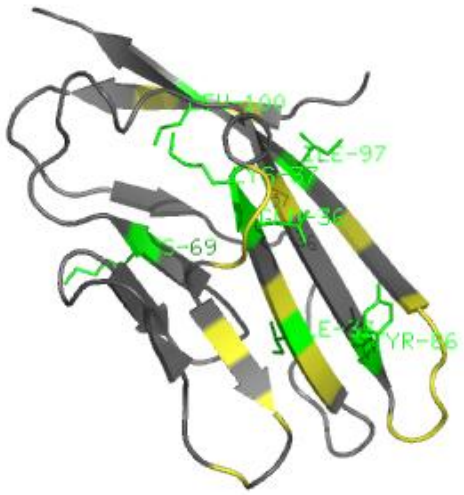
4.1.8. Relative Correlations between Fluctuations in Fastest Modes

The fluctuation behavior of amino acids is an informative property for the significance of the residues in a structure. The relative fluctuations of residues up to five fasted modes for protein 1QA9 are presented in Figure 4.8. (a). The residues that have high relative fluctuations with the rest of residues are identified as: 33, 36, 38, 46, 69, 81, 88, 98, and 100 and displayed in green in Figure 4.8. (b). As GNM is a coarse grained model, these results should be analyzed in a window of 2 or 3 residues. They are mostly located nearby the interface residues or in the core of the structure as shown in Figure 4.8 (b). On the other hand, the relative fluctuations of residues with the C and N termini in the slowest modes can also be used to identify the interface residues.

The correlations of the residues with the first and the last residues of protein 1QA9 are shown in Figure 4.9 (a). Residues 28, 41, and 91 are observed to display anticorrealation with the tail residues. These three residues are anchor or anchoring groove residues are shown in green shown in Figure 4.9. (b)

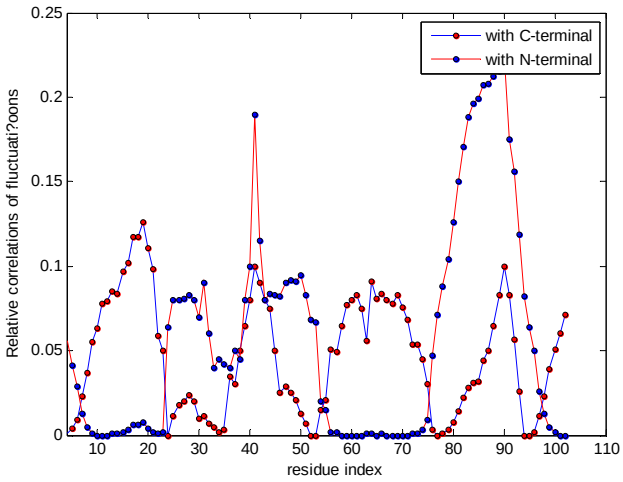


(a)

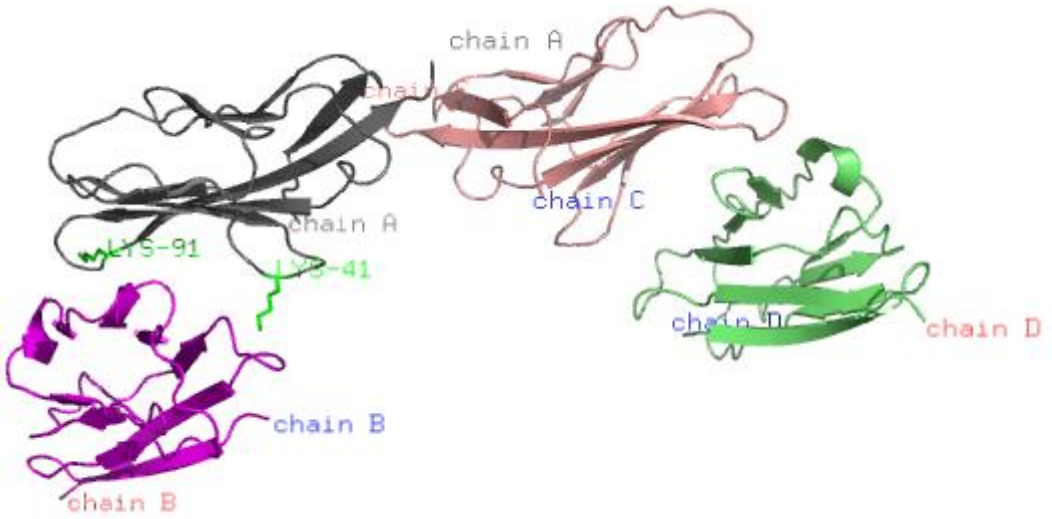


(b)

Figure 4.8. (a) Relative fluctuations of the residues, $< \Delta Rij^2 >$, in the fastest modes of the dynamics. (b) The residues that have high correlations in fluctuations in fastest modes.



(a)



(b)

Figure 4.9. (a) Relative fluctuations of the residues, $< \Delta R_{il}^2 > < \Delta R_{in}^2 >$ in slowest modes. (b) The complex structure of 1QA9.pdb and anchor or anchoring groove residues.

4.2. Heterogeneous versus Homogeneous Interactions

Protein-protein interactions can be classified according to different criterions (Lo Conte *et al.*, 1999). When the interaction occurs between identical protein chains it is called homogeneous interactions, and when two non-identical protein chains interact it is called heterogeneous interactions. In higher oligomers, homo- and hetero-interactions may

occur simultaneously. Heterogeneously interacting parts can be found either as complex or can function individually, while the homogeneous ones are in general as complex. In the present work this classification of interactions is made by using the information given in their PDB file. If a protein assumes a homo-complex with one chain and hetero-complex with another chain, then, it is discarded in the analysis.

Interfaces in heterogeneous complexes and homogeneous complexes are analyzed in terms of the properties that are described in section 4.1. The interfaces exhibit differences for some properties, namely, residue propensities, temperature factors and accessible surface area. The residues at the interfaces of the two types of complexes do not show significant differences are hydrophobicity, conservation, side chain polarity and charge.

4.2.1. Residue Propensity of the Homo and Hetero Interface

The amino acid propensities for interfaces of the two kinds of complexes display some differences. The frequencies of each amino acid at the interface of homo and hetero complex are presented in Figure 4.2.1. Most of the amino acids show insignificant differences or the same trend in both types of the interactions. Asp, Cys, Gln, and Lys show preferences for the hetero complexes while Ala, Gly, Pro and Thr are observed more at the interface of homo complexes.

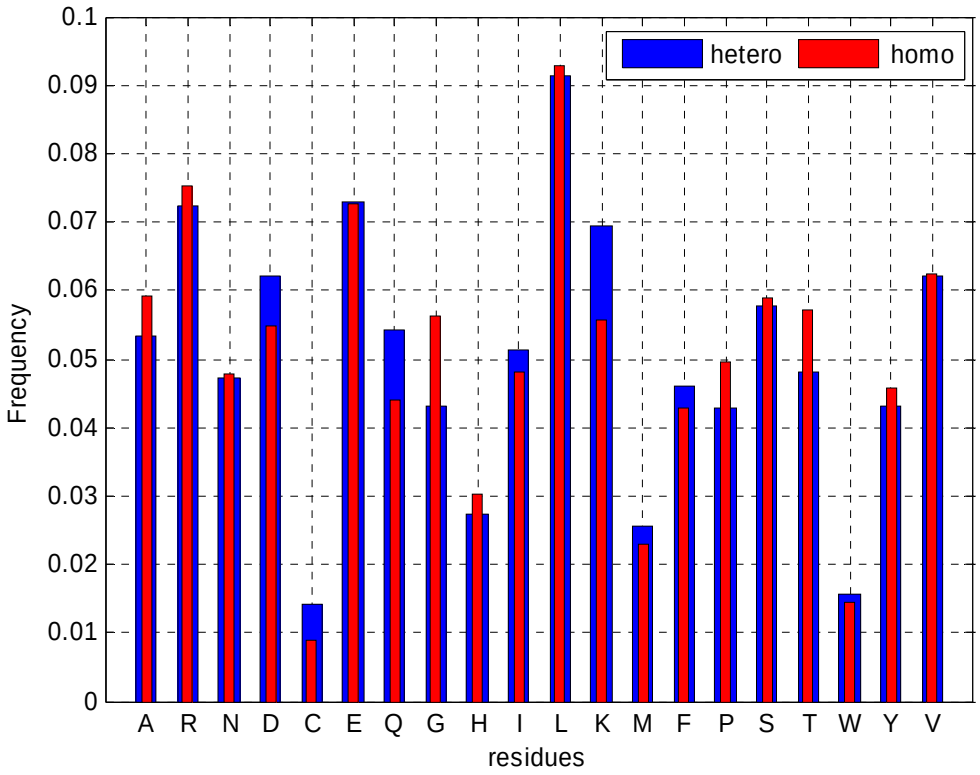


Figure 4.10. The distribution of residue types in homogeneous and heterogeneous interactions

4.2.2. Temperature Factor (B-factor)

The mobility of the residues involved at the interface of homo- and hetero-complexes shows some differences. As shown in Figure 4.2.2, the fluctuations of residues at interface of identical chains cluster at moderate values, while the residues at interface of non-identical chains show either too low or relatively higher mobility. This difference may be emanated from the shape of the interfaces. In heterogeneous complexes the interface region usually resembles lock and key. The residues that fall in the pockets at the interface exhibit low mobility because of the compactness of the structure around it. However, the shape of interfaces for homogeneous complexes is usually flat.

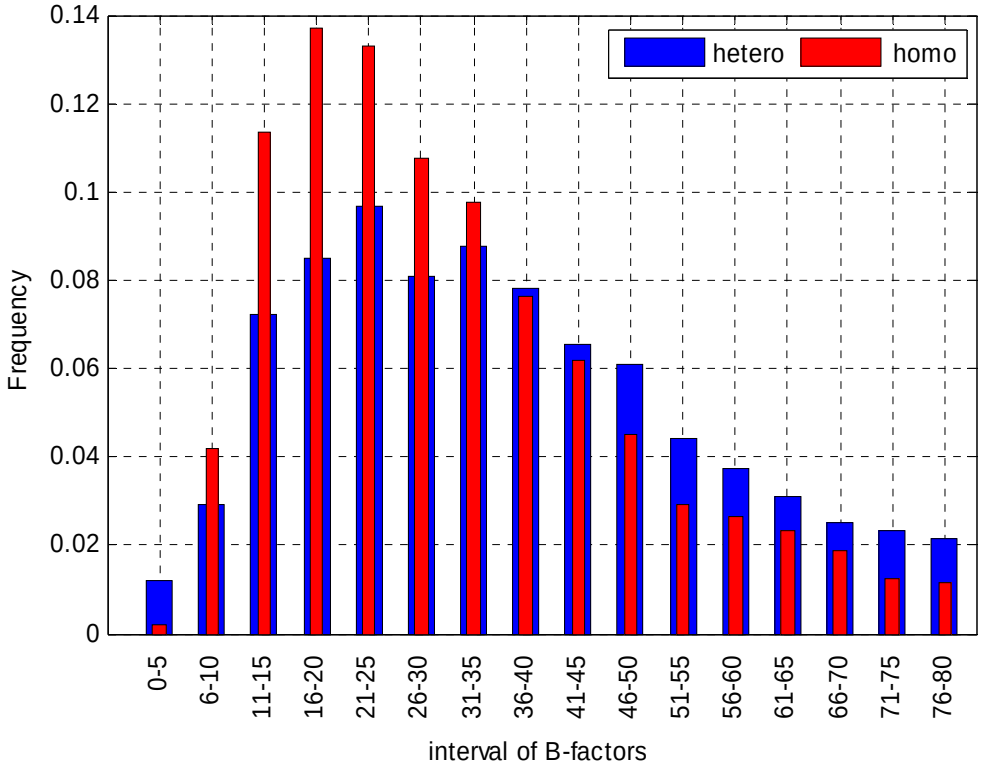


Figure 4.11. B factor in homogeneous and heterogeneous interactions

4.2.3. Accessible Surface Area

The distributions of solvent accessibility of the binding residues that involved in homo- and hetero-complexes are shown in Figure 4.12. The solvent accessibilities of homogeneous interfaces are relatively smaller compared to those of the hetero-complexes; yet the differences are not significant. This result is consisted with the mobility differences of two kinds of interfaces. It is expected that the amino acids with higher accessible surface area will exhibit higher flexibility.

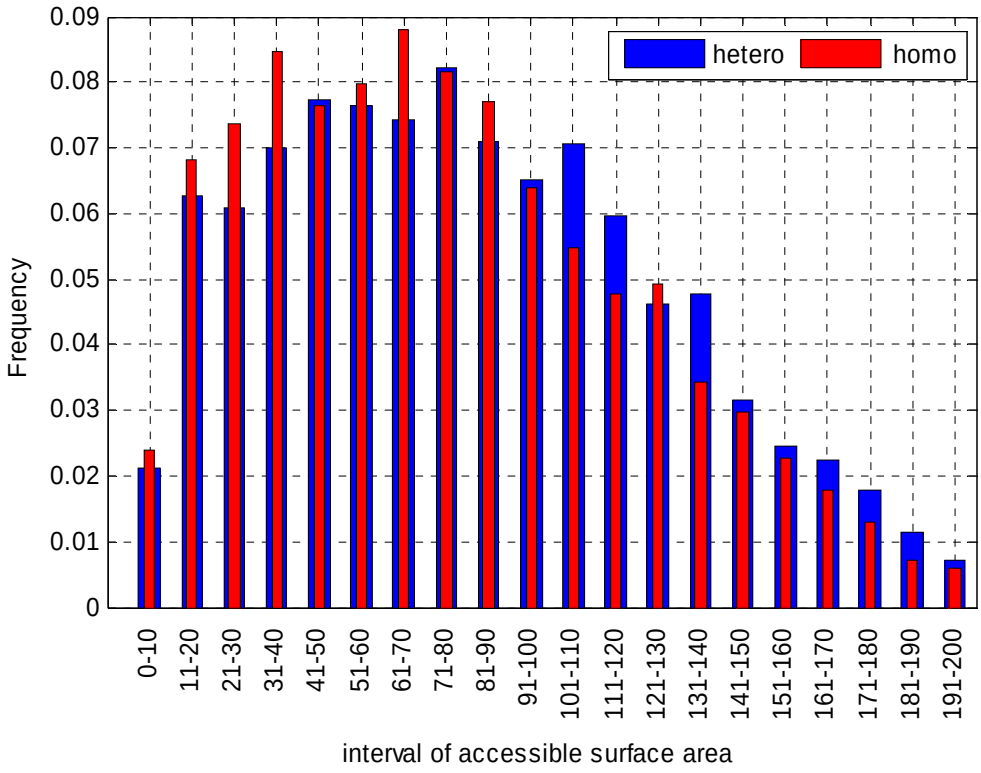


Figure 4.12. The distribution of accessible surface area values in homogeneous and heterogeneous interactions

4.2.4. Properties That Do Not Show Differences between Homo and Hetero Complexes

Chemical properties of the amino acids at homogeneous and heterogeneous interfaces do not exhibit different distributions.

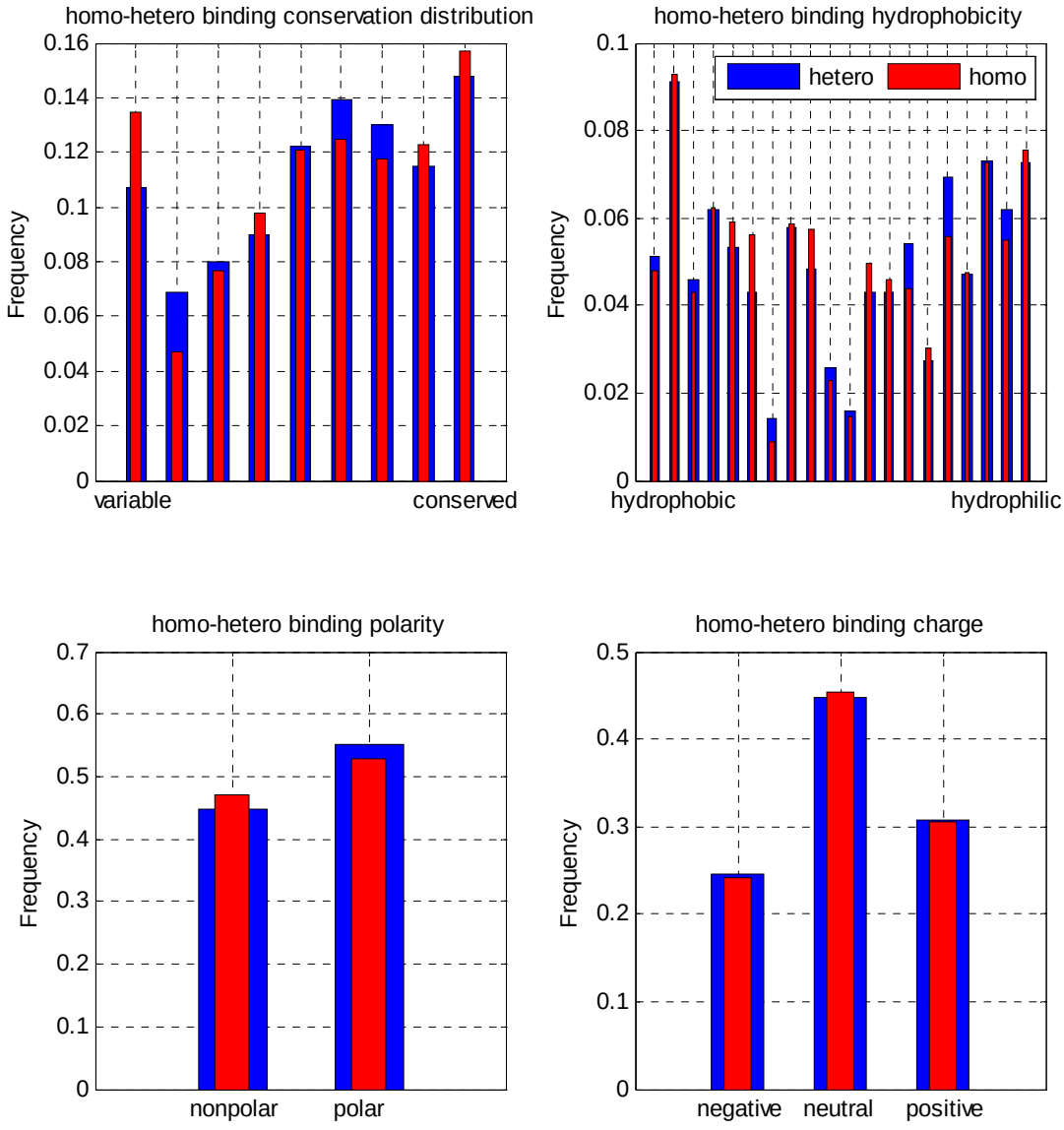


Figure 4.13. Indifferent properties between homogeneous and heterogeneous complexes

Side chain polarity, charge and hydrophobic character of these interactions are almost the same between the two as seen from the Figure 4.13. Due to this, the favored

amino acids in both types could not be grouped. The favored amino acids have different characters such as, Asp, Cys, Gln and Lys are found preferable at heterogeneous interfaces, while Asp is a negatively charged but Lys is a positively charged amino acids, and Cys is an hydrophobic residue but Asp is an hydrophilic one. The evolutionary conservation profiles of residues at both interface types do not show differences.

4.3. Transient versus obligatory interactions

Protein-protein interactions can also be divided into two classes according to the lifetime of the complex, namely, transient and permanent interactions. The permanent interactions are obligatory and usually very stable and thus only exist in the complex state. They are generally also functionally obligate. These permanent interactions are called as obligate interactions throughout this work. On the other hand, transient complexes can associate or dissociate according to the environment or external factors. They can function either as a complex or exist independently. The transient interactions can be weak that is formed and broken continuously or strong which is longer enough to complete a task in the cell. However, distinguishing an obligatory interaction from non-obligate one is not a trivial task. Several studies are proposed to separate these two types of interactions [Jones and Thornton (1997), Nooren and Thornton (2003), Zhu et al. (2006)]. They first have analyzed the distinguishing properties, such as; ratio of the interface area to protein surface area, conservation of the interface, amino acid composition of the interface, shape of the interaction region etc., of obligate and non-obligate interactions and then combined them to determine whether a complex is obligate or non-obligate using different methods.

In the presented work here, in order to separate the interactions whether obligate or transient, a web based server NOXclass [Zhu et al. (2006)] is used. It differentiates the binary interaction of the subunits based on their interface area, the ratio of the interface area to protein surface area, conservation of the interface, amino acid composition of the interface, the shape of the interface and the correlation between the amino acid composition of the interface and the protein surface. It uses a two stage support vector machines (SVM) to separate whether the interaction is obligate or non-obligate or just a result of crystal contact. It calculates the posterior probabilities that the query interaction belongs to obligate or non-obligate interaction types. For this work, the probability higher

than 80 % is taken into account. If a chain makes more than one interactions, when all of them are obligate, the chain is classified as obligatory complex, similar as non-obligate. If a chain exhibits two different characters in two different binary interactions, that chain is discarded.

4.3.1. Residue propensity at obligate and non-obligate interfaces

The interface type preferences of each of 20 amino acids are shown in Figure 4.14. Asn, Ile, Leu, Lys, Phe and Trp are mostly found at obligatory interfaces, whereas Cys, His, Met, and Thr are mostly found at non-obligatory interfaces. The results about the preferences of Ile, Leu, Cys, Met and Thr are consistent with Zhu et al. (2006). Especially the noteworthy differences obtained in the present work and the results obtained by Zhu et al. (2006) for the residues His, Lys, and Phe suggest the importance of the dataset used for training. Biological processes are usually complex enough that they do not be represented by just a hundred of examples. A unique feature about the obligatory interfaces is their size. The permanent interfaces are usually larger as stated by Neurvith et al. (2004).

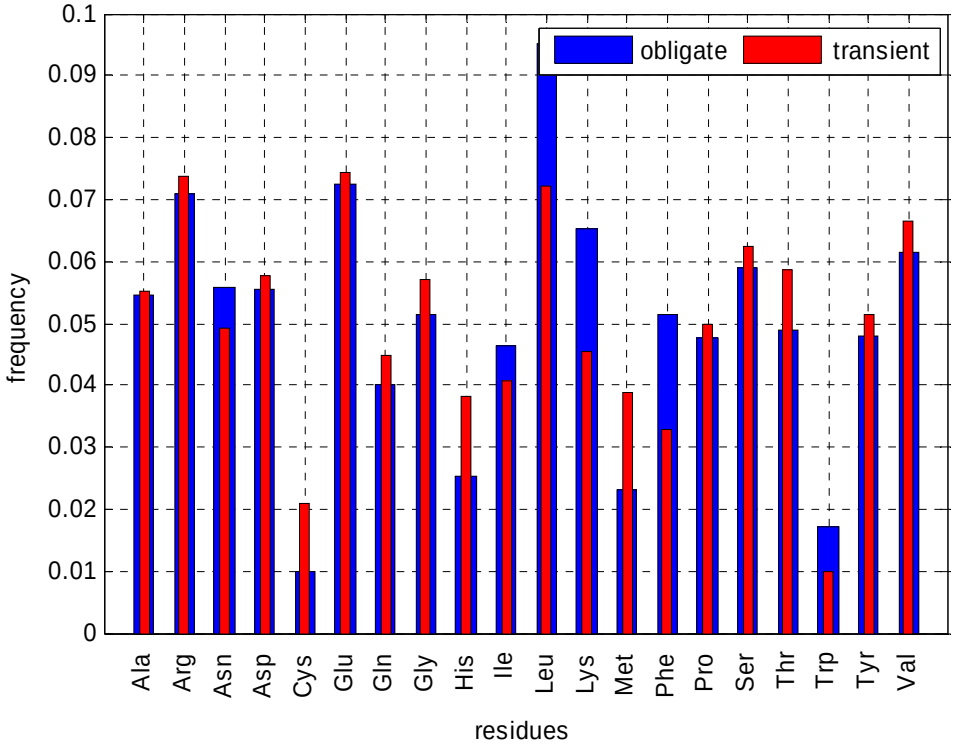


Figure 4.14. Residues in transient versus obligatory interactions

4.3.2. Side chain polarity and charge

Figure 4.15 (a) and (b) displays the difference between the two types of interfaces with respect to the polarity of the residues reflected by their side chains. There are undermined differences between the preference of polar and non-polar amino acids between the two interfaces. Nevertheless, it could be observed here that the polar but the neutral amino acids are abundant at interfaces of non-obligatory complexes while either the positively or the negatively charged polar residues are preferred at the obligate interfaces.

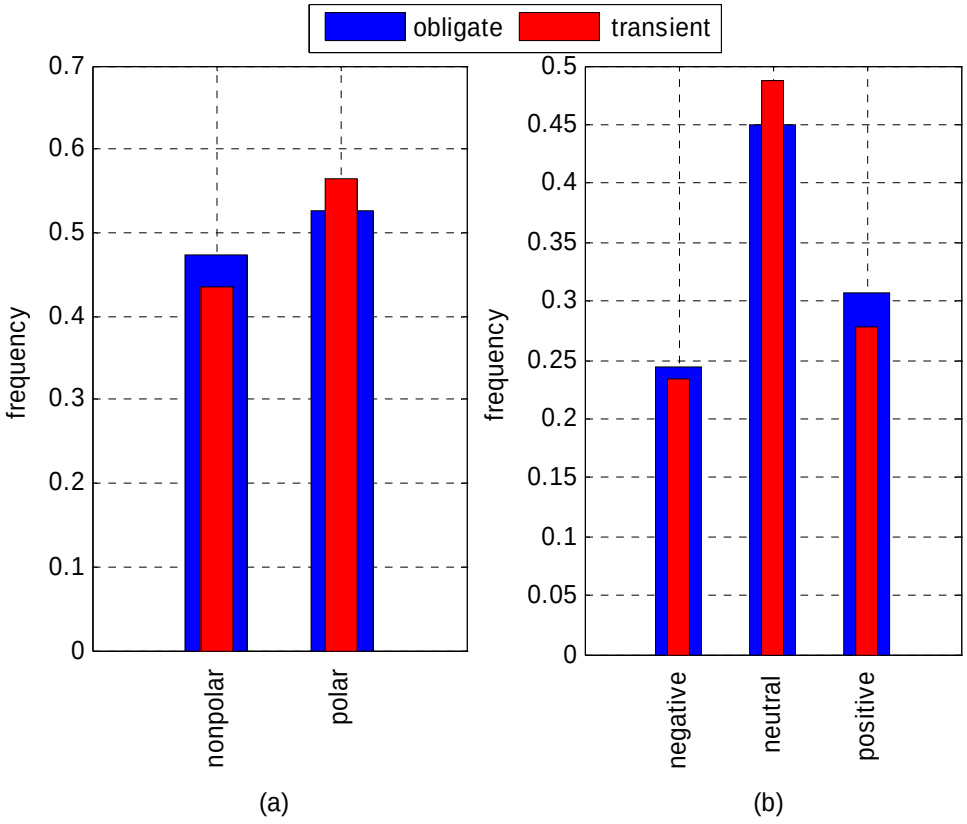


Figure 4.15. (a) Side chain polarity (b) Side chain charge in homogeneous and heterogeneous interactions

4.3.3. Hydrophobicity

Hydrophobicity of the interface residues involved in transient complexes or permanent complexes gives valuable information about the mechanism of the interaction. Figure 4.16 implies that the obligatory complexes are more hydrophobic compared to the transient ones as stated by Keskin et al. (2008) and Neurvith et al. (2004).

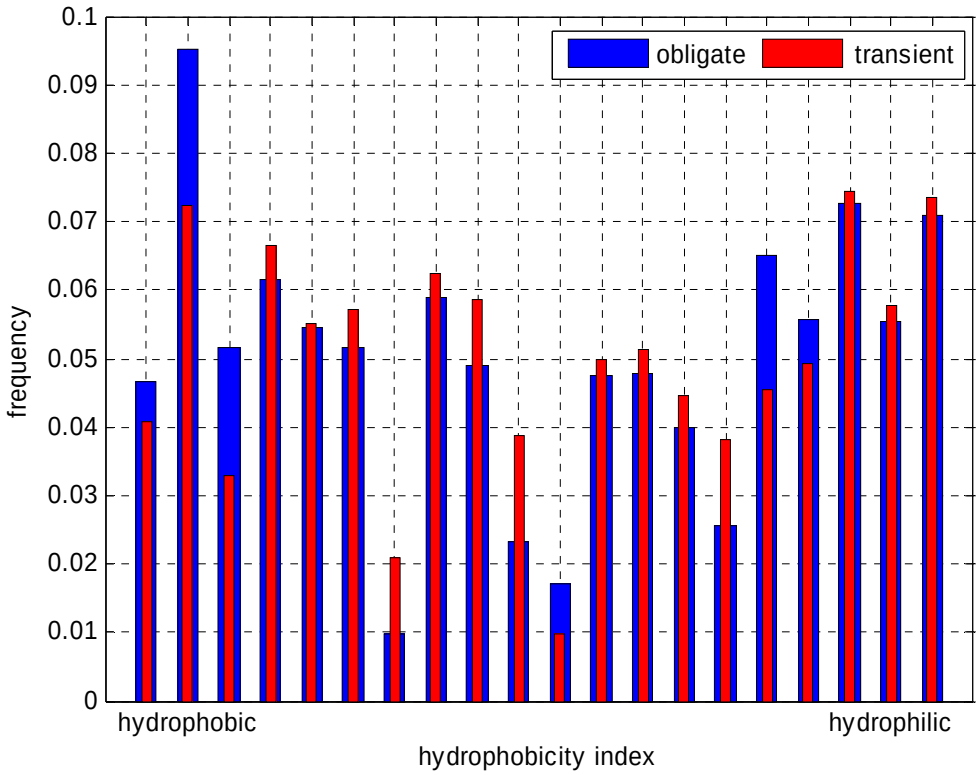


Figure 4.16. Hydrophobicity in transient versus obligatory interactions

4.3.4. Conservation, temperature factor and accessible surface area

The interface residues involved in transient and obligatory interactions exhibit insignificant differences in terms of conservation, mobility and solvent accessibility as illustrated in Figure 4.17.

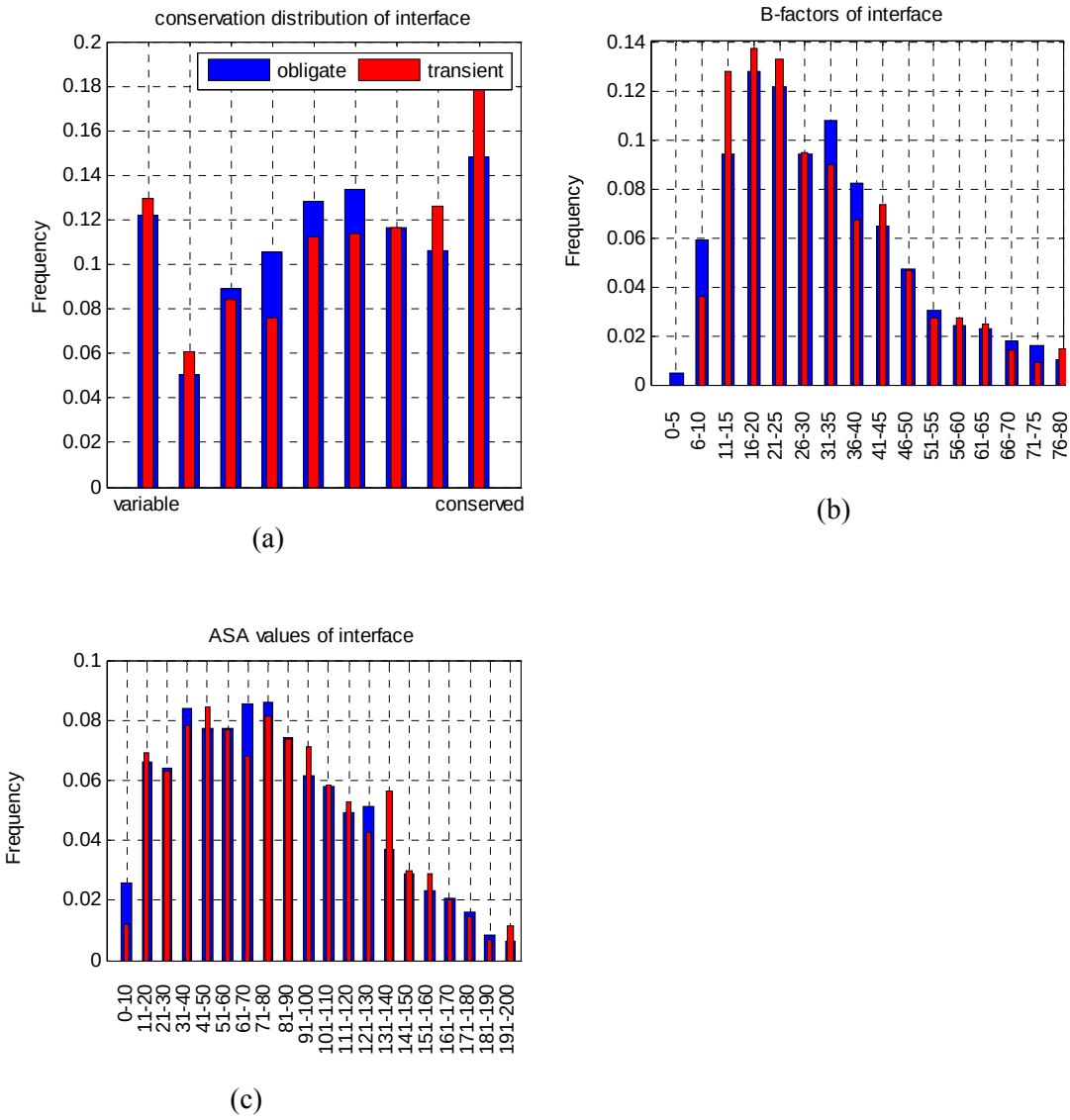


Figure 4.17. (a) Conservation (b) Mobility (c) Accessible surface area values in transient versus obligatory interactions

Figure 4.17(a) shows the conservation score distribution of interface residues for both types. The profiles are very similar, which is not a surprising as the fact that the binding residues are equally important for both. The binding residues are important for obligatory complexes to hold the complex structure as a single structure, and important for transient complexes for the next association. In Figure 4.17. (b) and (c) compares the mobility and the ASA of the interfaces residues, respectively. The latter figures suggest that low ASA and low mobility values are for the residues at obligatory interfaces.

4.4. Machine learning results

Up to here, the differences between the protein surfaces that are involved in protein-protein interactions and the remaining residues of the protein are characterized quantitatively. In this section, these differentiating properties are used in order to predict the location of a protein-protein binding sites on the structure of an unbound protein. For the prediction, two machine learning tools are used, namely, support vector machines (SVM) and multiple kernel learning (MKL).

Complexity is the hyper parameter of the SVM and it should be optimized. SVM is trained at 7 different complexity values such as, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100 and these values are optimized on validation set.

The dataset compiled for the training, validation and test consists of vectors representing the residue, its three sequence neighbors in both side, the six non-bonded spatially nearest neighbors in 10 Å cut-off distance and the aforementioned static and dynamic properties of these thirteen amino acids. There are 59748 instances that 12604 and 47144 of them are binding and nonbinding residues, respectively. First, the dataset is divided into two parts; training and test sets with desired percentages. The test set is randomly divided into ten equal parts. The training part is also divided randomly into two equal subgroups, and one group is used for training the algorithm and the other for validating the learning parameters. This later division is performed five times, and 5×2 times cross validation is performed. Each SVM runs ten times with ten different training, validation and test sets. The average validation and test accuracies and also the support vector percentages with their standard deviation over ten fold are reported for each trial.

4.4.1. Support vector machines (SVM) results

All the properties except the amino acids are numeric values. To represent the amino acids a 1-by-20 vector is used. The columns of the vector from 1 to 20 represents the amino acid types; A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V respectively and can only take value of 0 or 1. When the element takes value 1, it means that the amino acid is the corresponding one. The main reason of this representation is to put the amino acids equally distanced in space. SVM with linear kernel is trained and tested on this dataset. 20 % of the total set is used for training and validating the algorithm.

The vector representation treats the amino acids strictly different from each other, but this is not the case. There are some relations between them, and the scoring matrixes reflect these relations. In the present work, PAM250 (Point Accepted Scoring Matrix) is used. The results of trainings with and without PAM scores are given in Table 4.1.

Table 4.1. The SVM results obtained by using 20 % of the dataset for training and validation.

	Vector representation	PAM and vector
Selected complexity (C)	0.01	0.01
Validation Accuracy (avg. $\pm$ std. dev.)	80.89 ± 0.24	80.92 ± 0.19
Test Accuracy (avg. $\pm$ std. dev.)	80.67 ± 0.10	80.76 ± 0.09
Support Vector % (avg. $\pm$ std. dev.)	42.54 ± 0.36	42.48 ± 0.28

When the percentage of the dataset used for training and validation is increased to 40 % of the whole dataset, an increase in the accuracy is observed. The results are given in Table 4.2.

Table 4.2. The SVM results obtained by using 40 % of the dataset for training and validation.

	Vector representation	PAM and vector
Selected complexity (C)	0.1	0.1
Validation Accuracy (avg. ± std. dev.)	81.11 ± 0.13	81.14 ± 0.12
Test Accuracy (avg. ± std. dev.)	81.19 ± 0.11	81.30 ± 0.06
Support Vector % (avg. ± std. dev.)	41.71 ± 0.16	41.63 ± 0.17

The accuracy of the prediction is increased in both cases when more instances are used at both training and validation. The number of instances used for training and validation could not be increased further because of the limitations about the computer capacity and time at present. The maximum accuracy is obtained when the PAM score and the vector representation are used together.

SVM with 2nd degree polynomial kernel is also trained with best dataset which has the best accuracy and its results are given in Table 4.3.

Table 4.3. SVM with polynomial kernel results

Selected complexity (C)	0.0001
Validation Accuracy (avg. ± std. dev.)	76.99 ± 0.21
Test Accuracy (avg. ± std. dev.)	77.08 ± 0.42
Support Vector % (avg. ± std. dev.)	70.32 ± 0.36

Accuracy of the prediction is decreased when 2nd degree polynomial kernel is used. Linear kernel is better than the polynomial kernel for this problem. The complexity parameter is the smallest one used in this work. The accuracy may increase when smaller complexities are used, but when the complexity is decreased more underfitting will occur.

The main contribution of the present work is to take the dynamics into account in the binding site prediction. The dynamics of residues are calculated for each structure of the dataset by using the Gaussian Network Model. With the knowledge of the dynamics, SVM with linear kernel is trained and tested and the accuracy reached is reported above. To see the improvement rsulting from the usage of dynamics of the structure, the features calculated from GNM are excluded and SVM with linear kernel is trained. Without the contribution of the dynamics, the results are presented in Table 4.4.

Table 4.4. SVM with linear kernel results without dynamics of protein

Selected complexity (C)	1
Validation Accuracy (avg. ± std. dev.)	81.02 ± 0.15
Test Accuracy (avg. ± std. dev.)	80.97± 0.14
Support Vector % (avg. ± std. dev.)	42.39 ± 0.19

As seen, the prediction accuracy is decreased when the dynamic information in terms of the fluctuations and the correlations between fluctuations are removed. This outcome underlies the importance of the dynamics of the residues for determining the binding behavior.

4.4.2. Multiple kernel learning (MKL)

For the classification of residues, and also for determining the contribution of the grouped properties to the prediction, another machine learning tool, multiple kernel

learning (MKL), is used. The residue properties in the dataset are gathered into groups. Two different grouping, namely, residue based and property based are done. Training and validation is done using only the 20 % of the whole dataset in both grouping.

First grouping is based on the residues. Eleven groups are formed. The first group is the properties of the residue of concern. The second group contains the first sequence neighbor of the residue in both side and all the properties of them. The third and the fourth group is formed as the second group y-by taking the second and the third sequence neighbor respectively. Six of the remaining groups contain the properties of the six non-bonded closest neighbors in space and the last group is the packing vector. The results obtained from this trial are given in Table 4.5

Table 4.5. MKL results on residue based grouping

Selected complexity (C)	1
Validation Accuracy (avg. ± std. dev.)	80.46 ± 0.30
Test Accuracy (avg. ± std. dev.)	80.41 ± 0.17
Support Vector % (avg. ± std. dev.)	44.43 ± 0.17

There are small differences in the obtained accuracy between SVM and MKL. Another outcome that obtained from MKL is the contribution of each group to the resultant classification. The contributions of the groups to the final result are given in Table 4.6.

Table 4.6. Contribution of residue groups

Content of the groups		Contribution (avg ± std. dev)
properties of residue of interest		0.17 ± 0.01
packing vector		0.05 ± 0.01
sequence neighbors	primary	0.12 ± 0.01
	secondary	0.10 ± 0.01
	tertiary	0.16 ± 0.02
non-bonded structure neighbors in space	first	0.08 ± 0.01
	second	0.06 ± 0.01
	third	0.06 ± 0.01
	fourth	0.06 ± 0.01
	fifth	0.06 ± 0.01
	sixth	0.07 ± 0.01

The contribution of each of the groups are not lower than 0.05, suggesting contribution from all. The most important group is the residues' own properties. The non-bonded closest neighbors have almost equal contributions although the distances between them are less than 10 Å but different from each other.. . Sequence neighbors from left or right also contribute more or less in the same degree as the non-bonded spatial neighbors.

Another separation of the features into the groups is based on the properties that calculated for each of the residue. The accuracy of this new trial is given in Table 4.7.

Table 4.7. MKL results on property-based grouping

Selected complexity (C)	1
Validation Accuracy (avg. ± std. dev.)	80.81 ± 0.28
Test Accuracy (avg. ± std. dev.)	80.74 ± 0.13
Support Vector % (avg. ± std. dev.)	44.18 ± 0.34

The obtained accuracy is a bit higher than the preceding trial. The participation of each property to the final prediction are listed in Table 4.8.

Table 4.8. Contribution of properties

Content of the group	Contribution (avg $\pm$ std. dev)
Residue vector	0.36 ± 0.02
Residue PAM scores	0.02 ± 0.00
Temperature factors	0.02 ± 0.00
Conservation	0.05 ± 0.00
Hydrophobicity	0.03 ± 0.01
Side chain polarity and charge	0.02 ± 0.01
Accessible surface area and place	0.17 ± 0.01
Relative correlation of fluctuations in fast modes	0.1 ± 0.01
Relative correlation of fluctuations with tails in slow modes	0.15 ± 0.02
Alanine scanning	0.02 ± 0.01
packing	0.05 ± 0.01

The residue vector, conservation scores, accessible surface area and place (core or surface), relative correlation of fluctuations in fast modes, relative correlation of fluctuations with tails in slow modes, and packing vector have considerable contributions. The importance of hydrophobicity is lower than expected. Yet, the hydrophobicity, side chain polarity and charge are indirectly represented by the residue vector. As expected, the information obtained from the GNM has a great support to the final result.

4.5. Prediction Case Studies

In order to see the predictions of the algorithm on the structures, a dataset composed of randomly chosen 10 protein chains that are not used before is constructed and the resultant predicted binding residues are shown below.

4.5.1. Pyrimidine operon regulatory protein PYRR

Pyrimidine operon regulatory protein PYRR (PDB code: 1A4X) is a homo dimer. Chain A of this protein is used in the dataset and it contains 177 amino acids; 19 of them are defined as interface according to the ASA differences. In Figure 4.18. the SVM prediction results for 1A4X.pdb are shown. The predicted binding residue is in green and the interface residues obtained from accessible surface area difference are in yellow including the green one. The residue in green is an interface residue so it is a true-positive example. No misclassified examples for this protein.

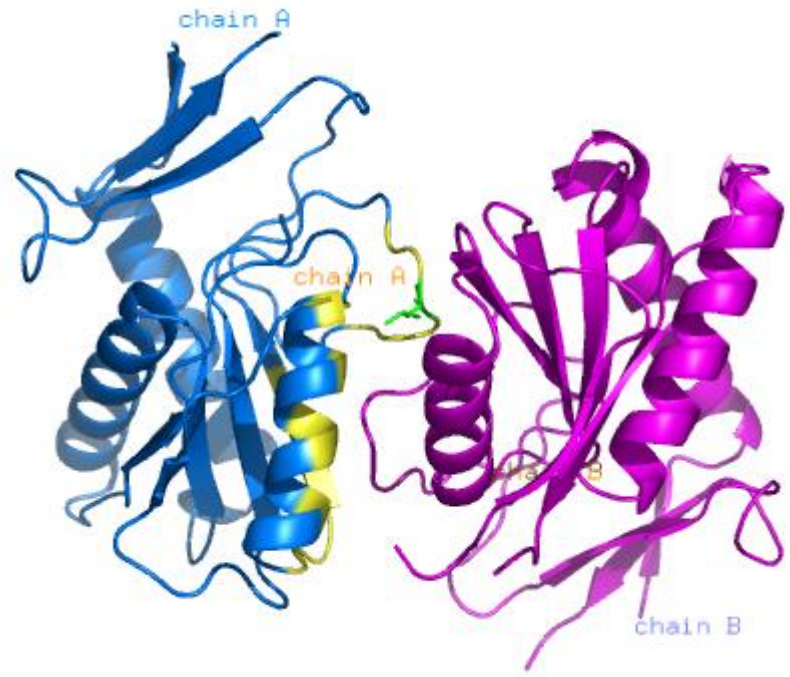


Figure 4.18. Complex structure and prediction results for 1A4XA.pdb

4.5.2. Glial cell-derived neurotrophic factor

Glial cell-derived neurotrophic factor (PDB code: 1AGQ) has four chains and chain A is used. Chain A has 85 amino acids and 34 of them classified as interface residue according to the ASA difference. The 17 residues in green are true classified binding residues. There are 18 residues that are classified as non-binding but the method predict them as binding and 11 residues classified as binding but SVM predicts them as non-binding. These misclassified residues are in red Figure 4.19.

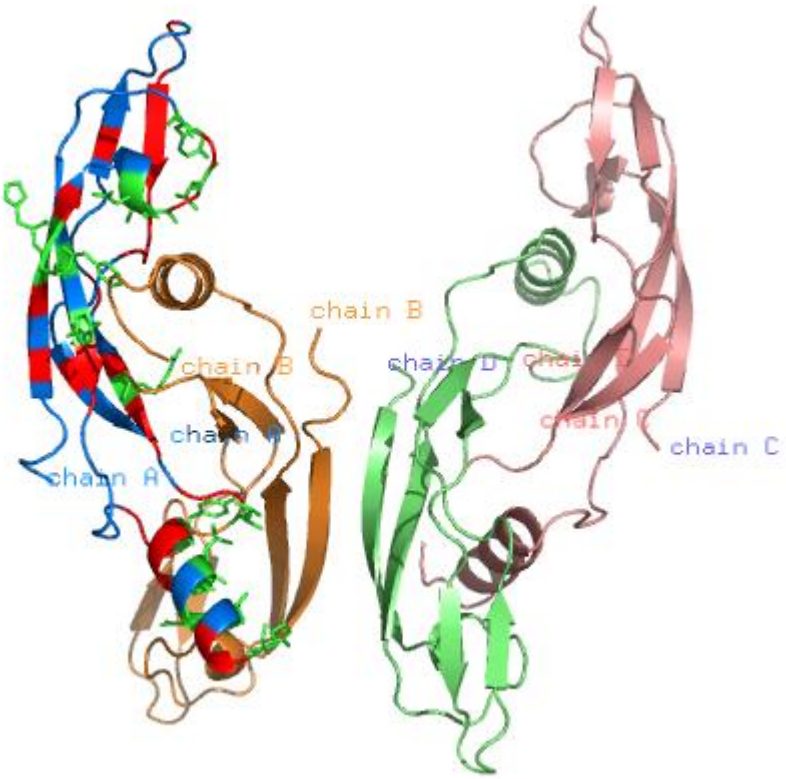


Figure 4.19. Complex structure and prediction results for 1AGQA.pdb

4.5.3. Protein (alcohol dehydrogenase)

The PDB id of this protein is 1B16 and it is homo dimer. Chain A of this protein is used and this chain contains 248 amino acids. The 10 residues in green are true classified binding residues shown in Figure 4.20. The rest of the interface residues that determined from the accessible surface area difference are in yellow and the number of this type of misclassified examples are 41. The number of misclassified non-binding residues is 2 and shown in red.

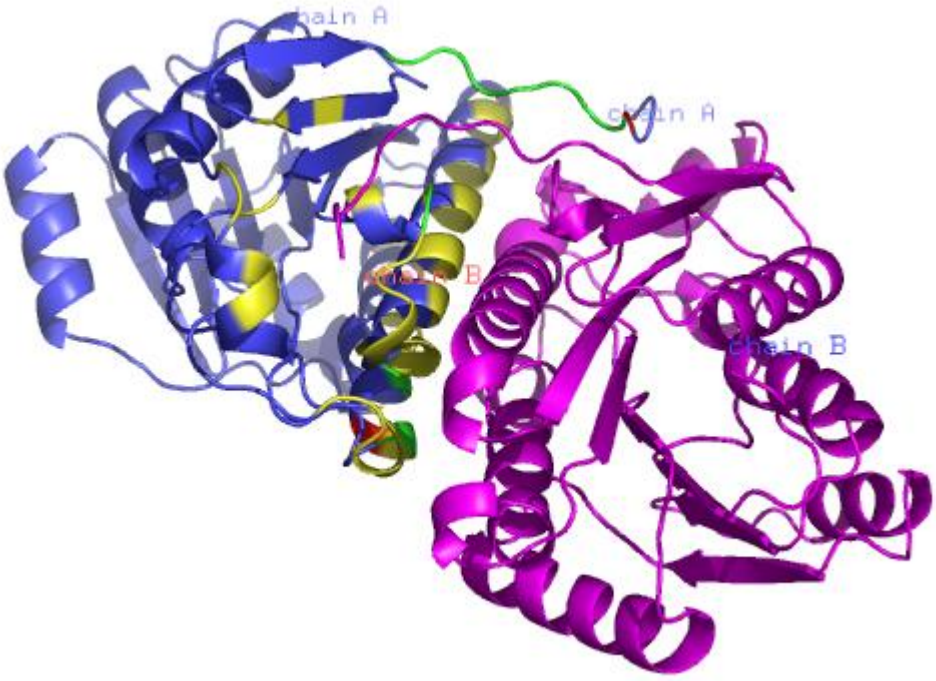


Figure 4.20. Complex structure and prediction results of Protein 1B16A.pdb

4.5.4. Dengue virus NS3 protease in complex with a Bowman-Birk inhibitor

Dengue virus NS3 protease in complex with a Bowman-Birk inhibitor (PDB code: 1DF9) has three chains. Chain C which is the Bowman-Birk inhibitor is used for the prediction. Chain C contains 45 residues and 24of them classified as interface residues according to ASA differences. SVM predict 2 of these surface residues as binding residue and these are displayed in green in Figure 4.21. There is no misclassified non-binding residue in this case but 21 misclassified binding residues are shown in yellow in Figure 4.21.

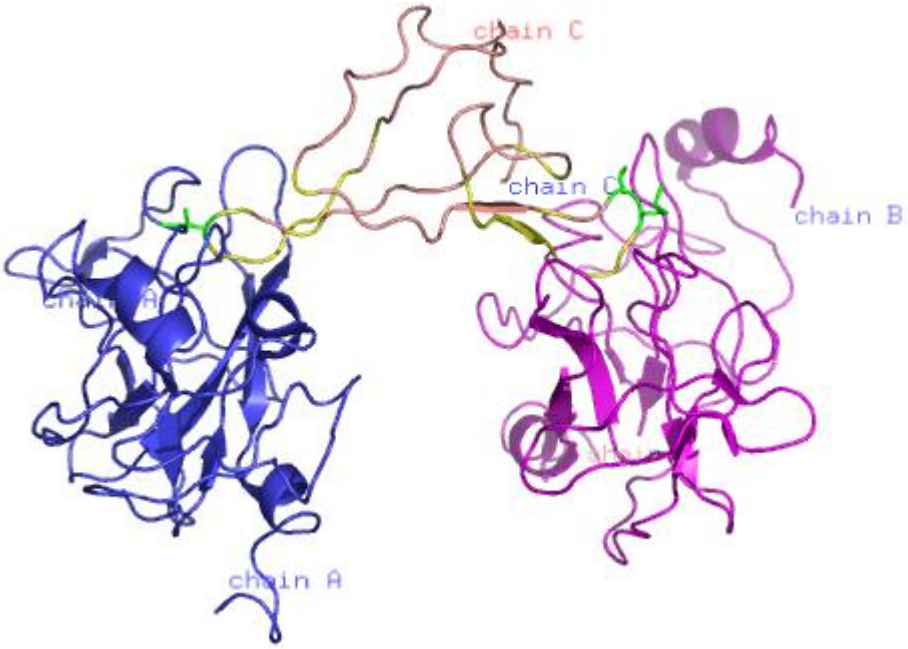


Figure 4.21. Complex structure and prediction results for 1DF9C.pdb

4.5.5. A chimera of beta-catenin and alpha-catenin

Structure of the dimerization and beta-catenin-binding region of alpha-catenin (PDB code: 1DOW) is shown in Figure 4.22. Chain A of this structure is taken for the prediction. Chain A contains 194 amino acids and 52 of them classified as interface residue according to their ASA differences. 13 of the binding amino acids are classified correctly by SVM and they are shown in green. The remaining 39 interface residues are shown in yellow. In this case there are only 3 misclassified non-binding residues and they are displayed in red.

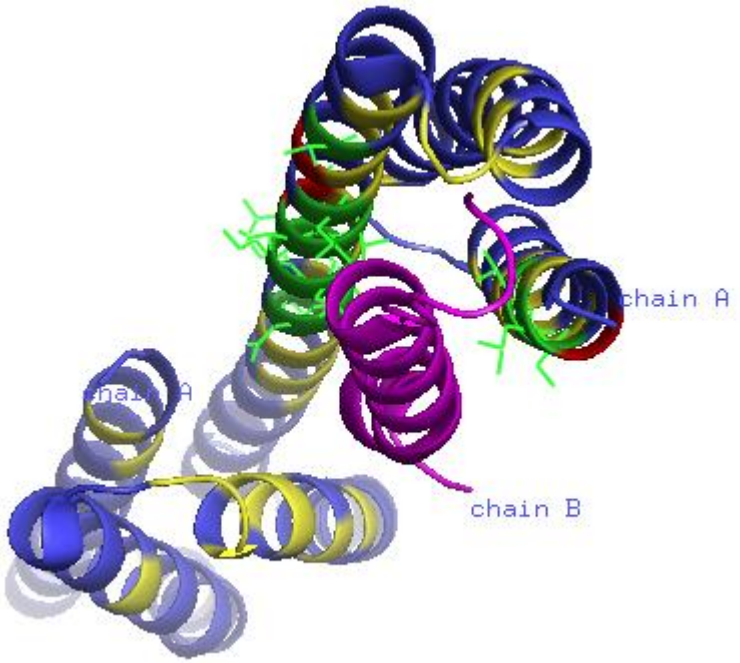


Figure 4.22. Complex structure and prediction results for 1DOWA.pdb

4.5.6. Rubredoxin

PDB id of Rubredoxin is 1E5D and it contains two chains. Chain A of this protein is taken for the prediction and it has 395 residues that 41 of them are interface residues as displayed by green in Figure 4.23. The wrong classified binding and non-binding residues are in yellow and red respectively. There are 41 misclassified binding residues and only one non-binding residue.

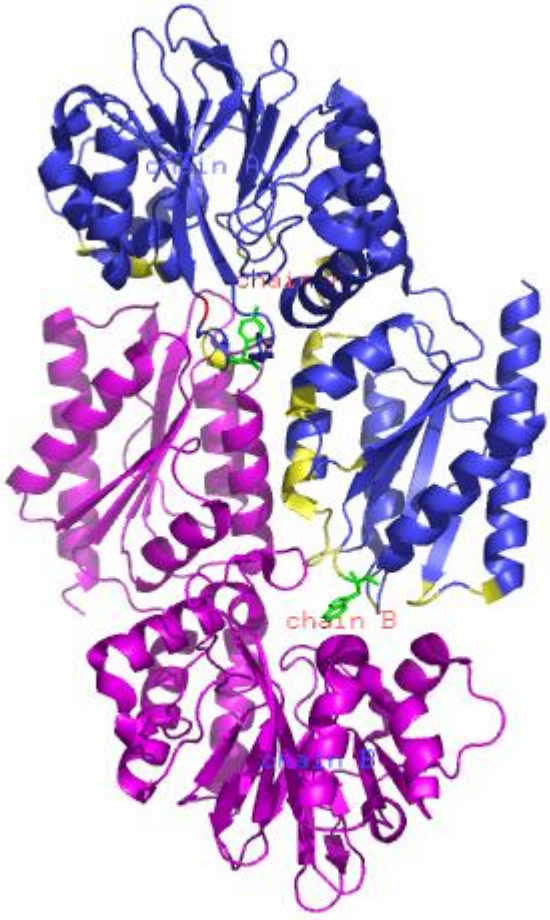


Figure 4.23. Complex structure and prediction results of 1E5DA.pdb

4.5.7. Inorganic pyrophosphatase

Inorganic pyrophosphate (PDB code: 1E9G) is a homo dimer. For the prediction chain A of this protein is used. There are 278 amino acids in this protein and 28 of them are classified as binding interface. 2 residues are predicted as binding residues and they are displayed in green in Figure 4.24. The misclassified 26 interface residues are shown in yellow. There is no misclassified non-binding residue in this protein.

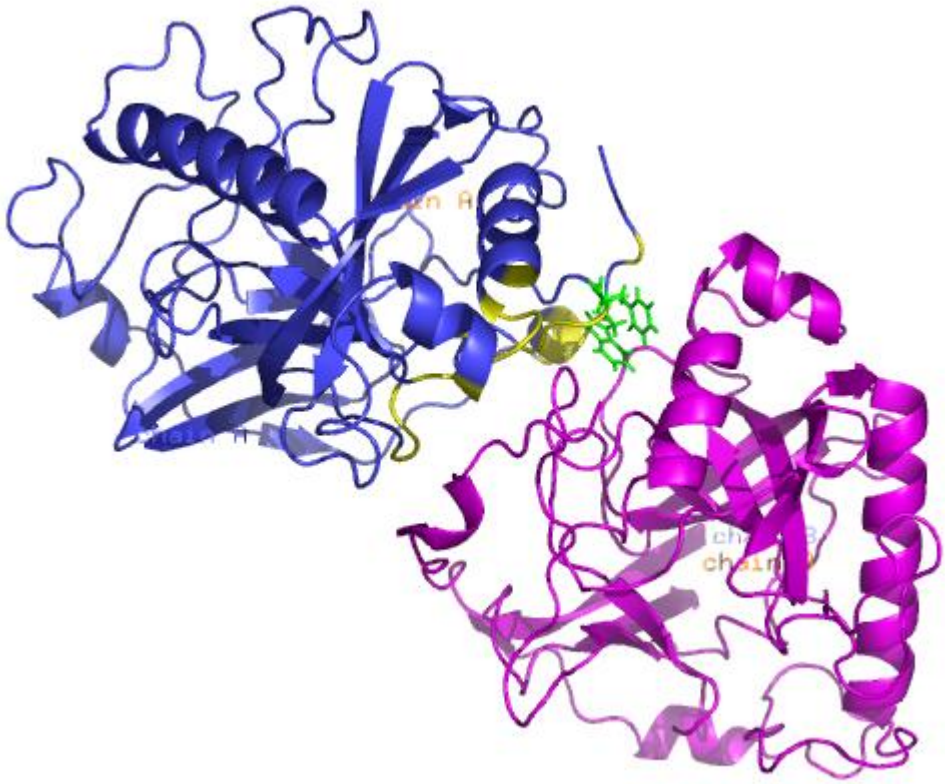


Figure 4.24. Complex structure and prediction results for 1E9GA.pdb

4.5.8. Internalin E-catalin complex

Interlin E-catalin complex is a hetero dimer. Both of the chains of this complex are used for the prediction. There are 454 amino acids in its chain A and 41 of them are classified as binding residues. SVM could not predict any binding residues for this chain. All the binding residues are wrongly predicted as non-binding residues. These misclassified binding residues are shown in yellow in Figure 4.25. On the other hand, there are 94 residues in chain B and 25 of them are classified as binding according to the accessible surface area differences. The correctly predicted number of binding residues is one and it is in green and one misclassified non-binding residue is in red. The misclassified number of binding residues is 24 and they are displayed in yellow.

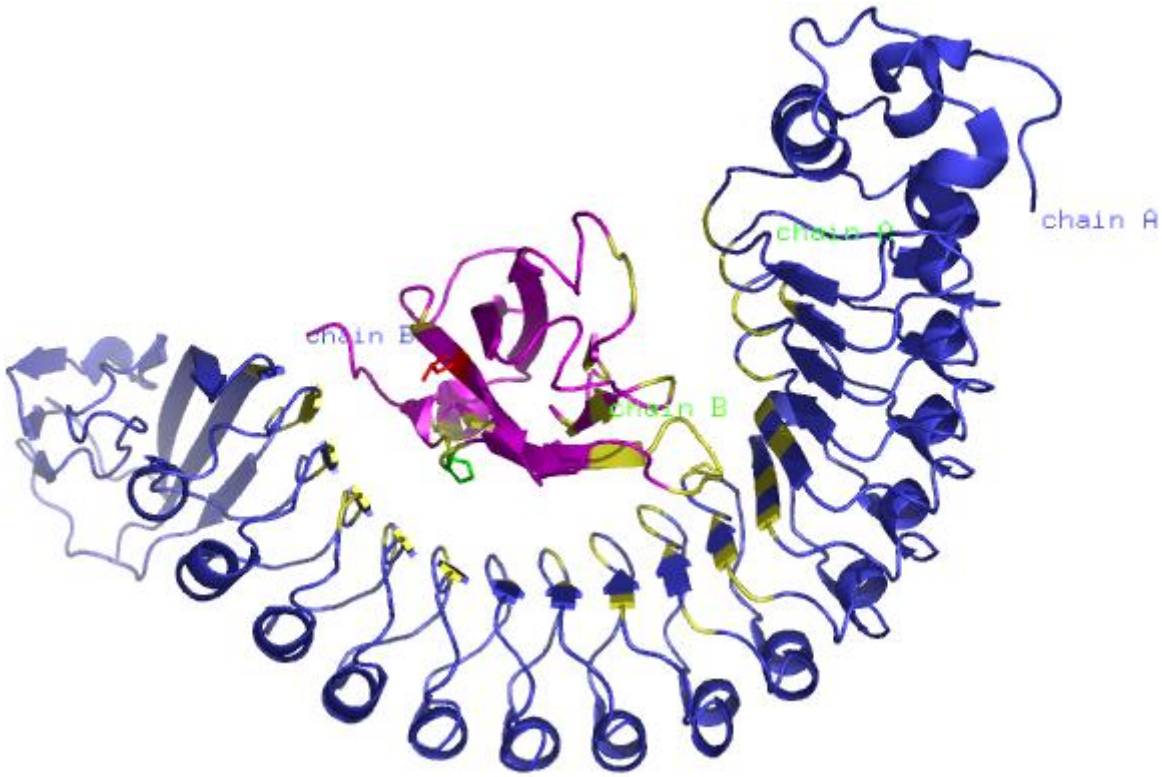


Figure 4.25. Complex structure and predicted results of both 1O6SA.pdb and 1O6SB.pdb

4.5.9. Dimeric hemoglobin

The dimeric hemoglobin protein (PDB code: 3SDH) is the last protein used in prediction. Chain A of this protein is used and it contains 139 amino acids. 25 of these residues involved in interface and 4 of them are correctly classified. These correctly classified residues are displayed in green in Figure 4.26. The number of misclassified binding residues is 21 and the misclassified non-binding residues are 5, they are in yellow and red in Figure 4.26 respectively.

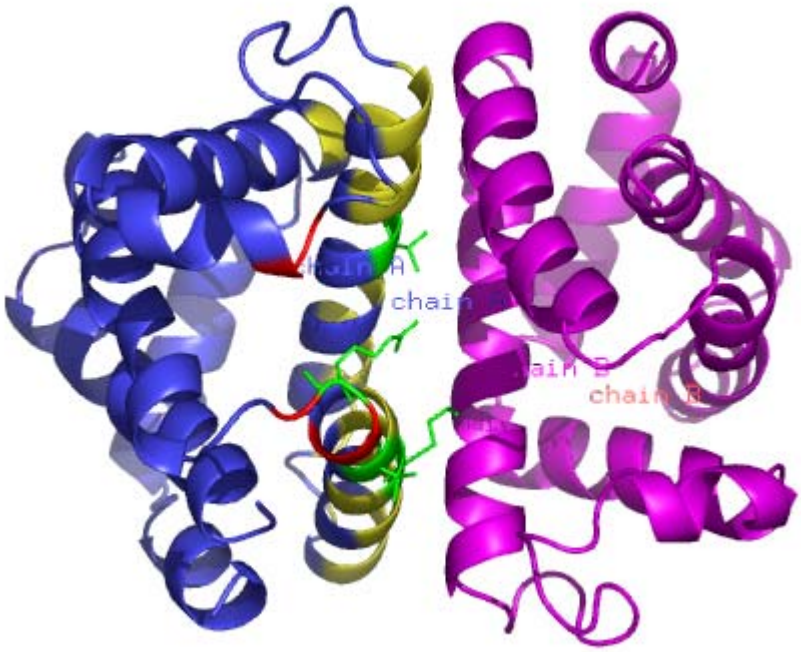


Figure 4.26. Complex structure and predicted binding residues of 3SDHA.pdb

5. CONCLUSIONS AND FUTURE WORK

5.1. Conclusions

Large amount of studies on prediction of protein binding sites and protein-protein interactions without knowing the binding partner have been done over the years. Some of them have used only the information obtained from sequence of the protein and some of them used both the sequence and the structural properties focusing on particular types of complexes. Few of them have considered the dynamics of the structures. Much progress has been made in our understanding of the driving forces of protein-protein interactions.

The focus of this work is to determine the distinctive properties between the protein interface residues and the rest of the protein and by using these features predicting the putative binding sites of the unbound protein using machine learning tools.

Analysis on amino acid preferences of protein interfaces showed that there are considerable differences in preferences of some amino acids. Arg, Cys, His, Ile, Leu, Met, Phe, Trp, Tyr and Val appears to be preferred at interface or core of protein, while Asn, Asp, Glu, Gln, Lys, Pro, Ser, Thr are favored at non-interface surface of protein. There are higher proportions of hydrophobic amino acids at interface and core of protein, while at the non-binding surface hydrophilic ones are favored at non-binding surface. Positively charged polar residues are preferred at binding interface while the non-polar or neutral polar residues and negatively charged polar residues are abundant at core and non-binding interface, respectively. Residue conservation has also observed to be higher at both interface and core compared to non binding protein surface. The accessible surface areas of residues at interface and non-interface are compared and it is observed that the binding residues accumulate either at too low or too high values in the former suggesting that there are binding residues at cavities and eaves.

The four different types of interactions, namely, homo, hetero, transient and permanent, are analyzed in terms of all the aforementioned static properties to see if they exhibit any significant character. First the homogeneous and the heterogeneous complexes

are compared and it is observed that although they have some distinctive preferences with respect to some properties, such as temperature factor, accessible surface area and amino acid types, these differences are not sufficient to treat them as two different classes. Then the interactions are analyzed in terms of the lifetime of complexes. Transient versus permanent complexes are compared according to their preferences. This separation exhibit more distinct characters compared to the preceding division. Permanent complexes are more hydrophobic than the transient ones and also different amino acids are preferable for these two kinds of interfaces. In spite of these differences, an analysis aimed to differentiate the binding residues from the rest of the residues of the protein may not treat them as two different classes. This present work did not separate the complexes when predicting the binding residues.

All the aforementioned properties reveal differences in terms of binding, core and non-binding surface residues; however such observations are insufficient to localize the protein binding sites. Binding residues cannot be uniquely identified by their electrostatic characteristics or shape of the interface. More distinguishing properties are needed, such as the contribution of the amino acids to the total free energy of the protein. The binding hot spots undertake the energetic stability of both the unbound protein structure and the structure of protein complex. These energetic amino acids are detected by in-silico alanine scanning mutagenesis and analyzed in terms of binding, core or non-binding surface. The residues that have high energy are located either at interface or core of the protein.

The dynamic characteristics of the amino acids may also assist to localize the binding residues. The amino acids at protein core and binding interface exhibit lower temperature factors compared to the non-binding surface residues already at unbound structure. Additionally, the fluctuations in both fastest and slowest modes of motion give information about both structurally and/or functionally important residues. The relative fluctuations of residues with the other residues in a structure at the two termini of the dynamic spectrum suggest the anchor and the anchoring groove residues that could highly be associated with binding residues. To this end, to consider the dynamic peculiarity of the residues in a structure together with other sequence and structure based properties would increase the prediction of binding interfaces.

All these attributes are necessary but may not be sufficient individually for determining the binding residues of an unbound protein. Combinations of them would possibly end up with a more successful prediction. Thus, support vector machines (SVM) and multiple kernel learning (MKL) are used by taking all these properties as features of the classification. The insertion of the dynamic characteristics of structures, namely, temperature factor and relative correlations between fluctuations in both fast and slow modes, improves the prediction accuracy and SVM with linear kernel evaluates the best accuracy. MKL yields about the same accuracy but it additionally gives the contribution of the features to the prediction by weighting the kernels. The three sequence neighbors from both side of the residue and the non-bonded closest six residues in 10 Å cut-off distance have almost the same contribution to the final prediction. On the other hand, residue type itself, evolutionary conservation scores, accessible surface area and its position of the amino acid (core or surface), relative correlations of fluctuations in fast modes, relative correlation of fluctuations with tails in slow modes, and packing are the most contributed attributes.

The maximum accuracy on the test set obtained during this work is 81.3 %.

5.2. Recommendations

Biological processes are usually complex enough that they do not be represented by just a hundred of examples. The dataset used for this work is consists of 263 proteins with 59748 amino acids, but because of the computational limitations maximum 20 % of the dataset is used for training and 20 % for validation. Both the number of proteins in the dataset and the number of examples used for training and validation should be increased further.

Although the accuracy obtained is higher than the previous ones, the coverage is low. This dilemma is caused by the unbalanced number of examples of the two classes in the dataset. Only, 21 % of the residues composing the dataset are interface residues. So the classifier is biased to predict a residue as non-binding. In order to overcome this situation, a new dataset that contains comparable amount of examples from both classes should be used for training and validation.

Here, the better accuracy is obtained by support vector machines with linear kernel than 2nd order polynomial kernel. However, the problem is complex in terms of features that linear separation may not work well. Other complex kernels different than polynomial or? Gaussian can be tried.

The main assumption of this work is that the structure of the protein does not change during decomplexation but this may not be the case. For a better understanding, the dataset should be consists of the proteins that have known structures in both bound and unbound states.

APPENDIX

Tablo A.1. Dataset

12ASA	1BVYF	1DQSA	1FM9D	1GL2C	1HZPA	1JXHA	1LD8B	1NKSA
1A0FA	1BYFA	1DTWA	1FO0A	1GL4A	1I1RB	1JY2N	1LDJA	1NMMB
1A2XA	1BYKA	1DX5I	1FO0B	1GL4B	1I2MB	1JY2P	1LDJB	1NMUA
1A38A	1BZYA	1E2AA	1FOEA	1GO3E	1I4DA	1JZDA	1LH0A	1NVTA
1A79A	1C28A	1E2TA	1FP3A	1GPWA	1I9BA	1JZDC	1LHPA	1O94D
1A88A	1CD9B	1E96B	1FS1B	1GPWB	1IA9A	1K1DA	1LI1A	1OSPO
1AA7A	1CG5A	1EAIC	1FSKA	1GT7A	1ICFI	1K20A	1LK5A	1PREA
1ABRB	1CHMA	1EBDC	1FTRA	1GVNA	1IG0A	1K2FA	1LL0A	1PRTA
1AD3A	1CI6A	1ED9A	1FUIA	1GX1A	1IK9A	1K3BA	1LM7A	1QA9A
1ADJA	1CI6B	1EERA	1FXKA	1GY9A	1IM9D	1K83C	1LQSR	1QBKB
1AFRA	1CMXA	1EERB	1FYHB	1GZ0A	1IREA	1K83H	1LR5A	1QFHA
1AIHA	1CRUA	1EK9A	1G0HA	1GZSB	1ISIA	1K83K	1LVOA	1QGWA
1AJSA	1CSEE	1EP3B	1G0SA	1H1YA	1ITUA	1K8KD	1M1EB	1QO0A
1AONO	1CYDA	1EUAA	1G3JA	1H2IA	1IXSB	1K8KE	1M2DA	1SGPE
1APYB	1D3BB	1EUVB	1G4YB	1H4LD	1J5SA	1K8KF	1M2OA	1SMTA
1AROP	1D4FA	1EV2E	1G57A	1H59B	1J7DA	1K8KG	1M4UA	1TBRR
1AT3A	1D4VA	1EWYC	1G5HA	1H6KA	1JB0D	1KACB	1M6PA	1TDTA
1AX4A	1D4XG	1F34B	1G72A	1H6KX	1JEQB	1KEYA	1M7GA	1TX4A
1AXIB	1D7AA	1F39A	1G73C	1H7EA	1JG5A	1KF6B	1MBXC	1TYFA
1AZSA	1D9EA	1F3UB	1G8EA	1H9SA	1JKGB	1KKMA	1MG2A	1XDTR
1AZZC	1DBQA	1F3VA	1G99A	1HCFX	1JMAB	1KQ4A	1MJGM	1YNJA
1B33A	1DCUA	1F45A	1GC1C	1HG3A	1JMVA	1KWSA	1MJHA	2EBOA
1B34B	1DF9A	1F75A	1GCQC	1HN2A	1JPYA	1KXPD	1MPYA	2TRCP
1B6SA	1DJ0A	1F80A	1GG2B	1HQ3A	1JQLA	1L0OA	1MR1C	
1BBHA	1DKGA	1F8MA	1GH6A	1HQ3D	1JT6A	1L0OC	1MZ9A	
1BD3A	1DM5A	1F8UB	1GH6B	1HSSA	1JTHA	1L1OB	1N1JB	
1BE3I	1DN1A	1F9AA	1GHQA	1HULA	1JV2A	1L1OC	1N9RA	
1BI7B	1DO8A	1FCDC	1GHQB	1HUXA	1JV2B	1L6WA	1NBAA	
1BO1A	1DOWB	1FCJA	1GL2A	1HX3A	1JW9B	1L9WA	1NBFA	
1BVNT	1DQNA	1FLTX	1GL2B	1HYNP	1JX2B	1LD8A	1NF3C	

Tablo A.2. Hetero complexes

1A2XA	1D4VA	1F8UB	1GL2C	1JV2A	1L0OC
1ABRB	1D4XG	1FM9D	1GL4A	1JV2B	1LD8A
1AJSA	1DN1A	1FO0A	1GL4B	1JW9B	1LD8B
1AROP	1DOWB	1FO0B	1H59B	1JX2B	1LDJA
1AXIB	1DTWA	1FXKA	1H9SA	1JZDC	1LDJB
1AZSA	1E96B	1G4YB	1I1RB	1K3BA	1M1EB
1B34B	1EBDC	1GC1C	1IM9D	1K83C	1M4UA
1BE3I	1EERA	1GCQC	1IREA	1K83H	1N1JB
1BI7B	1EP3B	1GG2B	1IXSB	1K83K	1OSPO
1BVNT	1EUAA	1GH6A	1J7DA	1K8KD	1QBKB
1BVYF	1EUVB	1GH6B	1JB0D	1K8KE	1QGWA
1CG5A	1EWYC	1GHQA	1JEQB	1K8KF	1SGPE
1CI6A	1F34B	1GHQB	1JKGB	1K8KG	1TX4A
1CI6B	1F3VA	1GL2A	1JMAB	1KACB	2TRCP
1CSEE	1F45A	1GL2B	1JQLA	1KXPD	

Tablo A.3. Homo complexes

12ASA	1C28A	1F39A	1H1YA	1JMVA	1LVOA
1A0FA	1CHMA	1F75A	1H2IA	1JPYA	1M2DA
1A79A	1CRUA	1F8MA	1H7EA	1JT6A	1M6PA
1A88A	1CYDA	1F9AA	1HG3A	1JXHA	1M7GA
1AA7A	1D4FA	1FCJA	1HN2A	1K1DA	1MJHA
1AD3A	1D7AA	1FP3A	1HSSA	1K20A	1MPYA
1ADJA	1D9EA	1FTRA	1HULA	1K2FA	1MZ9A
1AFRA	1DBQA	1FUIA	1HUXA	1KEYA	1N9RA
1AIHA	1DCUA	1G0HA	1HX3A	1KQ4A	1NBAA
1AT3A	1DJ0A	1G0SA	1HYNP	1KWSA	1NKSA
1AX4A	1DM5A	1G57A	1HZPA	1L6WA	1NVTA
1B6SA	1DO8A	1G5HA	1I9BA	1L9WA	1PREA
1BBHA	1DQNA	1G8EA	1IA9A	1LH0A	1QFHA
1BD3A	1DQSA	1G99A	1IG0A	1LHPA	1SMTA
1BO1A	1E2AA	1GT7A	1ISIA	1LK5A	1TDTA
1BYFA	1E2TA	1GX1A	1ITUA	1LL0A	1TYFA
1BYKA	1ED9A	1GY9A	1J5SA	1LM7A	2EBOA
1BZYA	1EK9A	1GZ0A	1JG5A	1LR5A	

Tablo A.4. Transient complexes

12ASA	1BYKA	1ED9A	1HUXA	1JT6A	1MJHA
1A38A	1CHMA	1EK9A	1HX3A	1JV2A	1MR1C
1A79A	1CI6A	1EP3B	1HYNP	1JX2B	1N9RA
1A88A	1CI6B	1F39A	1HZPA	1JXHA	1NBAA
1ABRB	1CRUA	1F3UB	1I1RB	1K1DA	1NKSA
1AD3A	1CYDA	1FCJA	1IA9A	1K3BA	1NVTA
1ADJA	1D4FA	1FP3A	1IG0A	1KXPD	1OSPO
1AIHA	1D4FA	1G0HA	1IM9D	1L1OC	1PREA
1AJSA	1D9EA	1G0SA	1IREA	1LD8A	1QA9A
1AT3A	1DBQA	1G99A	1ISIA	1LD8B	1QFHA
1AX4A	1DJ0A	1GH6A	1J5SA	1LI1A	1SMTA
1B34B	1DO8A	1GH6B	1J7DA	1LVOA	1YNJA
1BBHA	1DQNA	1H1YA	1JEQB	1M4UA	2EBOA
1BO1A	1DQSA	1HN2A	1JMVA	1M6PA	
1BYFA	1DTWA	1HULA	1JPYA	1M7GA	

Tablo A.5. Permanent complexes

1A0FA	1DKGA	1GCQC	1JKGB
1A2XA	1DOWB	1GL4A	1JMAB
1AROP	1DX5I	1GL4B	1JW9B
1AXIB	1E2AA	1GO3E	1KACB
1BE3I	1EUVB	1GVNA	1M1EB
1BI7B	1F34B	1H6KA	1M2DA
1BVNT	1F45A	1H6KX	1MBXC
1BVYF	1F8UB	1H9SA	1O94D
1CG5A	1FCDC	1HCFX	1SGPE
1CSEE	1G57A	1JB0D	1TBRR
1D4XG	1G73C	1JG5A	1TX4A
1DF9A	1G8EA	1JG5A	1XDTR

REFERENCES

Alpaydin E., 2004, *Introduction to Machine Learning*, The MIT Press, Cambridge.

Altschul, S. F., T. L. Madden, R. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, 1997, "Gapped blast and psi-blast: a new generation of protein database search programs", *Nucleic Acids Res.*, Vol. 25, pp. 3389-3402.

Bach, F., G. Lanckriet and M. Jordan, 2004, "Multiple kernel learning, conic duality, and the smo algorithm", *Proceedings of the 21st International Conference on Machine Learning*, pp. 41–48.

Bahar, I., A. R. Atilgan, M. C. Demirel and B. Erman, 1998, "Vibrational dynamics of proteins: Significance of slow and fast modes in relation to function and stability", *Physical Review Letters*, Vol. 80, pp. 2733-2736.

Bahar, I., A. R. Atilgan, and B. Erman, 1997, "Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential", *Folding and Design*, Vol. 2, pp. 173-181.

Bernstein, F. C., T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, 1977, "Protein Data Bank: a computer-based archival file for macromolecular structures", *Jounal of Molecular Biology*, Vol. 112, pp. 535-542.

Bogan, A. A. and K. S. Thorn, 1998, "Anatomy of hot spots in protein interfaces", *J. Mol. Biol.*, Vol. 280, pp. 1-9.

Bradford, J. R. and D. R. Westhead, 2005, "Improved prediction of protein-protein binding sites using a support vector machines approach", *Bioinformatics*, Vol. 21. No. 8, pp. 1487-1494.

Branden, C. and J. Tooze, 1991, *Introduction to protein structure*, Garland Pub. Inc., New York and London.

Brodner, A. J. and R. Abagyan, 2005, "Statictical analysis and prediction of protein protein interfaces", *Proteins*, Vol. 60, pp. 353-366.

Bromberg, Y. and B. Rost, 2008, "Comprehensive in silico mutagenesis highlights functionally important residues in proteins", *Bioinformatics*, Vol. 24, pp. i207-i212.

Burges, C. J. C., (1998), A tutorial on support vector machines pattern recognition, Data Mining and Knowledge Discovery Journal, Vol. 2, No. 2, pp. 121-167.

Burgoyne, N. J. and R. M. Jackson, 2006, "Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces", *Bioinformatics*, Vol. 22, pp. 1335-1342.

Chen, H. and H. –X. Zhou, 2005, "Prediction of interface residues in protein-protein complexes by a consensus neural network model: Test against NMR data", *Proteins*, Vol. 61, pp. 21-35.

Cheng, G., B. Qian, R. Samudrala and D. Baker, 2005, "Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design", *Nucleic Acid Res.*, Vol. 33. No. 18, pp. 5861-5867.

Chung, J. L., W. Wang and P. E. Bourne, 2006, "Exploiting sequence and structural homologs to identify protein-protein binding sites", *Proteins*, Vol. 62. No. 3, pp. 630-640.

Clackson, T. and J. A. Wells, 1995, "A hot spot of binding energy in a hormone receptor interface", *Science*, Vol. 267, pp. 383-386.

Crowley, P. B. and A. Golovin, 2005, "Cation-π interactions in protein-protein interfaces", *Proteins*, Vol. 59, pp. 231-239.

DeLano, W. L., M. H. Ultsch, A. M. de Vos, and J. A. Wells, 2000, "Convergent solutions to binding at a protein-protein interface", *Science*, Vol. 287, pp. 1279-1283.

Demirel, M. C., O. Keskin, 2005, "Protein interaction and fluctuations in a proteomic Network using an Elastic Network Model", *J. Biomolecular Structure & Dynamics*, Vol. 22, No. 4, pp. 381-386.

Dong, Q., X. Wang, L. Lin and Y. Guan, 2007, "Exploiting residue-level and profile level interface propensities for usage in binding sites prediction of proteins", *BMC Bioinformatics*, Vol. 8, pp. 147-159.

Edgar, R. C., 2004, "MUSCLE: multiple sequence alignment with high accuracy and high throughput", *Nucleic Acids Research*, Vol. 32, pp. 1792-1797.

Ertekin, A., R. Nussinov and T. Haliloglu, 2006, "Association of putative concave protein binding sites with the fluctuation behaviour of residues", *Protein Science*, Vol. 15, pp. 2265-2277.

Fariselli, P., F. Pazos, A. Valencia and R. Casado, 2002, "Prediction of protein-protein interaction sites in heterocomplexes with neural networks", *Eur J Biochem*, Vol. 269, pp. 1356-1361.

Glaser, F., T. Pupko, I. Paz, R. E. Bell, D. Bechor-Shental, E. Martz and N. Ben-Tal, 2003, "ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information", *Bioinformatics*, Vol. 19, No. 1, pp. 163-164.

Haliloglu T. and N. Ben-Tal, 2008, "Cooperative Transition between open and closed conformations in potassium channels", in press…

Haliloglu, T., E. Seyrek and B. Erman, 2008, "Prediction of binding sites in Receptor-Ligand complexes with the Gaussian Network Model", *Physical Review Letters*, Vol. 100, 228102.

Haliloğlu, T., I. Bahar and B. Erman, 1997, "Gaussian Dynamics of proteins", *Physical Review Letters*, Vol. 79, pp. 3090-3093.

Haliloglu, T., O. Keskin, B. Ma and R. Nussinov, 2005, "How similar are protein folding and protein binding nuclei? Examination of vibrational motion of energy hotspots and conserved residues", *Biophys. J.*, Vol. 88, pp. 1552-1559.

Hoskins, J., S. Lovell and T. M. Blundell, 2006, "An algorithm for predicting protein-protein interaction sites: abnormally exposed amino acid residues and secondary structure elements", *Protein Science*, Vol. 15, pp. 1017-1029.

Hu, Z., B. Ma, H. J. Wolfson and R. Nussinov, 2000, "Conservation of polar residues as hot spots at protein interfaces", *Proteins: Strcut. Funct. Genet.*, Vol. 39, pp. 331-342.

Ivanciuc, O., 2007, Applications of support vector machines in chemistry, Reviews in Comp. Chemstry, Vol.23, pp.291-400.

Janin, J., 1979, "Surface and inside volumes in globular proteins", *Nature*, Vol. 277, pp. 491-492.

Jones, S. and J. M. Thornton, 1995, "Protein-protein interactions: a review of protein dimer structures", *Prog. Biophys. Mol. Biol.*, Vol. 63, pp. 31-65.

Jones, S. and J. M. Thornton, 1996, "Principles of protein-protein interactions", *Proc. Natl. Acad. Sci. USA*, Vol. 93, pp. 13-20.

Jones, S. and J. M. Thornton, 1997, "Analysis of protein-protein interaction sites using surface patches", *Journal of Molecular Biology*, Vol. 272(1), pp. 121-132.

Kabsch, W., and C. Sander, 1983, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features", *Biopolymers*, Vol. 22, pp. 2577-2637.

Keskin, O., B. Ma and R. Nussinov, 2005, "Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues", *J Mol Biol*, Vol. 345, pp. 1281-1294.

Kessel, A., D. Shental-Bechor, T. Haliloglu and N. Ben-Tal, 2003, "Interaction of hydrophobic peptides with lipid bilayers: Monte Carlo simulation with M2δ", *Biophysical Journal*, Vol. 85, pp. 3431-3444.

Kuriyan, J. and W. I. Weis, 1991, "Rigid protein motion as a model for crystallographic temperature factors", *Proc. Natl. Acad. Sci. USA*, Vol. 88, pp. 2773-2777.

Lanckriet, G., T. D. Bie, N. Cristianini, M. Jordan and W. Noble, 2004, "A statistical framework for genomic data fusion", *Bioinformatics*, Vol. 20, pp. 2626–2635.

Landau, M., I. Mayrose, Y. Rosenberg, F. Glaser, E. Martz, T. Pupko, and N. Ben-Tal, 2005, "Consurf 2005: the projection of evolutionary conservation scores of residues on protein structures", *Nucleic Acid Research*, Vol. 33, pp. 299-302.

Li, M. H., L. Lin, X. L. Wang and T. Liu, 2007, "Protein-protein interaction site prediction based on conditional random fields", *Bioinformatics*, Vol. 23, pp. 597-604.

Li, X., O. Keskin, B. Ma, R. Nussinov and J. Liang, 2004, "Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in unbound states: implications for docking", *J. Mol. Biol*., Vol. 344, pp. 781-795.

Liang, S., C. Zhang, L. Song and Y. Zhou, 2006, "Protein binding site prediction using a empirical scoring function", *Nucleic Acid Research*, Vol. 34, pp. 3698-3707.

Lichtarge, O., H. R. Bourne and F. E. Cohen, 1996, "An evolutionary trace method defines binding surfaces common to protein families", *Journal of Molecular Biology*, Vol. 257, pp. 342-358.

Lichtarge, O., H. R. Bourne and F. E. Cohen, 1996, "An evolutionary trace method defines binding surfaces common to protein families", *J. Mol. Biol*., Vol. 257, pp. 342-358.

Lo Conte, L., C. Chothia and J. Janin, 1999, "The atomic structure of protein-protein recognition sites", *J. Mol. Biol*., Vol. 285, pp. 2177-2198.

Murakami, Y. and S. Jones, 2006, "SHARP2: protein-peotein interaction prediction using patch analysis", *Bioinformatics*, vol. 22, pp. 1794-1795.

Neuvirth, H., R. Raz and G. Schreiber, 2004, "Promate: A structure based prediction program to identify the location of protein-protein binding sites", *Journal of Molecular Biology*, Vol. 338, pp. 181-199.

Nooren, I. M. and J. M. Thornton, 2003, "Structural characterization and functional significance of transient protein-protein interactions", *Journal of Molecular Biology*, Vol. 325(5), pp. 991-1018.

Ofran, Y. and B. Rost, 2003, "Predicted protein-protein interaction sites from local sequence information", *FEBS Lett*., Vol. 544, pp. 236-239.

Panchenko, A. R., F. Kondrashov and S. Bryant, 2004, "Prediction of functional sites by analysis os sequence and structure conservation", Vol. 13, pp. 884-892.

Parthasarathy, S. and M. R. N. Murthy, 1997, "Analysis of temperature factor distribution in high-resolution protein structures", *Protein Science*, Vol. 6, pp. 2561-2567.

Porollo, A. and J. Meller, 2007, "Prediction based fingerprints of protein-protein interactions", *Proteins: Structure, Function, and Bioinformatics*, Vol. 66, pp. 630-645.

Qin, S. and H. –X. Zhou, 2007, "meta-PPISP: a meta web server for protein-protein interaction site prediction", *Bioinformatics*, Vol. 23, pp. 3386-3387.

Rakotomamonjy, A., F. Bach, S. Canu, Y Grandvalet, 2007, "More efficiency in multiple kernel learning", *Proceedings of the 24st International Conference on Machine Learning (ICML)*.

Rose, G. D., A. R. Geselowitz, G. J. Lesser, R. H. Lee and M. H. Zehfus, 1985, "Hydrophobicity of amino acid residues in globular proteins", *Science*, Vol. 229, pp. 834-838.

Rost, B. and C. Sander, 1994, "Conservation and prediction of solvent accessibility in protein families", *Proteins: Struct. Funct. and Genet*., Vol. 20, pp. 216-226.

Schölkopf, B., A. J. Smola, 2002, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, Cambridge, London.

Sonnenburg, S, G. Rätsch, C. Schäfer, 2006, "Learning interpretable SVMs for biological sequence classification", *BMC Bioinformatics*, Vol. 7 (Suppl 1).

Sonnenburg, S., G. Raetsch, C. Schaefer and B. Scholkopf, 2006, "Large scale multiple kernel learning", *Journal of Machine Learning Research*, Vol. 7, pp. 1531–1565.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice", *Nucleic Acids Res*., Vol. 22, pp. 4673-4680.

Thorn, K. S. and A. A. Bogan, 2001, "ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions", *Bioinformatics*, Vol. 12, No. 3, pp. 284-285.

Van Dijk, A. D. J., S. J. de Vries, C. Domingues, H. Chen, H. –X. Zhou and A. M. J. J. Bonvin, 2005, "Data driven docking: HADDOCK's adventure in CAPRI", *Proteins*, Vol. 60, pp. 232-238.

Vapnik, V. N., 1998, *Statistical Learning Theory*, J. Wiley and Sons, New York.

Wolfenden, R., L. Andersson, P. Cullis and C. Southgate, 1981, "Affinities of amino acid side chains for solvent water", *Biochemistry*, Vol. 20, pp. 849-855.

Yan, C. H., V. Hannover and D. Dobbs, 2004, "Identification of interface residues in protease-inhibitor and antigen-antibody complexes: a support vector machine approach", *Neural Comp. Appl*., Vol. 13, pp. 123-129.

Yan, C., D. Dobbs and V. Hannover, 2004, "A two stage classifier for identification of protein-protein interface residues", *Bioinformatics*, Vol. 20 (suppl. 1), pp. 1371-1378.

Young, L., R. L. Jernigan and D. G. Covell, 1994, "A role for surface hydrophobicity in protein-protein recognition", *Protein Science*, Vol. 3, pp. 717-729.

Zhang, Z. and M. G. Grigorov, 2006, "Similarity networks of a protein binding sites", *Proteins*, Vol. 62, No. 2, pp. 470-478.

Zhou, H. X. and Y. Shan, 2001, "Prediction of protein interaction sites from sequence profile and residue neighbour list", *Proteins*, Vol. 44, pp. 336-343.

Zhou, H.-X. and S. Qin, 2007, "Interaction site prediction for protein complexes: a critical assessment", *Bioinformatics*, Vol. 23, No. 17, pp. 2203-2209.

Zhu, H., F. S. Domingues, I. Sommer, and T. Lengauer, 2006, "NOXclass: prediction of protein-protein interaction types", *BMC Bioinformatics*, Vol. 7:27.