ALMA MATER STUDIORUM - UNIVERSITA DI BOLOGNA

TITOLO TESI

PREDICTION OF PROTEIN-PROTEIN INTERACTION SITES WITH A NEW PROBABILISTIC METHOD

Presentata da: Saeideh NAZERI

Coordinatore:

Chiar.ma Prof.ssa Rita Casadio

Relatore:

Chiar.mo Prof. Pier Luigi Martelli

Correlatori:

Chiar.mo Prof. Piero Fariselli

Dr. Castrense Savojardo

Esame finale anno 2016

Abstract

In this study we applied of powerful discriminative model to predict interacting residues on the surface of the protein, Grammatical-Restrained Hidden Conditional Random Field (GRHCRF) has already reached to admirable results to address sequence labeling problems in different aspects of biology such as structure prediction of β-barrel membrane proteins. The main advantage of this method is that we can exert our prior knowledge into the model avoiding computational complexity compare to other generative and discriminative models.

We performed our method on dataset derived from a rich source of protein complexes benchmark 5.0 and we defined our features base on structural information of protein sequence. We made comparison with other state of the art predictors and we ranked a acceptable position compared to other methods.

Acknowledgements

First of all, I would like to express my profound gratitude to my Professor Rita Casadio for giving me the opportunity to study at Bio-computing Group of university of Bologna and for her precious support and guidance during the course and final thesis.

I would also like to thank my tutors Prof. Pier Luigi Martelli and Prof. Piero Fariselli for their valuable advices and help.

Special thanks to my advisor Dr. Castrense Savojardo for his special support and tolerance through the process of researching and writing this thesis.

Finally, I would like like to thank my Parents, Hossein and Pari, and my sister Hilda, this accomplishment would not have been possible without them.

Contents

Introduction ……………………………………………………………………………..…..9

1 Proteins, Genome, Proteome, Interactome ………………….….…………………...11

1.1 Protein definition ……………………………………………………………….12

1.1.1 Nucleic acids ………………………………………………………...13

1.1.2 Peptide chain ………………………………………………………….15

1.2 Protein structure ………………………………………………………………...16

1.2.1 Primary structure ……………………………………………………..17

1.2.2 Secondary structure …………………………………………………..18

1.2.3 Tertiary structure ……………………………………………………..19

1.2.4 Quaternary structure ………………………………………………….22

1.3 Protein function and evolution …………………………………………………23

1.4 Protein structures on the Web …………………………………………………23

1.4.1 Database of protein structures ………………………………………..24

1.4.2 Database of Protein structural Classification ………………………24

1.4.3 Database of classification of protein function ………………………25

2 Protein protein Interactions ……………………………………………………...27

2.1 PPI definition …………………………………………………………………..27

2.2 PPI types and characteristics ……………………………………………………28

2.2.1 Homo-oligomeric and Hetero-oligomeric complexes ……………...29

2.2.2 Obligate and non-obligate complexes ……………………………….29

2.2.3 Transient and permanent complexes ……………………………….30

2.3 Physiochemical properties of PPI sites …………………………………………30

2.4 Interacting residue definition and features …………………………………..31

2.5 PPI Predictors ……………………………………………………………………….32

2.5.1 Intrinsic-based predictors ……………………………………………..33

2.5.1.1 Sequence-based interface predictors ………….……..33

2.5.1.2 Structure-based predictors ………………..…………...33

2.5.2 Template-based predictors ……………………………………………...34

2.5.3 Partner-specific interface predictors ………………………………… 34

2.6 Databases of protein interactions ………………………….…………………...35

3 Probabilistic methods for sequence analysis ……………………………………………38

3.1 Probabilistic methods ……………………………………………………………..39

3.1.1 Generative models ……………………………………………………………...40

3.1.2 Discriminative models …………………………………………………...41

3.2 Grammatical-Restrained Hidden Conditional Random Fields ……………..42

3.2.1 GRH-CRF formulation ………………………………...…………44

3.2.2 Parameter estimation ……………………………………………………46

3.3 Decoding algorithms ……………………………………………………………………...48

4 Implementation Of CRF_GRF method for protein interaction prediction …...…...52

4.1 Dataset …………………………………………………...……………………………52

4.2 Grammar definition ……………………………………………………………………..55

4.3 Accuracy measures for PPI residue prediction 56

4.3.1 10-Fold cross validation 56

4.3.2 Fundamental definitions 57

4.4 Evaluation of Performance 58

4.4.1 Performance on BM_609_16 and BM_609_5 59

4.4.2 Comparison with performance of predictors using experimental structures 62

Conclusion 64

Bibliography 66

List of Figures

1.1 Omic sciences and their interaction. The flow of biological information is bidirectional.. 12

1.2 Amino acid Structure , R group (in red) is different in each amino acid……………..…………13

1.3 : 20 amino acid residue chemical structure are clustered to five groups, the shaded red part shows distinctive R-group also named side chain ………………………………………………………....15

1.4 : formation of peptide bond between 2 amino acid with release of H2O ……………………….16

1.5: partial view of four different levels of protein structure in this case hemoglobin…………..17

1.6(a): α-helix secondary structure of protein chains (b)zig-zag shape of β-sheet ………..……..19

1.7: formation of disulfide bonds between two cystein residue which highly stabilize protein structure……………….......…………………………………………………………………………………….....21

3.1: Schematic view of supervised machine learning, it can be summarized to two steps training and testing ………………………………………………………………………………………………....39

4.1: A sphere with 12 Angestrom radix is depicted around each residue and we mark the residues on that area which are on the surface……………………………………………………....55

4.2: The graphical view of two grammars defined for interacting residue prediction, grammar A include 3 states while grammar B has 11 states and more complex connection….. ….. 56

List of Tables

TTable 1.1 : The name and abbreviation of 20 different amino acid are listed……………………....14

Table 2.1 : Sample predictors for methods described in features section………….....………….....35

Table 2.2 : Databases of protein interactions…………………………………………………………....36

Table 4.1 : Performances of grammar A (with 3 states) on dataset BM_609_16, grammar B (with 11 states) on BM_609_16 ,grammar A on BM_609_5 and grammar B on BM_609_5…………………………………………………………………………………...……....61

Table 4.2 : Performance of different predictors on bench mark 4.0 dataset ……………………….62

Table 4.3: Performance of HomPPI predictor on 4 different datasets which contains different interaction types released on 2011……………………..……………………………….…………..63

Introduction

Protein-protein interactions are among the most important processes in biology, playing fundamental roles in the immune system, signaling pathways and enzyme inhibition. Proteome-wide studies have revealed that most proteins interact with other proteins[1,74]. To complement experimental approaches which include different approaches like intrinsic-based and template-based methods[22]. These approaches typically attempt to use state of the art machine learning methods to predict residues or patches in proteins that participate in interactions. Thus most of the methods has shown variation and different performance on different datasets. Some of the methods uses more than 100 features in dataset, it seems it is needed to take step higher than just emphasizing on different feature definitions which it seems they have reached to their saturation limit[24].

This fact motivated us to apply a powerful computational method called: Grammatical Restrained Hidden Conditional Random Fields, which have already performed better than similar predictors in addressing sequence labeling problem in different biological challenges like structure prediction of β-barrel membrane proteins[68]. The advantage of this method is that we can insert our biological knowledge or pattern discoveries in form of grammar to force the model to train from those rules and increase the discrimination power without increasing the time and space computational complexity.

Beside the computational method we developed the training and test sets that are non-redundant and rich source of protein complexes derived from benchmark 5.0. The features for classifying residues is base on structural information of the protein sequences. Finally we compared our results with reliable predictors and in most of the cases our method performed better or equal.

Chapter 1

1. Proteins, Genome, Proteome, Interactome

Cells are the building blocks of all living organisms, it is known as a complex system possess extraordinary attributes like building its owns constituents, selfreplicating, responding to its environment, recovering from damage and producing energy from organic compound or sunlight[2]. These function also knowns as cellular pathways are result of millions of interactions between advanced macromolecules[1]. There is so much variety in macromolecules in terms of their precursor, shape, size and specialized function. Beside all the variation in macromolecules in a cell, they share one basic fundamental, they are all result of polymerization of subunit molecules and these monomers determine what will be the category of the final result[3].

Macromolecules in cell can be analyzed from different dimensions, each of the dimensions called “omic” technology which adopt a systematic approach to study the biological events that make up a cell, tissue or organism[11](figure 1.1).

The genome is the total DNA of a cell or organism, it measure differences in DNA sequence between individuals and the expression of thousands of genes simultaneously. The transcriptome reflects the gene expression as mRNA macromolecule that are template for protein synthesis in a process called translation[12].

The proteome is defined as the set of all expressed proteins in a cell or tissue , the interactome is complete map of protein interactions that can occur in an organism. The knowledge in this area is only 10% of all PPIs in human[26], however the small portion of known part is valuable for explaining biological processes such as the molecular level links between diseases and proteins[22,23].

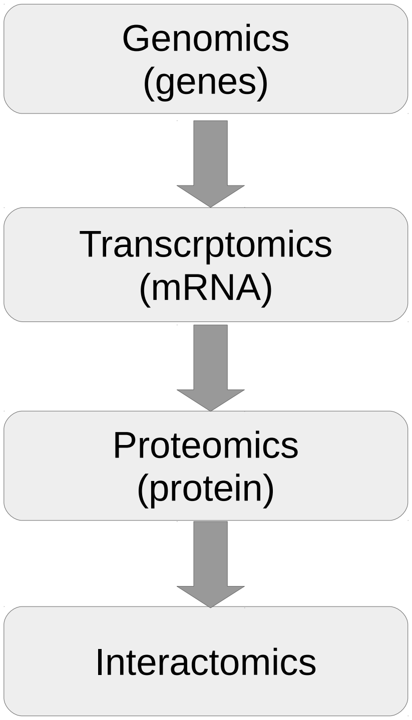


FIGURE 1.1: Omic sciences and their interaction. The flow of biological information is bidirectional.

In this chapter we focus on proteom level and we continue with more detailed definition of proteins and their structures to illustrate the basic concepts for our final target which is the prediction of interactions among proteins.

1.1 Protein definition

Proteins most abundant macromolecules in the cell which by far are the most structurally complex and functionally sophisticated molecules known. They have a major role in accomplishing cell functions. In a process called protein synthesis, proteins are made in the cell to carry out astonishing attributes of living system. Some type of proteins are quiet well-known with names like enzymes which promote many chemical reactions, other proteins carry messages from one cell to another, specialized proteins act as antibodies, toxins, hormones.

From a chemical point of view proteins are linear hetero-polymers of simpler organic molecules called amino acids. The chain of amino acids known as protein sequence is able to spontaneously achieve a stable and active three-dimensional (3D) structure in a polar solvent, the process by which the unstructured string of amino acids acquires its correct three-dimensional structure to achieve the biologically active native state is called protein folding[9].

1.1.1 Amino acids

20 amino acids(AA) are the alphabet of writing proteins, the basic structure of all the AA is: one carboxyl group(COO) and one amino group(H_3N) which both are connected to one carbon(in blue), called α-carbon. The difference between AA is base on their side chain (R-group)[1].

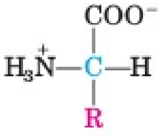


FIGURE 1.2: Amino acid Structure , R group (in red) is different in each amino acid.

physiochemical properties of the distinctive R group can help us to understand protein structure in higher levels, as we see in figure 1.3 the shape of 20 residues is depicted in 5 main category base on their similarity in physiochemical attributes like polarity, electro static charge and shape. These 5 groups are known with the names aliphatic, aromatic, polar, positively charged and negatively charged[3].

Polarity is one one the most important features of a residue which specify its behavior against solvent. Water is high polar solvent which forms hydrogen bonds with polar molecules in biochemistry it is known as solubility. If a residue is non-polar, it repels water and this action is called hydrophobicity (a type of phobia against water), in the other words high polar residues are soluble in water and have ability to form hydrogen bonds with solvent molecules, clearly they are hydrophilic. However there is a kind of relativity in hydrophobic power of the residues. For example Aromatic Residues are also considered as hydrophobic, hence compare to aliphatic group they have less force. The peak in hydrophilic attribute belong to residues which are negatively or positively charged, their tendency to form hydrogen bond with water is even higher than polar residues[1].

Amino acid	**Code**	**Amino acid**	**Code**	**Amino acid**	**Code**	**Amino acid**	**Code**
Alanine	A	Glycine	G	Methionine	M	Serine	S
Cysteine	C	Histidine	H	Asparagine	N	Threonine	T
Aspartic acid	D	Isoleucine	I	Proline	P	Valine	V
Glutamic acid	E	Lysine	K	Glutamine	Q	Tryptophan	W
Phenylalanine	F	Leucine	L	Arginine	R	Tyrosine	Y

Table 1.1 : The name and abbreviation of 20 different amino acid are listed.

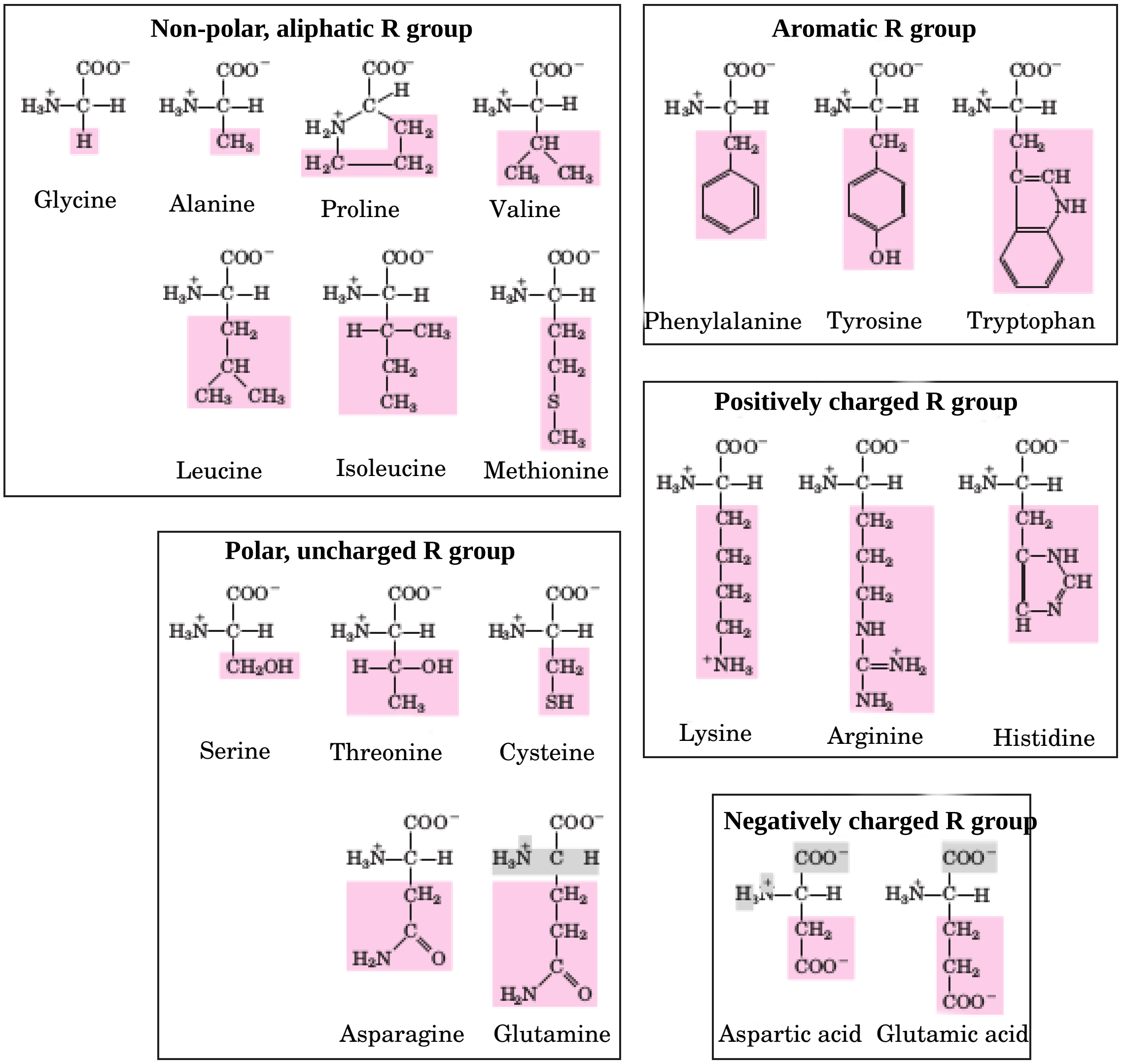


FIGURE 1.3 : 20 amino acid residue chemical structure are clustered to five groups, the shaded red part shows distinctive R-group also named side chain

1.1.2 Peptide chain

Each amino acid can connect to another amino acid versus a special covalent bond called peptide bond which is shown figure 1.4, while the formation of this bond one molecule of water(H_2O) is released, thats why the AA inside poly-peptide chains are called Residues[1].

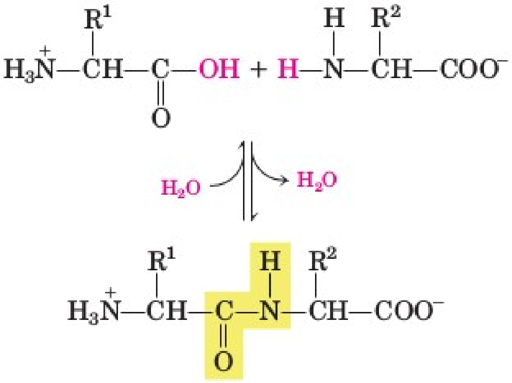

R¹
H
R²
$H_3\overset{+}{N}$—CH—C—OH + H—N—CH—COO⁻
O
H_2O
H_2O
R¹
H
R²
$H_3\overset{+}{N}$—CH—C—N—CH—COO⁻
O

FIGURE 1.4 : formation of peptide bond between 2 amino acid with release of H2O

Poly-peptide chain length can vary from 3 to thousands, the physiochemical properties of residue has direct impact on the properties of the chains, in other words amino acid composition determines the characteristic of poly-peptide chains[3]. Each protein contains from one or many chains, It is not all the story about proteins. Other type of chemical constitutes can be attached to proteins, constitutes like lipids, sugars or metals like iron, zinc or copper molecules which plays an important role in their biological function, however the main structure of the proteins is consist of sequence of residues[1]. The sequence of amino acid is the first level of protein structure, known as primary structure. There is much can be said about protein structure in next section.

1.2 Protein structure

Protein structure are studied at four different level which is called: primary, secondary, tertiary and quaternary structure, there is a type of conceptual hierarchy among different levels[2]. In figure 1.5 we see four different structure level of a well known protein called hemoglobin.

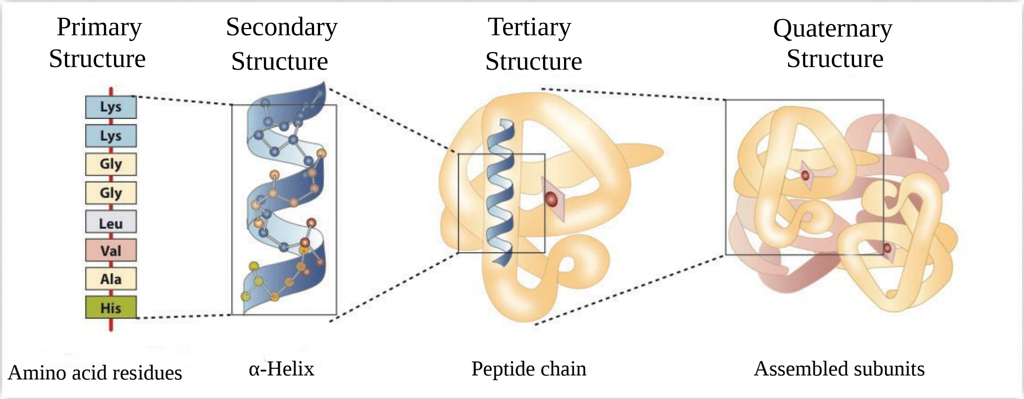


FIGURE 2.5: partial view of four different levels of protein structure in this case Hemoglobin

First level is simply sequence of amino acids for each chain. Next level is called Secondary structure also known as local conformation shows spatial arrangement of adjacent residues in primary structure. Tertiary structure shows three-dimensional view of a folded chain and finally quaternary is arranged view of assembled chains in space[3].

1.2.1 Primary structure

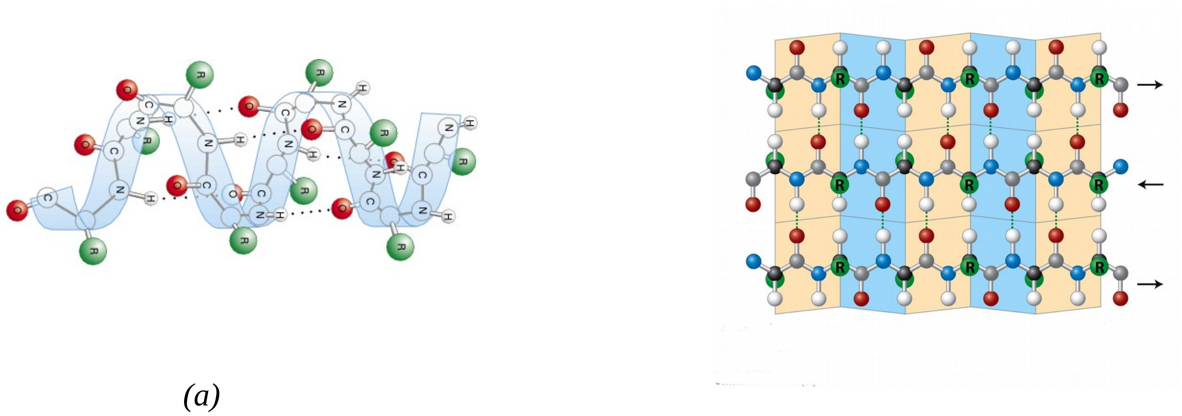
Primary structures are source of rich information to study proteins, already millions of protein sequences from different tissues are available. Protein sequences are widely used to study evolution and homology in proteins. Amino acid composition determines higher level of protein structure like secondary and tertiary structure, this is the reason there are so much study and computational methods which try to predict protein secondary and higher structure from sequence alone[5]. In a peptide chain two terms are widely used: N-terminal and C-terminal which represent two ends of a chain, the one end which has free amino group is N-terminal and the other end in which there is free carboxyl-group is C-terminal.

Protein primary structure is available in different online databases, one of the reliable websites is uniporot: http://www.uniprot.org/. Uniprot stands for universal protein resource, with comprehensive information about protein sequences and their annotation. In most of the databases of proteins, sequences are represented as a text file called FASTA, their simple and easily understandable format make them ideal for programmers to extract sequence information[8].

1.2.2 Secondary structure

Secondary structures are stable and prominent local conformations of primary structure of a protein chain which show the geometrical positioning of contiguous residues. Generally a distinctive regular and ordered arrangement is observed in neighbor residues, the most common and observed structures are α-helices, β-sheets and random coils. The main force that create them is hydrogen bonds in position of polar chemical groups such as the C=O and N-H atoms in the peptide bond[4].

The most stable arrangement is α helix, in which backbone of the chain wound around a imaginary horizontal axis and side chains (R groups) of the residues stand out from the ribbon shape of the back bone.(figure 1.6) .



(a)

(b)

FIGURE 1.6(a): α-helix secondary structure of protein chains , R-group side chains protrude outward, hydrogen bonds showed with dots give the structure considerable stability (b)zig-zag shape of β-sheet conformation from the top , the hydrogen bonds are shown as dots in neighboring chains

Another type of secondary structure which is widely observed is β-sheets which form a zig-zag shape the hydrogen bonds cross-links between adjacent chains residues of adjacent chains.(figure 1.6)

Other type of secondary structure are known as names like loop or turns, these conformations links the segments of β-sheet and α-helices and in general they are formed in locations when chain change its direction or simply turns. There are different type of turns with different length and in their shape in most of the time is irregular, one of the most common type is known as β-turns which connects β-sheet segments, they mostly are observed at the surface of proteins. Turns let protein structure have flexibility or plasticity which is very important in formation of tertiary structure [5].

1.2.3 Tertiary Structure

Tertiary structure demonstrates the overall view of folded protein that The polypetide chain spontaneously bend into highly stable twisted shape. physiochemical properties of the residues and weak covalent interactions between them and solvent lead to the final three-dimensional structure. Unlike protein secondary structure which focuses on local arrangement of neighbor residues, tertiary structure consider the whole shape of the chain, it involve exact location of molecules of the backbone in three dimensional[6]. This three dimensional has key role in many protein function and evolution studies as well as protein interactions. There are some secrets or scientific wonders about these sophisticated structures.

First in 1972 Christian Anfinsen showed that the folding process is autonomous, so it does not require any additional factors or input of energy. This important fact was rewarded by noble prize of chemistry. Base on Anfinsen discovery, protein fording is spontaneous process[1]. Second fact is if protein degrade and denature to unfolded chain by an unfavorable environment(like rise in temperature), it can reversibly gain its first structure if the environment become appropriate again. Third , in folding process protein sequence gain maximum stability or lowest energy level in thermodynamic point of view. Now the main question is how proteins get such a stability? So far we demonstrated that how hydrogen bonds arrange the secondary structure, for tertiary structure also many intra-molecular non-covalent and covalent interactions shape the form of protein 3D structure. The non-covalent interactions known as hydrogen bonds, ionic, hydrophobic, and van der Waals interactions are much weaker than covalent bonds however they are highly important and effective. Here we describe some of important interactions and their manner for more profound understanding of the protein tertiary and quaternary structure.

- Disulfide bonds: Cysteine is polar uncharged residue and its polarity is due to sulfhydryl side chain, the side chain has ability to oxidized to form a covalent bond with another sulfhydryl side chain of cystein (figure 1.7), The dimerc residues which linked by disulfide bond has high hydrophobic behavior even the cystein is polar and hydrophilic. The structure of proteins are so influenced by disulfide bonds, they can link different part of the chain or two chains in protein[1].

- Hydrophobic interactions: Are the forces that hold together the non-polar segments of the protein. They stabilize protein by minimizing the number of water molecules required to surround hydrophobic portions and put the more hydrophobic portions in interior part of the macromolecule. The high polar solvent increase the strength of hydrophobic interactions[1].

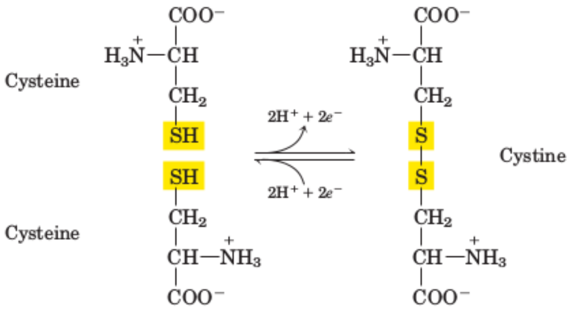


FIGURE 1.7: formation of disulfide bonds between two cystein residue which highly stabilize protein structure

- Salt bridge: When two oppositely charged residue from an ion pair, it is called as salt bridge, negatively and positively charged residue groups (refer to figure 1.2) participate in this type of interaction[1].
- Van der Waals interactions: Random movement of the electrons around atoms creates transient electric dipole in side chain of the residues, therefore two dipoles with opposite electric charge weakly attract each other and get closer, these is referred as Van der Waals interactions[1].

1.2.4 Quaternary structure

When proteins have more than one chain (generally they have), the arrangement of the chains in 3-dimensional space is called quaternary structure, the chains (which are called subunits) interact and link to each other and form a aggregate complex structures. All the mentioned intra molecular interaction like Hydrophobic-interactions, Disulfide bonds, salt bridges, van der Waals interactions forms the final stabilized structure. At the end of protein structure definitions there are some simple definitions on highest level of structure in different point of view:

- **Hetero-dimer vs homo-dimer:** Subunits (chains) in quaternary can be identical known as homo-dimers or non identical known as hetero-dimers. When the subunits are the same, symmetry is observed in protein structure and repetitive subunits can be superimposed on each other with geometrical rotation. However in some proteins the symmetric repetitive subunits consist of more than one chain which is called protomers[6].

- **Fibrous vs Globular proteins**: Fibrous proteins consist of repeating single type secondary structure elements which give them high flexibility and strength, Keratin and collagen are two popular examples. On the other hand globular contain several type of secondary structure with the compact forms of the twisted chain which proteins can have high range and diversity in their shapes. Globular proteins include enzymes, transport proteins, motor proteins, regulatory proteins and proteins with many other functions. This comparison can be the first clear sample of how protein structure and protein function are related[9].

- **Motif vs Domain**: Motifs in general means repetitive patterns, in protein structure, it is some highly observable patterns in topology of secondary structures They are found in four main class : All α-helix, All β-sheet, α + β in which α-helices and β-sheets separated in different parts of the poly-peptide chain and α/β in which α-helices and β-sheets intervals are together[6].

Domains are distinctive structural part pf the proteins which have visible different fold and carry a specific function, they are known as independent bold segments which they can perform they function even if they split from the rest of protein. The domains are highly repetitive which means many proteins can have similar domains[7].

1.3 Protein function and evolution

Knowing the three-dimensional structure of a protein is an important part of understanding how the protein functions. Domains and motifs are fundamental terms in explaining of evolutionary of proteins, All levels of biology undergo evolution and proteins as well, The study of molecular evolution generally focuses on families of closely related proteins[9]. Many examples of recurring domain or motif structures are available, and these reveal that protein tertiary structure is more reliably conserved than primary sequence[3]. The comparison of protein structures can thus provide much information about evolution. Proteins with significant primary sequence similarity, and/or with demonstrably similar structure and function, are said to be in the same protein family. A strong evolutionary relationship is usually evident within a protein family[9].

1.4 Proteins on the Web

The knowledge about biological data is very far from complete. Nevertheless, it is of impressive size and it is constantly and rapidly growing. For this reason, information about biological molecules is generally collected into integrated databases publicity available through the World Wide Web.

1.4.1 Database of protein structures

The Protein Data bank website, http://www.rcsb.org/ , is one of the most reliable archive of 3D view of proteins with many visualization and searching tools, The high reliability of the website is because all the macromolecular structures are derived from X-ray diffraction and NMR studies. Each structure is assigned a unique four-character identifier or PDB ID. The data from the PDB files provide only a series of coordinates detailing the location of atoms and their connectivity. Viewing the images requires easy-to-use graphics programs such as RasMol and Chime that convert the coordinates into an image and allow the viewer to manipulate the structure in three dimensions. The PDB website has instructions for downloading other viewers[10].

1.4.2 Database of protein structural classification

The SCOP put proteins in the correct evolutionary framework based on conserved structural features. Two similar enterprises, the CATH (class, architecture, topology and homologous super-family) and FSSP (fold classification based on structure-structure alignment of proteins) databases, make use of more automated methods and can provide additional information. Structural motifs become especially important in defining protein families and super-families. Improved classification and comparison systems for proteins lead inevitably to the elucidation of new functional relation ships. Given the central role of proteins in living systems, these structural comparisons can help illuminate every aspect of biochemistry, from the evolution of individual proteins to the evolutionary history of complete metabolic pathways[9].

1.4.3 Database of classification of protein function

The Gene Ontology Consortium has produced a systematic classification of gene function, in the form of a dictionary of terms and their relationships. Organizing concepts of the Gene Ontology project include three categories[6]:

- Molecular function: A function associated with what a protein from the biochemical point of view.
- Biological process: A component of the activities of a living system, mediated by other macro-molecules from the cell’s point of view.
- Cellular component: The assignment of site of activity or partners.

Chapter 2

2 Protein Protein Interactions

Proteins interact with other proteins, DNA, RNA and small molecules to perform their cellular tasks. Protein interactions tell us how proteins come together to construct metabolic and signaling pathways in order to fulfill their functions[26]. The most reliable methods to determining protein interfaces are X-ray crystallography and mutagenesis. These techniques are expensive and time consuming[20]. Therefore, over the past 25 years, there has been a rapid development of computational methods aiming protein interaction prediction. However, one should keep in mind that computational methods will only be better and cover more interactions with the help of reliable experimental data. Its is believed that a collective effort between the experiments and computations can make it possible to have a near complete set of interactions[19,20,21].

2.1 PPI Definition

The first step needed is to define precisely what protein–protein interactions. Commonly they are understood as physical contacts with molecular docking between proteins that occur in a cell or in a living organism in vivo[25]. The physical contact considered in PPIs should be specific, not just all proteins that bump into each other by chance. It also should exclude interactions that a protein experiences when it is being made, folded, quality checked or degraded.

2.2 PPI Types and characteristics

Protein−protein interaction types are diverse ranging from transient or permanent non-obligate interactions to obligate interactions. Different types of complexes with specific functions can be observed. We first define the reasons that why protein interactions is a challenging task and what makes their interaction explanation difficult:

- Proteins are dynamic: they move and and have a limited flexibility and vibration which affect their interaction to other molecules[24,26].
- they have ordered and disordered regions: disordered regions are unstructured amino acid compositions that cannot provide a stable folded structure[27]. The proteins with disordered regions can bind to several different proteins by adapting a conformation compatible with partner proteins[28].
- Context dependency: Another essential element for defining PPIs is the biological context. Not all possible interactions will occur in any cell at any time. Instead, interactions depend on cell type, cell cycle phase and state, developmental stage, environmental conditions[25].
- conformational changes after interactions: Many protein structures undergo conformational changes on binding another protein. This means that the use of features derived from static molecular structures may not be enough to describe potential interacting surfaces where conformational flexibility plays a key role. This additional aspect introduces a degree of complexity that is very difficult to take into account with computational methods[14,15].

In the following we will have a review on different interaction types ans basic definitions, and finally a representation of existing methods and applications.

2.2.1 Homo-oligomeric and hetero-oligomeric complexes

This type of classification is straightforward and we have already explained in protein quaternary structure. If the proteins in a complex are identical (interactions occurring between identical protein chains), they form a homo-oligomer, whereas if the PPI takes place among nonidentical chains then it forms a hetero-oligomer. Homo-oligomers are mostly symmetric and provide a good scaffold for stable macromolecules. The stability of hetero-oligomers, on the other hand, varies[29].

2.2.2 Obligate and non-obligate complexes

In order to classify interactions as obligate/non-obligate, one needs to know the affinity and stability of the proteins in the complex and monomeric states, proteins (monomers) of a complex are unstable on their own in vivo then this is an obligate interaction, whereas the components of non-obligate interactions can exist independently[29,30]. Obligate interactions are named as two-state folders. Protein components fold and bind at the same time to form stable complexes. The individual proteins cannot exist as stable, folded structures, but they are stable in the complex form. The components of the non-obligate interactions are three-state folders; they first fold and then come together to form the complex. Most of the stable machineries in the cell are examples of obligate complexes[26].

2.2.3 Transient and permanent complexes

Protein interactions can be classified based on the lifetime of the complex. This classification is relevant only to non-obligatory interactions. Permanent interactions are usually very stable; once two proteins interact they permanently stay as a complex[29]. Transient interactions associate and dissociate temporarily in vivo. Binding between hormone−receptors, signal transduction, inhibition of proteases and chaperone-assisted protein folding are examples of transient interactions. These types of interactions dominate signaling and regulatory pathways as they provide a mechanism for the cell to quickly respond to extracellular stimuli and relay the signals when needed[26].

2.3 Physiochemical properties of PPI sites

As it has been mentioned before, the interactions between proteins occur via inter facial patches which are located on the surface of the proteins. Hence, it is required to have an exact definition of surface region, base on th fact that Interacting residues are subset of surface residues, this definition highly affects the accuracy of predictions. Considering the studies done up to now, Relative Solvent Accessibility (RSA) which is defined as exposure of a residue to a solvent, can be an index to identify surface residues. In the samples with RSA>16% they can be considered as surface one. The threshold ranges from 5% to 16%, where the higher the value of threshold, the lower the number of surface residues. Having said that, this is important to mention that all the information of a residue in this study is limited CA, for simplification and avoiding overloaded complexity[20,24].

Physiochemical properties of protein−protein interfaces include structural and chemical-properties. These should be examined to understand the nature of the intermolecular interactions. the shape of the binding interface, complementarity of two binding-sites, types of secondary structures and energy distribution in the interface regions are some of the properties of binding sites[31].

2.4 Interacting residue definition and features

The most general definition for interacting residue is as follows: “Two residues are in contact if the elucidation distance of their CA atoms are less than a defined threshold” which is distance-based definition. Furthermore, there is an other method, called “surface-based”, which is based on the measurement of conformational changes of residues after interaction. This can be done by measuring Accessible Surface Area (ASA) of residues in unbound condition and comparing with ASA of protein complex (bound state). If the difference of ASA is higher than fixed threshold, are considered as interacting residues[15,26,32]**.**

Comparing interacting and noninteracting residues reveals that they have different behaviors which make them distinguishable. Many studies that already have been performed on different types of datasets are divided into different categories base on the type of properties of protein information chosen for prediction like intrinsic methods which uses sequence or structure properties of the proteins. Other methods like template base approaches use protein fold information or templates comes out from multiple sequence alignment (MSA), partner specific methods target specific type of interactions like anti-body antigen. Most of the predictors use intrinsic methods that cover many properties applicable for sophisticated computations. We list some important residue features that are used in many predictors[20,35]:

- Evolutionarily conserved residues using MSA(Multiple Sequence Alignment)
- Simple predicted RSA using SABLE [20,34], the stand-alone version with default parameters.

- The difference in solvent accessibility (dSA), in an unbound structure between the predicted accessibility with SABLE and observed accessibility calculated with DSSP [20,35].
- Electrostatic potential, extracted from the STING server [20,36].
- Residue interface propensity, for each of the 20 amino acids based on the training set calculated as a fraction that each surface amino acid contributed to the interface compared to the fraction that each amino acid contributed to the whole protein surface [20,37].
- Hydrophobicity taken from the amino acid Index database [20,38].
- Depth and protrusion index, interface residues tend to have a higher average depth index(are more deeply buried), while a higher side chain protrusion(leading to the observed increase in solvent accessible surface area) [43].

2.5 PPI Predictors

An interface residue predictor receives as input a set of proteins then predicts a subset of residues on the proteins surface that are involved in intermolecular interactions. When comparing the true interacting residues with the prediction, it is standard to calculate the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) (Chapter4.3). These four values give rise to a variety of performance metrics can be used to assess the quality of the predictor.

The field of protein–protein interface prediction has classified into many different approaches. Methods might use intrinsic features of the sequence or the structure, evolutionary relationships or use an existing complex as a reference template. Predictors make use of many distinct quality measures, different training and testing data sets, thus a fair comparison between them is difficult [20,33].

2.5.1 Intrinsic-based predictors

2.5.1.1 Sequence-based interface predictors

Sequence-based interface predictors use features from primary structure of the query proteins to recognize interfaces and thus, can be applied to almost any protein. They can be very helpful because of large available databases of protein sequences compared to structures. Interface residues are more conserved than the rest of the protein surface [17,18] and these conserved positions are identified from multiple sequence alignments (MSAs) often with phylogenetic trees assisting the procedure[5, 18,19]. Most common features in these predictors are hydrophobicity distribution, composition propensity to be an interface residue and physico-chemical properties [16,17].

2.5.1.2 Structure-based predictors

Structural features are important discriminative attributes for protein interface prediction. These features are associated with the atomic coordinate of the proteins, such as secondary structure [24,25], solvent-accessible surface area [26,27], geometric shape of the protein surface [26] and crystallographic B-factor [24]. In recent years increasing number of solved structures has improved the performance of 3D-based interface predictors. In these predictors, the query 3D structure is either used to identify interface residues in close proximity to each other. Usage of 3D structural and evolutionary properties tends to improve results over predictions based on sequence alone. It appears that feature-based methods have reached saturation, and the inclusion of more properties does not improve predictive performance[20].

2.5.2 Template-based predictors

Interfaces are conserved among homologous complexes [39], motivating the first category of template-based methods, which relies on homologous complexes. However such homologous structures are not always available. Therefore the second category of template-based predictors uses structurally, but not necessarily evolutionarily, similar complex templates. In this category, the main features that contribute to the quality of predictions are the structure-based MSAs and the binding partner information. Although homologous template-based predictors improve the predictions over intrinsic based methods, they are limited to those proteins where homologous complex structures exist.

2.5.3 Partner-specific interface predictors

The methods discussed before predict interfaces for one sequence protein, but proteins may show variant interface patterns due to their binding partner (e.g. antibodies [40]). Therefore, partner-specific predictors identify interacting residue pairs between two query proteins that are assumed to interact. Two major groups of partner-specific methods are **intrinsic-based methods**, **docking-based methods** predictors.

- **Intrinsic-based methods** uses set of features from 3D structural (or their com-bination with sequence features) that is being computed for training and testing consid-ering partner-specific features such as propensities and electrostatic complementarity [20,35,40].

- **protein–protein docking** predicts potential interfaces from docked poses of the two query proteins and detect interfaces based on contact energy and frequency scores (for a review on docking see [41,42]).

Software	method	Reference
PSIVER	Structure base	[51]
PINUP	Structure base	[52]
ProMate	Structure base	[54]
RAD-T	Structure base	[59]
ISIS	Structure base	[50]
Cons-PPISP	Structure base	[53]
HomPPI	Template-based	[55]
IBIS	Template-based	[56]
T-PIP	Template-based	[57]
PredUS	Template-based	[58]
PrISE	Template-based	[60]
PAIRpred	Partner-specific-Intrinsic-based method	[61]
DoBi	Partner-specific-Docking	[62]
RCF	Partner-specific -Docking	[63]

Table 2.1 : Sample predictors for methods described in features section

2.6 Databases of protein interactions

Publicly accessible databases of protein–protein interactions greatly simplify the analysis of various types of data on protein interactions. Several databases that are currently available (Table 2.2) provide access to both experimental data and the results of diverse computational methods of inference. Some databases also identify the most reliable subsets of the interaction data. Further development of interaction databases is crucial for standardization of the interaction datasets and data exchange formats, as well as for the integration of the data-bases with other bioinformatics resources[23].

Database	URL	Reference
DIP	*http://dip.doe-mbi.ucla.edu*	[42]
BIND	*http://www.bind.ca*	[43]
MINT	*http://cbm.bio.uniroma2.it/mint*	[44]
MIPS	*http://mips.gsf.de*	[45]
The GRID	*http://biodata.mshri.on.ca/grid/servlet/Index*	[23]
LiveDIP	*http://dip.doc-mbi.ucla.edu/ldip.html*	[46]
PREDICTOME	*http://predictome.bu.edu*	[47]
STRING	*http://www.bork.emblheidelberg.de/STRING*	[48]
InterDOM	*http://InterDom.lit.org.sg*	[49]

Table 2.1 : Databases of protein interactions

Chapter 3

3 Probabilistic methods for sequence analysis

Over a century ago relation of biology and computation sounds impossible, now a days they are completely mixed and a form a interdisciplinary science called computational biology, this modern science has variety of applications in all levels of biology by using the data analytical methods to perform pattern recognition, classify data to categories, simulators to mirror behavior of living system, construction of networks of interactions, evolution trees and etc, all the samples highly need computational/mathematical approaches[22].

For Protein interaction prediction we address "sequence labeling problem" in which there are two chains of sequences for each protein the fist one is residues of protein sequence known as input data and each residues of input data is presented by a vector of features, we simply call them *X*. Second there is a "label" for each residue shows if it is interacting or non interacting shortly by "*I*" or "*N*", this is output data and we call it *Y*, in two classes of Interacting and non-Interacting. As we see sequence labeling is actually a type of classification. Input data is relational and has two characteristics: first: statistical dependencies exist between the entities we wish to model, second: each entity often has a rich set of features that can aid classification [64]. All these facts lead us to use supervised machine learning methods.

Supervised machine learning is a type of automation phenomenon which we can see the steps in figure 3.1. All the process can be summarized in two steps "*training*" and "*testing*", training is a type of automation process in which the parameters of the model is fixed by the learning algorithm from labeled input data. So the model can be used to perform classification (or assigning label) to new unseen input data, this process is prediction step. The detail of the steps can be clarified better if we continue with a real model .

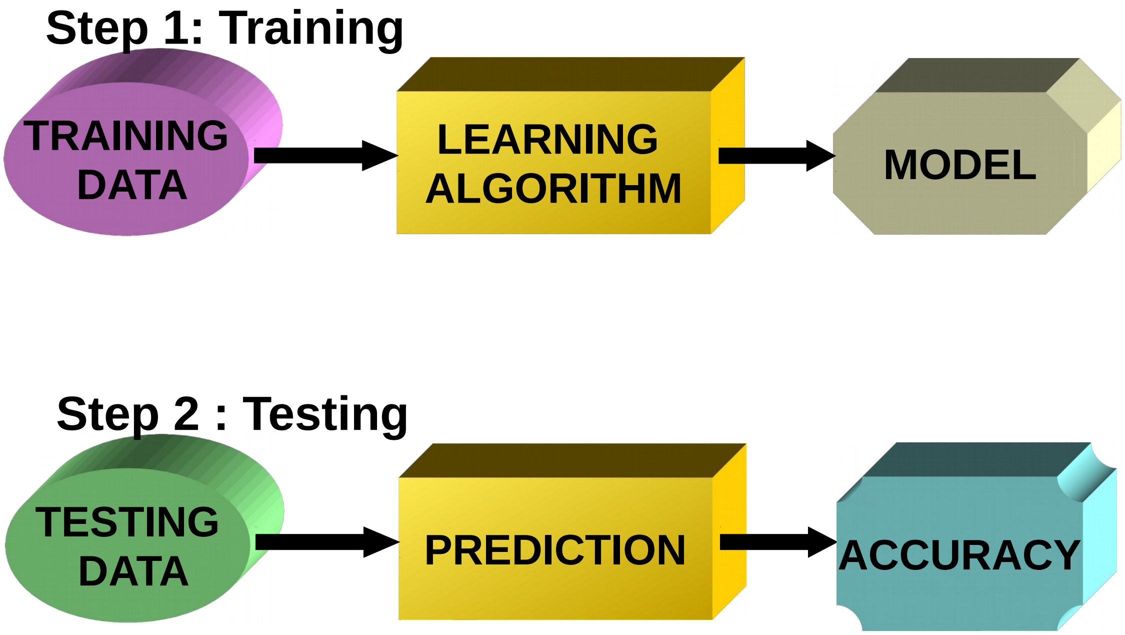


FIGURE 3.1: Schematic view of supervised machine learning , it can be summarized to two steps training and testing

3.1 Probabilistic methods

Probabilistic methods are effective tools building accurate models for large amount of data when there is high degree of uncertainty and lack of theory. One important framework of probabilistic methods is graphical models which graphically represent probability distributions on a set of variables. In particular, when there is high variation in variables data, graphical models can depict the model effectively by reducing complexity. Each node in the graph is associated to a random variable in the model while edges encode dependencies between variables. As a whole, the graph provides a complete description of the model in terms of random variables and probabilistic relationships that exist between them[65]. Probabilistic relationships can be explained by Bayesian framework and generally are divided to two main category Generative models and Discriminative models. The framework for both models is to provide an effective tool to build accurate models based on available data and to reason about them.

3.1.1 Generative models

Generative models are presented by joint probability distribution, also known as naive bayes rule, which explained by simple formula $p(x,y) = p(x|y).p(y) = p(y|x).p(x)$ for a single variable. These models learn of input x and label y, and for prediction it assign the probable label y by calculation of $p(y|x)$. In graphical model the joint probability should be applied on set of variables, thus for a graph with K nodes, the join distribution is[66]:

$$\mathrm{p}(x) = \prod_{k=1}^{K} . \ \mathrm{p}(x_k|y_k) \qquad (3.1)$$

In which y_k are the nodes that x_k is connected to them or there is edge between x_k

and y_k. A well-known methods of this type is *HMM (Hidden Markov Models),* this method involves a chain of states, such as ‘matching to an *I* position in a multiple sequence alignment’, ‘matching to an *N* position’, insertion and deletion. Each state can emit an amino acid from the alphabet of 20 or be silent (like in a deletion state). The chain of states is hidden but the chain of amino acids, i.e. the protein sequence, is observed. The hidden Markov model gives the probability, $p(y_i = I|x)$, that residue *i* within the protein sequence a is in the interface state. To model the joint distribution $p(y,x)$ tractably, an HMM makes two independence assumptions. First, it assumes that each state depends only on its immediate predecessor, that is, each state y_t is independent of all its ancestors $y_1 , y_2 , . . . , y_{t-2}$ given its previous state y_{t-1}. Second, an HMM assumes that each observation variable x_t depends only on the current state y_t. With these assumptions, we can specify an HMM using three probability distributions: first, the distribution $p(y_1)$ over initial states, second, the transition distribution $p(y_t \mid y_{t-1})$ and finally, the observation distribution $p(x_t \mid y_t)$. That is, the joint probability of a state sequence y and an observation sequence x factorizes as[67]:

$$p(y,x) = \prod_{t=1}^{T} . \; p(y_t | y_{t-1})\mathrm{p}(x_t | y_t) \qquad (3.2)$$

This method is off to be named as a good discriminative model because of a problem called “label bias” which is conditional dependencies among y_t , $y_{t-1.}$ Moreover it requires modeling the distribution $p(x)$, which also here complex dependencies should be defined among features of input data. However, we can not name this model ineffective, generative models can perform very well when dependencies among model parameters is clear, protein interaction prediction is not the appropriate case.

3.1.2 Discriminative models

It models *p*(*y*|*x*) directly by direct map of input *x* to *y*. These models properly covers the problem of generative models. There is no need to define *p*(*x)* explicitly therefore, it does not expend modeling effort on the observations, furthermore, the conditional probability of the label sequence can depend on arbitrary, nonindependent features of the observation sequence without forcing the model to account for the distribution of those dependencies. Unlike HMM the probability of a transition between labels may depend not only on the current observation, but also on past and future observations, if available. one of the most popular sample of these models is *conditional random field(CRF) with t*he distribution equation as [64]:

$$p(y|x) = \frac{1}{Z(x)} \exp\{ \sum_{k=1}^{K} . \ \lambda_k f_k (y_t , y_{t-1}, x_t)\} \qquad (3.3)$$

Where λ_k is parameter vector and f_k is a set of real-valued feature functions. Each local feature may depend on all x_t ,this assumption yields a Markovian label sequence. *Z(x)* is an instance-specific normalization function which is[64] :

$$Z(x) = \sum_{y} . \ \exp\{ \sum_{k=1}^{K} . \ \lambda_k f_k (y_t , y_{t-1}, x_t)\} \qquad (3.4)$$

Note that the normalization constant *Z(x)* sums over all possible state sequences, an exponentially large number of terms. It can be computed by with dynamic programming algorithms which is able to compute the most likely label sequence.

3.2 Grammatical-Restrained Hidden Conditional Random Fields

Even though CRF has applied broadly to address sequence labeling problem, it is possible to enhance their performance by including hidden variable into the model which are usually not directly observable at training time[68].

Another limitation related to the fully observability of CRFs models is that they are not well-suited for those sequence analysis problems which can be successfully addressed only by designing a regular grammar in order to provide meaningful results. These problems typically arise in Computational Biology. The training sets generally consist of pair of observed and label sequences and a few subset of the labels are compatible to the grammar requirements. Fully-observable models such as CRFs perform inefficiently in this case. On the contrary, in HMMs it is possible to model a huge number of concurring paths compatible with the grammar and with the experimental labels without increasing the time and space computational complexity.

The two mentioned constraints motivate us to introduce a deviation of CRF method inspired by HMM called: Grammatical-Restrained Hidden Conditional Random Fields (GRH-CRFs), in which we can benefit all the capabilities of CRF plus definition of grammar in the labels, it means base on our prior knowledge we know that there is some constraints in labels can be defined as regular rules to avoid the results that are irrelevant or biologically point of view impossible. The model can not learn these grammars in training process from observation dataset because the reasons of generating labels are hidden states (like HMM), hence the advantages of both discriminative and generative models are integrated[69].

The idea that might be supposed here is that grammars can be defined in CRF method too, if observed sequence relabeled to achieve one to one correspondence between states and labels. This idea is true but it limits the path to be unique for each sequence. In GRH-CRF there is no one to one correspondence between states and labels therefore there would be a huge variant paths for each observed sequence. For optimal extract of path information there are algorithms base on the dynamic programming that efficiently recognize the most probable path among all the possible path having generated the observed sequence. Two well-known examples of these algorithms are Viterbi and Posterior-Viterbi and the most probable path is called “Viterbi-Path”[68,69].

3.2.1 GRH-CRF formulation

We address the problem of mapping an observation sequence $x = (x_1 , . . . , x_L)$ to a label sequence $y = (y_1 , . . . , y_L)$. we also introduce hidden state sequences $h = (h_1 , . . . , h_L)$. Each label in the sequence belongs to a finite set Y of possible labels. Similarly, hidden states are members of a set of possible states $\mathcal{H}$. Furthermore, labels are associated to disjoint sets of states. In other words, each state $h \in \mathcal{H}$ belongs to a subset $\mathcal{H}_\mathrm{y} \subset \mathcal{H}$ of states associated to the label $y \in Y$ and we have:

$$\mathcal{H} = \bigcup_{y \in Y} \mathcal{H}_\mathrm{y} \qquad (3.5)$$

We maintain the sets $\mathcal{H}y$ disjoint in order to keep the inference problem in the model tractable. We want to restrict our model so that a sequence of hidden states is allowed only if it is in agreement with a regular grammar $\mathcal{G}$. We define grammar constraints as follows:

$$\Gamma(h,h')=\begin{cases}1 & \text{if the transition } h' \longrightarrow h \text{ is allowed by } \mathcal{G} \\ 0 & \text{otherwise}\end{cases} \qquad (3.6)$$

We define potential functions which take into consideration grammatical constraints:

$$\Psi_j(h_j , h_{j-1} , y_j , x) = \exp\{ \sum_{k=1}^{K} . \ \lambda_k f_k(h_j , h_{j-1}, x) \} . \Gamma(h_j, h_{j-1}) . 1_{\{hj \in \mathcal{H}yj\}} \qquad (3.7)$$

Where, using the indicator function 1 $\{h_{\text{j}} \in \mathcal{H}_{yj}\}$ we also require that the state at time j is compatible with the corresponding label. The probability distribution of label sequence given an observable sequence $p(y|x)$ can be defined including hidden states as follows:

$$P(y|x) = \frac{p(\mathrm{y} , \mathrm{h}|\mathrm{x})}{p(\mathrm{h}|\mathrm{y} , \mathrm{x})} \qquad (3.8)$$

The joint probability of a label sequence y and an hidden-state sequence h given an observation sequence x is:

$$\mathrm{p}(y , h|x) = \frac{\prod_{\mathrm{j}=1}^{\mathrm{L}} \Psi_j(h_j , h_{j-1} , y_j , x)}{\mathrm{Z}(x)} \qquad (3.9)$$

where, in analogy with HCRFs, *Z(x)* is a partition function obtained summing over all possible label and states sequences as follows:

$$\mathrm{Z}(x) = \sum_y \sum_h \prod_{\mathrm{j}=1}^{\mathrm{L}} \Psi \ (h_j , h_{j-1} , y_j , x) \qquad (3.10)$$

The probability of an hidden-state sequence given a label sequence and an observation sequence is:

$$p(h|y\ ,\ x) = \frac{p(y\ ,\ h|x)}{p(y|x)} = \frac{\prod_{j=1}^{L} \Psi\ (h_j\ ,\ h_{j-1}\ ,\ y_j\ ,\ x)}{Z(y\ ,x)} \qquad (3.11)$$

where the partition function *Z(y, x)* is obtained by keeping fixed the label sequence and summing over all possible corresponding state sequences as follows:

$$Z(y\ ,\ x) = \sum_{h} \prod_{j=1}^{L} \Psi_j\,(h_j\ ,\ h_{j-1}\ ,\ y_j\ ,\ x) \qquad (3.12)$$

Given the above distributions we formally define the Grammatical-Restrained Hidden CRF:

Definition 3.2.1: *Let x, h and y be random variables over observation, hidden state and label sequences, respectively. Let $\Theta = \{\lambda_k\} \in \mathbb{R}^K$ be a parameter vector and $F = \{f_k(h, h', y, x)\}$ from k=1 to K be a set of real-valued feature functions defined over hidden state pairs, labels and the entire observation. A linear-chain Grammatical-Restrained Hidden Conditional Random Field is a conditional probability distribution of the form:*

$$P(y|x) = \frac{p(y\ ,\ h|x)}{p(h|y\ ,\ x)} = \frac{Z(y\ ,x)}{Z(x)} \qquad (3.13)$$

where *Z(y,x)* and *Z(x)* are instance-specific partition functions as 3.12 and 3.10 formulas.

3.2.2 Parameter estimation

Given a training dataset D = $\{(x^{(i)}\ ,\ y^{(i)})\}$*i=1* to *N* of independent and identically distributed labeled observation sequences, the parameters *Θ* of a GRH-CRF model can be obtained by maximum likelihood estimation. The log-likelihood of data is:

$$\begin{aligned}\ell(\Theta;\mathcal{D}) &= \log\prod_{i=1}^{N} p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}) \\ &= \log\prod_{i=1}^{N}\frac{Z(\mathbf{y}^{(i)},\mathbf{x}^{(i)})}{Z(\mathbf{x}^{(i)})} \\ &= \sum_{i=1}^{N}\log Z(\mathbf{y}^{(i)},\mathbf{x}^{(i)}) - \sum_{i=1}^{N}\log Z(\mathbf{x}^{(i)})\end{aligned} \tag{3.14}$$

Taking the first derivative with respect to parameter λ_k of the objective function we obtain:

$$\frac{\partial\ell(\Theta;\mathcal{D})}{\partial\lambda_k} = \underbrace{\sum_{i=1}^{N}\frac{\partial}{\partial\lambda_k}\log Z(\mathbf{y}^{(i)},\mathbf{x}^{(i)})}_{\mathcal{C}} - \underbrace{\sum_{i=1}^{N}\frac{\partial}{\partial\lambda_k}\log Z(\mathbf{x}^{(i)})}_{\mathcal{F}} \tag{3.15}$$

where, in analogy with the Boltzmann machines and HMMs for labeled sequences [70], C and F can be seen as clamped and free phases. After simple computations we can rewrite the derivative as:

$$\frac{\partial\ell(\Theta;\mathcal{D})}{\partial\lambda_k} = E_{p(\mathbf{h}|\mathbf{y},\mathbf{x})}[f_k] - E_{p(\mathbf{h},\mathbf{y}|\mathbf{x})}[f_k] \tag{3.16}$$

where the $E_{p(h|y,x)}$ [f$_k$] and $E_{p(h,y|x)}$ [f_k] are the expected values of the feature function f_k computed in the clamped and free phases, respectively. To avoid over-fitting, we regularize the objective function using a Gaussian prior, so that the function to maximize has the form of:

$$\ell(\Theta;\mathcal{D}) = \sum_{i=1}^{N}\log Z(\mathbf{y}^{(i)},\mathbf{x}^{(i)}) - \sum_{i=1}^{N}\log Z(\mathbf{x}^{(i)}) - \sum_{k}\frac{{\lambda_k}^2}{2\sigma^2}$$

(3.17)

and the corresponding gradient is:

$$\frac{\partial \ell(\Theta; \mathcal{D})}{\partial \lambda_k} = E_{p(\mathbf{h}|\mathbf{y},\mathbf{x})}[f_k] - E_{p(\mathbf{h},\mathbf{y}|\mathbf{x})}[f_k] - \frac{\lambda_k}{\sigma^2} \qquad (3.18)$$

3.3 Decoding Algorithms

Decoding is the task of assigning labels *y* to an unknown observation sequence *x*. Viterbi algorithm is routinely applied as decoding for the linear-chain CRFs, since it finds the most probable path of an observation sequence given a CRF model [71]:

$$y^* = \underset{y}{\operatorname{argmax}}\ p(y|x) \qquad (3.19)$$

In CRF models with hidden variables, the Viterbi algorithm is used to search the hidden state space rather than the label space. In this context, the algorithm is particular effective when there is a single strong highly probable hidden state path, while when several paths compete (have similar probabilities), posterior decoding may perform significantly better. Whit this approach, for each position the label is assigned according to its posterior probability $p(y_j = y|x)$ as follows:

$$y_j^* = \max_{y} p(y_j = y|x) \qquad (3.20)$$

which is equivalent to:

$$y^* = \underset{y'}{\operatorname{argmax}} \prod_{j=1}^{L} p(y_j = y_j'|x) \qquad (3.21)$$

However, the selected label sequence of the posterior decoding may not be allowed by

the grammar. A simple solution of this problem is provided by the posterior-Viterbi decoding, that was previously introduced for HMMs [72]. Posterior-Viterbi, exploits the posterior probabilities and at the same time preserves the grammatical constraint.
The algorithm works by considering, for any position j in the sequence and label y, the posterior probability of reaching some state h associated with the label y. This probability is given as follows:

$$p(h_j \in \mathcal{H}_y|\mathbf{x}) \;=\; \sum_{h \in \mathcal{H}_y} p(h_j = h|\mathbf{x}) \qquad (3.22)$$

i.e. summing over posterior probabilities p(h_{j} = h|x) of all possible states $h \in H_{\mathrm{y}}$. After this step, at any given position, we have the posterior probabilities of reaching different disjoint regions of the hidden state space that correspond to the different labels. By applying a Viterbi search over these posteriors we can obtain the labeling as in Equation 3.22 and at the same time preserving the grammatical constraints.
The first step can be accomplished using the Forward-Backward algorithm as described for the free phase of parameter estimation. In particular, the posterior probability of a hidden state is obtained as:

$$p(h_j = h|\mathbf{x}) \;=\; \frac{\alpha_j^{\mathcal{F}}(h)\beta_j^{\mathcal{F}}(h)}{Z(\mathbf{x})} \qquad (3.23)$$

Let $\mathrm{M}_{(\mathrm{h,\,j})}$ be the matrix obtained as:

$$M(h,j) \;=\; \sum_{h' \in \mathcal{H}_{\Lambda(h)}} p(h_j = h'|\mathbf{x}) \qquad (3.24)$$

where $\Lambda(h) = y$ is the function that returns for each state the associated label y. The matrix M is such that if h, h' are associated to the same label y then $M(h, j) = M(h', j)$, $\forall j$.

By performing a grammatical constrained Viterbi search over the matrix *M* we can solve the maximization problem in Equation 3.22 We define $\rho_j(h)$ as the product of posterior probabilities of the partial state sequence $h_1, \ldots, h_{j-1}, h_{j=h}$ of length j ending in state h while $\pi_j(h)$ is a traceback pointer. The algorithm proceeds as follows:

1. Initialization:

$$\rho_0(h) = \begin{cases} 1 & \text{if } h = \mathbf{BEGIN} \\ 0 & \text{otherwise} \end{cases} \tag{3.25}$$

2. Recursion:

$$\rho_j(h) = \max_{h'} \rho_{j-1}(h')\Gamma(h', h)M(h, j) \tag{3.26}$$

$$\pi_j(h) = \operatorname*{argmax}_{h'} \rho_{j-1}(h')\Gamma(h', h)M(h, j) \tag{3.27}$$

3. Termination and Traceback:

$$h^*_{n+1} = \mathbf{END} \tag{3.28}$$

$$h^*_j = \pi_{j+1}(h^*_{j+1}) \quad for \quad j = n, n-1, \ldots, 1$$

$$h^*_0 = \mathbf{BEGIN}$$

The labels are assigned to the observed sequence according to the state path h^*.

Chapter 4

4 Implementation of CRFGRF method for protein interaction prediction

In this chapter two new topological model was described whose parameters were discriminatively trained using Grammatical-Restrained Hidden CRFs. The model was trained and tested using 10-fold cross validation on a newly generated dataset derived from benchmark 5.0 proteins. To highlight the advantages of using a GRHCRF formulation, the model was compared with other approaches previously released for PPI prediction. In all the experiments, the model achieved performances comparable or superior to other state-of- the-art methods.

4.1 Dataset

There is high diversity in the available databases in PPI, nevertheless finding a dataset with updated information containing all types of protein interaction types is not easy. Here after a long research on choosing appropriate dataset, Benchmark 5.0 was chosen which has been broadly used for training and testing protein docking algorithms, developing ranking algorithms, formulating energy functions and performing protein structure analysis(*https://zlab.umassmed.edu/benchmark/*)[74,75].

Since 2005 different version of benchmark was released base on this fact that there was high growth in protein structures in protein data bank (PDB). Benchmark is valuable source of protein complexes from different aspects, (1) It covers high range of different interaction types; (2) it contains all the structures in both bound and unbound state with high resolution; (3) non redundant; (4) most of the complexes have experimentally-measured binding affinities and binding free energies [75].

We filtered the dataset using various quality criteria: (1)No sequence redundancy (no sequence of the data set has a sequence identity ≥ 25% to any other sequence of the data set); (2) the complex structure needed to be determined by X-ray crystallography with resolution 3.25 Å or better; (3) the sequences in bound and unbound chains are equal; (4) chains needed to consist of at least 50 residues. Finally we could achieve 609 of non redundant chains with 125922 residues, in this study we refer to out dataset as BM609. BM is abbreviation of benchmark and 609 is number of chains.

We evaluated several properties from the structure files. The change in solvent accessible surface area (ΔASA) upon complex formation was calculated by DSSP. DSSP is a program which is more known for secondary structure assignment, moreover it can measure solvent exposure of proteins if the atomic coordinates in Protein Data Bank format is provided. Therefor we run DSSP two times for each of the chains in bound and unbound state and we marked the residues on the surface which their solvent exposure (ASA) has increased.

As we mentioned in chapter 2 the residues with RSA value higher than a fixed threshold are defined as surface area, here we consider residues RSA> 16% as surface(BM_609_16), as we know if we lower this threshold the higher part of the protein will be considered as surface area so we also defines a data set with RSA>5% (BM_609_5) to see how it affect our prediction results.

After marking interacting residues on the surface we extracted two features for each residue from the dataset. we call the first feature interacting neighbors :

we imagine a sphere with radix of 12 Å around each residue and we mark the residues on that area which are on the surface (figure 4.1), the window size limits the number of the closest neighbors therefore at the end we have have a list of closest residues, the length of the list is the window size and the items are ordered by their distance.

Before going further we mention two assumptions :

- For each protein in the dataset, a profile based on a multiple sequence alignment was created using the PSI-BLAST program on the non-redundant dataset of sequences. PSI-BLAST runs were performed using a fixed number of cycles set to 3 and an e-value of 0.001, therefore in the list of contacting residues, we put the corresponding profile vector value.
- residue actually means Cα atom and the distance is euclidean value of two Cα-Cα atom

For the second feature, as we mentioned in chapter 2 one of the discriminating feature is: The difference in solvent accessibility (dSA), in an unbound structure between the predicted accessibility with SABLE and observed accessibility calculated with DSSP [20,35].The SABLE server can be used for predicting real valued relative **S**olvent **A**ccessi**B**i**L**iti**E**s of amino acid residues in proteins.We calculated the average of the difference value dSA for neighbor residues came out from the first feature. RSA straightforwardly characterizes the local environment of residues in protein structures and identifies surface, interior and interface regions in proteins, therefore the normalized value of dSA for each residue considering the structural neighbors can be an effective scale to determine the ability of a residue to interact with other residues.

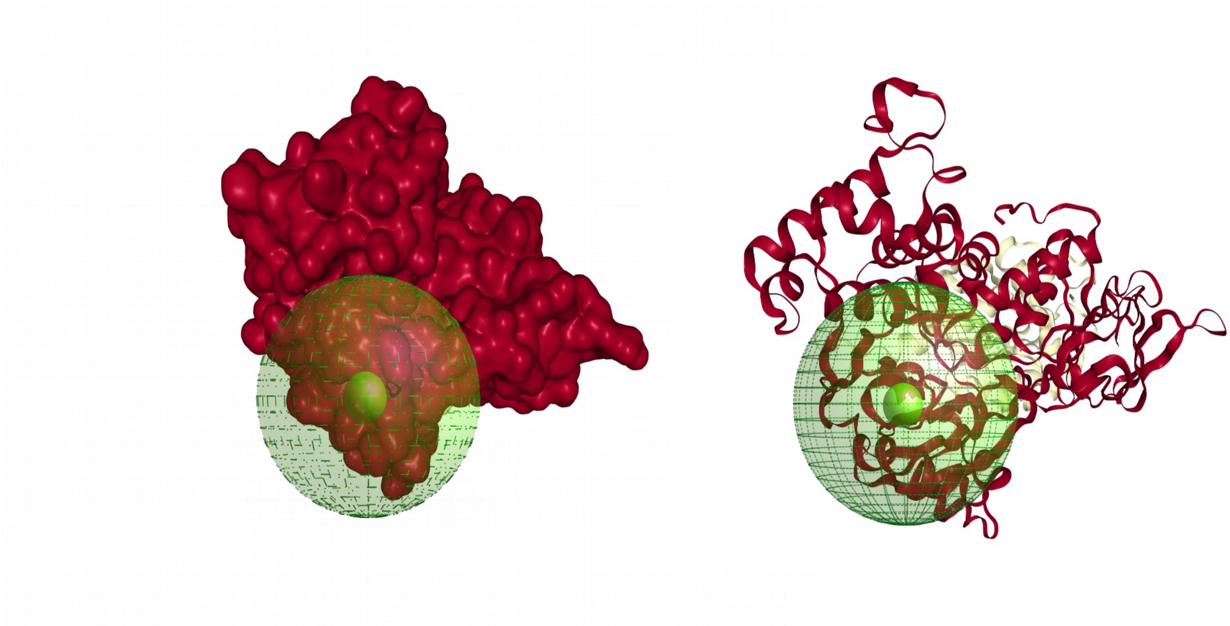


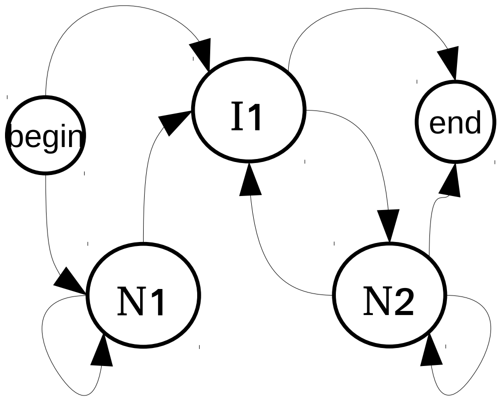
FIGURE 4.1 : A sphere with 12 Angestrom radix is depicted around each residue and we mark the residues on that area which are on the surface

4.2 Grammar definition

The topology of two grammars is depicted by directed graphs in figure (4.2). Each node represents the label state and their dependency to other states is shown by edges. Grammar A has only 3 states and grammar B covers 11 states. The topology is constructed base on statistical analyses on interacting and non interaction labels of training set. This is the powerful point of this method which can only emphasize on patterns and relations on interacting and non interacting labels of training set without relevance to biology. Beside this studies on biological rules are welcomed and applicable by this method however in this moment more emphasize is on mathematical relations.

Both of the grammars defines rules for 2 label states Interacting and non-interacting, as statistics shows only 15% of residues are interacting in training set and most of the time they are surrounded by non interacting residues, grammar A tries to more capture short islands of interacting sites scattered around chains while Grammar B tries to capture longer interacting sites in proteins that have a few but long interfaces.

Grammar A:



Grammar B:

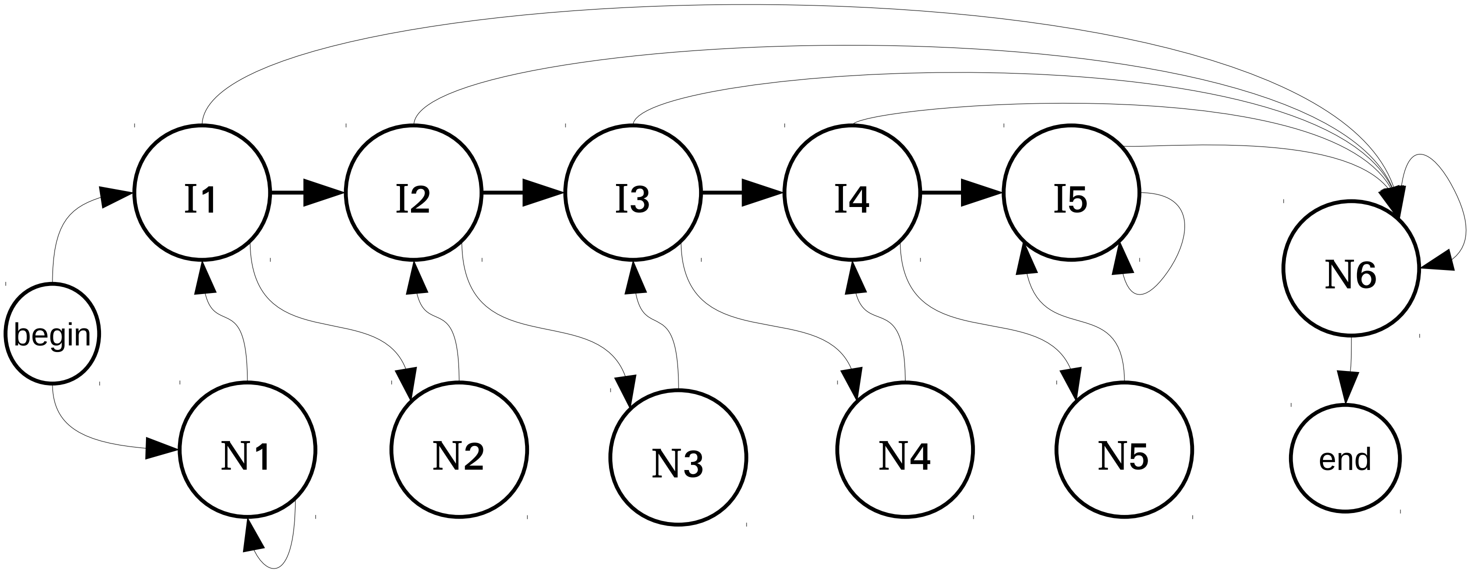


FIGURE 4.2: The graphical view of two grammars defined for interacting residue prediction, grammar A include 3 states while grammar B has 11 states and more complex connections.

4.3 Accuracy measures for PPI residue prediction

4.3.1 10-Fold Cross Validation

Cross Validation is a technique for estimating the performance of a predictive model. In other word, it is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. The training set is split into k smaller sets and the following procedure is followed for each of the k “folds”. First, A model is trained using K-1 of the folds as training data and then the resulting model is validated on the remaining part of the data (i.e., it is used as a test set to compute a performance measure such as accuracy)[20,22].

The performance measure reported by k-fold cross-validation is then the average of the values computed in the loop. The 10-fold cross-validation performed on the BM609 dataset to test the efficiency of both grammars. The original dataset was partitioned into 10 nearly equal subsets with same distribution of interacting and noninteracting and surface residues. In each validation, one subset was used for testing while the rest was used for training.

4.3.2 Fundamental definitions

Predicting interacting residues formulates as a binary classification problem, where each protein residue can be either interacting (I) or non-interacting (N). Evaluation of the classification performance generally considers those cases that are correctly and incorrectly predicted for each class, which is quantified by the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) [75].

- **Sensitivity** (also known as True Positive Rate, Recall or TPR) evaluate the effectiveness of the predictor for each class. TPR measures the fraction of correctly predicted interacting residues:

$$\text{TPR} = \frac{TP}{TP+FN} \qquad (4.1)$$

- **Specificity** (SPC) evaluates the fraction of predicted interacting residues forming an interacting in the experimental complex structure:

  $$\text{SPC} = \frac{TP}{TP+FP} \qquad (4.2)$$

  Specificity in many articles [22,76], introduced as TN/(TN+FP) and formula 4.2 is introduced as Precision (also Positive Predictive Value, PPV) for avoiding exaggerated results because high rate of non-interacting residue we referenced to formula 4.2 as specificity[75].

- **Accuracy** (ACC) evaluates the effectiveness of a predictor by the fraction of correct predictions:

  $$\text{ACC} = \frac{TP+TN}{TP+FP+TN+FN} \qquad (4.3)$$

- **Matthew's Correlation Coefficient** (MCC) is a measure that balances the sensitivity and specificity, evaluating the strength of the correlation between predicted and the actual classes. Its values range from -1 to 1, where 1 corresponds to a perfect prediction, 0 to a random prediction and -1 to a perfectly inverse prediction:

  $$\text{MCC} = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(FP+FN)(TN+FN)}} \qquad (4.4)$$

- **F1:** Measures designed to balance between false negative and positive rates called F1 :

$$F1 = 2 * \frac{SPC.TPR}{SPC+TPR} \qquad (4.5)$$

4.4 Evaluation of performance

One the most important part of prediction is that how we can assess our prediction results, the clear approach is to compare our results with other methods, however it can not be always a perfect way because different methods use different datasets base on different type of protein interactions we have already explained in chapter 2, and different feature definitions . e.g. Table 4.6 shows that how one predictor (HomePPI) can perform differently on different datasets[55].

In table 4.5 we have collected the results of well known predictors on bench mark 4.0 data set which is the closest one to our dataset[76]. For the comparison we more consider MCC value, because ACC can not be a good evaluator due to the low proportion of interacting residues which typically in most of datasets range from 10% to 30% of the surface (In our data set is 15%). With this small fraction, simply by labeling all the residues as non-interacting the high accuracy is achievable. Matthew’s correlation coefficient (MCC) is a better criterion which ends this artifact.

4.4.1 Performance on BM_609_16 and BM_609_5

The final results are reported in four sections in tables (4.1), refer to what we have already explained we have two Datasets with names BM_609_16 and BM_609_5, and for each of them we have applied two grammars A and B with four different window sizes 5, 9, 13 and 17 which determines the number of neighbor residues in feature one explained in 4.1.

The best MCC result is 34% belong to Grammar B for BM_609_16 dataset with window size 17. All the tables shows that with increase of the window size the prediction results improve for both grammar A and B, this is quiet meaningful, the reason is the structural neighbor residues are ordered by distance (as we already explained as first feature in dataset) and when the window size is small the neighbors are more residues that are close in primary structure, though the larger window size can target contacting residues scoped in by tertiary structure.

The two later sections in table 4.1 show the results for other dataset in which the RSA value threshold is decreased to 5% and higher proportion of residues will be considered as surface area, this change slightly decrease the performance around 3% (best MCC result is 31% for Grammar B and window size 17), expanding the surface area can not improve the performance, however in this dataset set also we see the increase of the window size the results has improved, likewise the grammar B has performed better in all four different window sizes and different datasets. Grammar B with considering longer areas of interacting residues can perform better compared to grammar A which more focus on short scattered interface areas.

The results shows low rate for sensitivity and better results for specificity which means the number of predicted interacting residues is not long enough compared to the length of sequence thus the predicted ones was on the right positions. With the growth of window size the specificity does not change however a a gradual improvement happen in sensitivity which means has no cost on specificity and machine has got better performance and this improvement has been reflected in MCC value.

Finally, the decoding phase was performed using the Posterior-Viterbi and Viterbi algorithms we have already explained , Posterior-Viterbi outperformed with a large gap compared to Viterbi so in all tables we just report results decoded from posterior-viterbi.

Dataset		TPR	SPC	ACC	MCC	Time(min)
BM_609_16 Grammar A	Window 17	31%±0.06	64%±0.06	75%±0.04	31%±0.06	500
	Window 13	29%±0.04	64%±0.01	75%±0.02	30%±0.06	400
	Window 9	27%±0.02	63%±0.04	75%±0.04	29%±0.01	300
	Window 5	25%±0.06	63%±0.06	74%±0.01	27%±0.04	200
BM_609_16 Grammar B	Window 17	32%±0.01	67%±0.02	76%±0.04	34%±0.02*	1500
	Window 13	30%±0.07	66%±0.08	76%±0.01	32%±0.06	1300
	Window 9	28%±0.06	66%±0.01	75%±0.06	30%±0.08	1000
	Window 5	25%±0.04	65%±0.06	75%±0.03	29%±0.03	700
BM_609_5 Grammar A	Window 17	25%±0.01	65%±0.08	77%±0.03	30%±0.01	500
	Window 13	23%±0.05	64%±0.07	76%±0.04	28%±0.08	400
	Window 9	21%±0.01	64%±0.04	76%±0.01	26%±0.06	300
	Window 5	18%±0.02	64%±0.07	76%±0.01	24%±0.05	200
BM_609_5 Grammar B	Window 17	27%±0.08	66%±0.03	77%±0.05	31%±0.01	1500
	Window 13	25%±0.03	65%±0.03	77%±0.06	29%±0.04	1300
	Window 9	23%±0.06	65%±0.05	77%±0.05	28%±0.01	1000
	Window 5	18%±0.03	65%±0.08	76%±0.01	25%±0.05	700

Table 4.1 : Performances of grammar A (with 3 states) on dataset BM_609_16, grammar B (with 11 states) on BM_609_16, grammar A on BM_609_5 and grammar B on BM_609_5.

** : the best result is for grammar B on dataset BM_609_16with window size 17*

4.4.2 Comparison with performance of predictors using experimental structures

Among all the predictors and results available on different articles table 4.2 is very valuable [76] because of reporting all assessment values: True Positive Rate, Specificity, False Positive Rate, Accuracy and MCC on benchmark 4.0 dataset which gives us a view of our performance and better assessment.

If we compare our best MCC value with MCC results of other web servers it get the position 4 from 12 web servers that is interestingly good ranking position and comparable with other methods in general. However if we look at the results of Home PPI on different datasets categorized by protein interaction type we see that th MCC value has increased to 65% for obligate homo-dimers dataset.

Dataset	Web server	TPR	SPC*	ACC	MCC
Benchmark 4.0	Pseudo-meta	69%±0.02	41%±0.07	88%±0.07	48%±0.01
	PredUs	70%±0.01	30%±0.02	30%±0.02	38%±0.03
	eFindSite PPI	39%±0.06	45%±0.09	90%±0.05	37%±0.05
	cons-PPISP	27%±0.09	33%±0.08	88%±0.08	24%±0.07
	SPPIDER	34%±0.00	20%±0.08	82%±0.07	17%±0.03
	ProMate	52%±0.06	21%±0.00	68%±0.04	16%±0.05
	WHISCY	13%±0.00	33%±0.04	90%±0.00	16%±0.04
	PIER	06%±0.06	34%±0.02	90%±0.06	11%±0.08
	VORFFIP	53%±0.01	33%±0.07	57%±0.09	11%±0.07
	PSIVER	64%±0.05	11%±0.08	54%±0.06	10%±0.03
	InterProSurf	43%±0.05	16%±0.03	67%±0.07	10%±0.00

Table 4.2 : Performance of different predictors on bench mark 4.0 dataset.

Finally, it is important to keep in mind that even predictors with similar overall performance often disagree about the interface in individual proteins. The true interface is usually in common by one or more predictors. Moreover, some predictors seem to be better suited for some kinds of complexes than others: WHISCY performs better on enzymes than on inhibitors while PIER shows a reverse behavior[73]. This suggests that not only properties but also interface predictors can be complementary and can be combined for improved prediction of protein interfaces.

Dataset	**TPR**	**ACC**	**MCC**	**Source**
transient enzyme-inhibitor complexes	58.0	85.0	44.0	[55]
transient non-enzyme- inhibitor complexes	48.0	84.0	42.0	
Obligate hetero-dimers	71.0	86.0	60.0	
Obligate homo-dimers	73.0	91.0	65.0	

Table 4.3: Performance of HomPPI predictor on 4 different datasets which contains different interaction types released on 2011.

Conclusion

Different methods already have been introduced for protein interaction prediction however as we saw it is quiet necessary the new algorithms and methods. The results of GRF-CRF compared to other other method and considering its first version was quiet acceptable however its capacity for improvement is undeniable.

The results of the performance for different datasets showed that definition of the grammars can enhance the performance of the predictors. Moreover even the number features in our case is not high, they had role in improvement of classification power of the algorithm. Additionally, the examined existing methods behave in different manner on different sets of proteins, while our method is able to have stable results given group of proteins. This means that the existing methods are suitable for making predictions for a given group of proteins, but are not able to make general decisions. Our models are trained using various proteins, but if the models are focused on a specific group of proteins, then the prediction power will be much better.

To improve the results in future several decisions can be made: first it is possible to include additional sequence and structural base features, regarding other classifiers, this method has very few features. Second searching for other state of the methods to analyze surface of proteins like Voroni algorithms which consider more geometric partitioning rather than the use of an absolute distance cutoff, can be helpful to have more sophisticated and intelligent grammars.

Bibliography

[1] *Nelson DL, Cox MM (2005). Lehninger's Principles of Biochemistry (4th ed.). New York, New York: W. H. Freeman and Company.*

[2] *Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, Scott MP, Zipurksy SL, Darnell J (2004). Molecular Cell Biology (5th ed.). New York, New York: WH Freeman and Company.*

[3] Creighton, T.E. (1992) Proteins: Structures and Molecular Properties, 2nd edn, W. H. Freeman and Company, New York.

[4] Perticaroli S, Nickels JD, Ehlers G, O'Neill H, Zhang Q, Sokolov AP (October 2013). "Secondary structure and rigidity in model proteins". Soft Matter. ***9*** *(40): 9548–56.* doi:10.1039/C3SM50807B. PMID 26029761

[5] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N and Yeh LS. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32:D115-119 (2004)

[6] Anfinsen CB. Priciples that govern the folding of proteins chain, Science 181, pp. 223-230 (1973).

[7] Bateman A, Birney E, Durbin R, Eddy SR, Howe KL and Sonnhammer ELL. The Pfam Protein Families Database. Nucleic Acids Res. 28:263-266 (2000)

[8] Baldi P and Brunak S. BIOINFORMATICS: The machine learning approach. Second edition. A Bradford Book. The MIT Press (2001).

[9] Thornton J.M., Orengo C.A., Todd A.E., and Pearl F.M.G. (1999) Protein folds, functions and evolution. J Mol Biol 293:333–342.

[10] Bowie, J.U., Lüthy, R., and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170.

[11] Richard P Horgan / Louise C Kenny (2011) 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics 13:189–195.

[12] Theodorescu D, Mischak H. Mass spectrometry based proteomics in urine biomarker discovery. Wld J Urol 2007;25:435–43. doi:10.1007/s00345-007-0206-3

[13] Petricoin E, Zoon K, Kohn E, Barrett J, Liotta L. Clinical proteomics: translating benchside promise into bedside reality. Nat Rev 2002;1:683–95. doi:10.1038/nrd891

[14] Castrense Savojardo 1,2 , Piero Fariselli 1,2 , Damiano Piovesan 1 , Pier Luigi Martelli 1 , and Rita Casadio: Machine-Learning Methods to Predict Protein Interaction Sites in Folded Proteins

[15] Iakes Ezkurdia, Lisa Bartoli, Piero Fariselli, Rita Casadio, Alfonso Valencia and Michael L. Tress : Progress and challenges in predictingprotein protein interaction sites,2009. doi:10.1093/bib/bbp021

[16] Georgina Mirceva , Andrea Kulakov: Improvement of protein binding sites prediction by selecting amino acid residues' features ,2014 Elsevier

[17] Govindarajan Sudha , Ruth Nussinov , Narayanaswamy Srinivasan : An overview of recent advances in structural bioinformatics of proteineprotein interactions and a guide to their principles 2014 Elsevier

[18] Sjoerd J. de Vries and Alexandre M.J.J. Bonvin: How Proteins Get in Touch: Interface Prediction in the Study of Biomolecular Complexes, 2008, 9, 394-406

[19] Huan-Xiang Zhou , and Sanbo Qin :Interaction-site prediction for protein complexes: a critical assessment ,2007 , doi:10.1093/bioinformatics/btm323

[20] Reyhaneh Esmaielbeiki, Konrad Krawczyk, Bernhard Knapp, Jean-Christophe Nebel and Charlotte M. Deane: Progress and challenges in predicting protein interfaces, 2015,doi: 10.1093/ bib/bbv027

[21] Matthew Z. Tien , Austin G. Meyer, Dariya K. Sydykova , Stephanie J. Spielman , Claus O. Wilke : Maximum Allowed Solvent Accessibilites of Residues in Proteins, 2013, doi: 10.1371 /journal.pone. 0080635

[22] Tristan T Aumentado-Armstrong , Bogdan Istrate and Robert A Murgita : Algorithmic approaches to protein-protein interaction site prediction, 2015,doi: 10.1186/s13015-015-0033-9

[23] Lukasz Salwinski and David Eisenberg: Computational methods of analysis of protein–protein interactions, 2003, doi: 10.1016/S0959-440X(03)00070-8

[24] Javier De Las Rivas and Fontanillo: Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks ,2014, 10.1371/journal.pcbi.1000807

[25] Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.;Romero, P.: Intrinsically Disordered Protein. J. Mol. Graphics Modell. 2001, 19, 26−59.

[26] Ozlem Keskin, Nurcan Tuncbag and Attila Gursoy :Predicting Protein−Protein Interactions from the Molecular to the Proteome Level , 2016 , doi: 10.1021/acs.chemrev.5b00683

[27] Mackay JP, Sunde M, Lowry JA, Crossley M, Matthews JM (2007) Protein interactions: is seeing believing? Trends Biochem Sci 32: 530–531.

[28] Chatr-Aryamontri A, Ceol A, Licata L,Cesareni G (2008) Protein interactions: integration leads to belief. Trends Biochem Sci 33: 241–242; author reply 242–243.

[29] Pogoryelov, D.; Krah, A.; Langer, J. D.; Yildiz, O.; Faraldo Gomez, J. D.; Meier, T. Microscopic Rotary Mechanism of Ion Translocation in the F(O) Complex of Atp Synthases. Nat. Chem. Biol. 2010, 6, 891−899.

[30] Levy, E. D.; Teichmann, S. Structural, Evolutionary, and Assembly Principles of Protein Oligomerization. Prog. Mol. Biol. Transl 2013, 117, 25−51.

[31] Bogan, A. A.; Thorn, K. S. Anatomy of Hot Spots in Protein Interfaces. J. Mol. Biol. 1998, 280, 1−9.

[32] Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997;272: 133–43.

[33] Ezkurdia I, Bartoli L, Fariselli P, et al. Progress and challenges in predicting protein-protein interaction sites. Brief Bioinform 2009;10:233–46.

[34] Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. Prot Struct Func Bioinform 2004;56:753–67.

[35] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–637.

[36] Bradford JR, Westhead DR. Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics 2005;21:1487–94.

[37] Neshich G, Borro LC, Higa RH, et al. The diamond STING server. Nucl Acids Res 2005;33:W29–35.

[38] Kawashima S, Ogata H, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res 1999;27.

[39] Ma B, Elkayam T, Wolfson H, et al. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc NatlAcad Sci 2003;100:5772–7

[40] Ahmad S, Mizuguchi K. Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. PLoS One 2011;6:e29104.

[41] Rodrigues JP, Bonvin AM. Integrative computational modeling of protein interactions. FEBS J 2014;281:1988–2003.

[42] Xenarios I, Salwinski L, Duan XQJ, Higney P, Kim SM, Eisenberg DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 2002, 30:303-305.

[43] Bader GD, Donaldson I, Wolting C, Ouellette BFF, Pawson T, Hogue CWV: BIND - The biomolecular interaction network database. Nucleic Acids Res 2001, 29:242-245.

[44] Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: MINT: a Molecular INTeraction database. FEBS Lett 2002, 513:135-140.

[45] Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B:MIPS: a database for genomes and protein sequences. Nucleic Acids Res 2002, 30:31-34.

[46] Duan XJ, Xenarios I, Eisenberg D: Describing biological protein interactions in terms of protein states and state transitions: the LiveDIP database. Mol Cell Proteomics 2002, 1:104-116.

[47] Mellor JC, Yanai I, Clodfelter KH, Mintseris J, DeLisi C: Predictome: a database of putative functional links between proteins. Nucleic Acids Res 2002, 30:306-309.

[48] von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B: STRING: a database of predicted functional associations between proteins. Nucleic Acids Res 2003, 31:258-261.

[49] Ng SK, Zhang Z, Tan SH, Lin K: InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. Nucleic Acids Res 2003,31:251-254.

[50] Ofran Y, Rost B. ISIS: interaction sites identified from sequence. Bioinformatics 2007;23:e13–16.

[51] Murakami Y, Mizuguchi K. Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. Bioinformatics 2010;26:1841–8.

[52] Liang S, Zhang C, Liu S, et al. Protein binding site prediction using an empirical scoring function. Nucleic Acids Res 2006;34:3698–707.

[53] Chen H, Zhou HX. Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. Proteins 2005;61:21–35.

[54] Neuvirth H, Raz R, Schreiber G, et al. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol 2004;338:181.

[55] Xue LC, Dobbs D, Honavar V. HomPPI: a class of sequence homology based protein-protein interface prediction methods. BMC Bioinformatics 2011;12:244.

[56] Shoemaker BA, Zhang D, Thangudu RR, et al. Inferred Biomolecular Interaction Server—a web server to analyze and predict protein interacting partners and binding sites. Nucleic Acids Res 2010;38:D518–24.

[57] Esmaielbeiki R, Nebel J-C. Scoring docking conformations using predicted protein interfaces.

BMC Bioinformatics 2014;15:171.

[58] Zhang QC, Deng L, Fisher M, et al. PredUs: a web server for predicting protein interfaces using structural neighbors.Nucleic Acids Res 2011;39:W283–7.

[59] Bendell CJ, Liu S, Aumentado-Armstrong T, et al. Transient protein-protein interface prediction: datasets, features, zalgorithms, and the RAD-T predictor. BMC Bioinformatics 2014;15:82

[60] Jordan RA, Yasser ELM, Dobbs D, et al. Predicting proteinprotein interface residues using local surface structural similarity. BMC Bioinformatics 2012;13:41.

[61] Minhas A, ul Amir F, Geiss BJ, et al. PAIRpred: partnerspecific prediction of interacting residues from sequence and structure. Proteins 2014;82:1142–55.

[62] Guo F, Li S, Wang L, et al. Protein-protein binding site identification by enumerating the configurations. BMC Bioinformatics 2012;13:158.

[63] Hwang H, Vreven T, Weng Z. Binding interface prediction by combining protein–protein docking results. Proteins 2014;82:57–66.

[64] Charles Sutton, Andrew McCallum. An Introduction to Conditional Random Fields for Relational Learning

[65] Bishop, C. M. Pattern Recognition and Machine Learning (InformationScience and Statistics), 1st ed. 2006. corr. 2nd printing ed. Springer, Oct. 2007.

[66] Pearl, J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, 1 ed. Morgan Kaufmann, Sept. 1988.

[67] Roman Klinger , Katrin Tomanek ,Classical Probabilistic Models and Conditional Random Fields

[68] Fariselli, P., Savojardo, C., Martelli, P., Casadio, R., et al. Grammatical-restrained hidden conditional random fields for bioinformatics applications. Algorithms for Molecular Biology 4, 1 (2009), 13.

[69] Piero Fariselli, Castrense Savojardo, Pier Luigi Martelli and Rita Casadio , Grammatical-Restrained Hidden Conditional Random Fields for Bioinformatics applications , 2009 , doi: 10.1186/1748-7188-4-13*

[70] Krogh, A. Hidden Markov Models for Labeled Sequences. In In Proceedings of the 12th IAPR ICPR'94 (1994), IEEE Computer Society Press, pp. 140–144.

[71] Lafferty, J., McCallum, A., and Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of ICML01 (2001), pp. 282–289.

[72] Fariselli, P., Martelli, P., and Casadio, R. A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-betamembrane proteins. BMC Bioinformatics 6(Suppl.4), S12 (2005).

[73] Calem J Bendell , Shalon Liu , Tristan Aumentado-Armstrong , Bogdan Istrate , Paul T Cernek , Samuel Khan , Sergiu Picioreanu , Michael Zhao and Robert A Murgita ,Transient protein-protein interface prediction: datasets, features, algorithms, and the RAD-T predictor ,2015

[74] Thom Vreven 1 , Iain H. Moal , Anna Vangone , Brian G. Pierce , Panagiotis L.Kastritis , Mieczyslaw Torchala , Raphael Chaleil , Brian Jiménez-García , Paul A.Bates , Juan Fernandez-Recio , Alexandre M.J.J. Bonvin , and Zhiping Weng :Updates to the integrated protein-protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2, doi:10.1016/j.jmb.2015.07.016

[75] Bin Liu , Xiaolong Wang , Lei Lin , Buzhou Tang , Qiwen Dong and Xuan Wang: Prediction of

protein binding sites in protein structures using hidden Markov support vector machine, 2009 , doi:10.1186/1471-2105-10-381

[76] Surabhi Maheshwari and Michal Brylinski , Predicting protein interface residues using easily accessible on-line resources, 2015 , doi: 10.1093/bib/bbv009