Chapter 12
Using Structural and Physical–Chemical Parameters to Identify, Classify, and Predict Functional Districts in Proteins—The Role of Electrostatic Potential

Goran Neshich, Izabella Agostinho Pena Neshich, Fabio Moraes, Jose Augusto Salim, Luiz Borro, Inacio Henrique Yano, Ivan Mazoni, Jose Gilberto Jardine and Walter Rocchia

Abstract In this chapter, we will overview the role of the local protein structure environment (which we will call here: nano-environment) in maintaining the functional purpose of different protein districts (defined as protein structure sites delimited by their functional objectives). Namely, we suggest that the local environment at each protein point and/or region reflects, not only its constitutional/structural role, but also its contribution to providing necessary and required characteristics for the functional objective that such particular site is supposed to have. For instance, protein–protein communication is executed through protein interfaces, and amino acid residues belonging to that site must have some specific characteristics which do not only differentiate them from the free surface residues, but also make possible that two very specific proteins may engage, bind and by doing so, perform their function. Similarly, enzyme function is normally related to activity of its catalytic site residues (CSRs). Obviously, these very peculiar residues are embedded in a very specific nano-environment (defined also by the contribution of CSR). Consequently, the enzyme function could be described in terms of characteristics of the CSRs and their surroundings. Based on the above considerations, and assuming that the local nano-environment is not only defining the protein district function, but it is also a concept for which we can design specific metrics to quantify it, and a specific set of properties to describe it, we studied the role of different descriptors and found that, together with hydrophobicity, electrostatic potential is of fundamental importance. As we will better detail in the course of this work, the electrostatic potential might

G. Neshich (✉) · I.H. Yano · I. Mazoni · J.G. Jardine
Embrapa Agricultural Informatics, Campinas, Brazil
e-mail: Goran.Neshich@embrapa.br

I.A. Pena Neshich · J.A. Salim · L. Borro
Unicamp, Campinas, Brazil

F. Moraes
UNESP, Sao Jose do Rio Preto, Brazil

W. Rocchia
CONCEPT Lab—CompuNet, Istituto Italiano di Tecnologia, Genova, Italy

W. Rocchia and M. Spagnuolo (eds.), *Computational Electrostatics for Biological Applications*, DOI 10.1007/978-3-319-12211-3_12

not always be the top ranked property defining the nano-environment of interest, but it is, however, always present, contributing significantly in carving proper protein district characteristics for specific structure/function purposes.

Keywords Protein structure · Protein districts · Structure–function relationship · Nano-environment · Physical–chemical properties · STING database · Electrostatic potential · Protein interfaces · Protein specificity · Secondary structure elements · Catalytic site residues · Hydrophobic effect

12.1 Introduction

Java Protein Dossier [1] is a concept database and visualization tool for protein structures. JPD is a part of the STING [2, 3] platform, which provides one of the most comprehensive collections [4] of physical–chemical parameters describing protein sequence, structure, stability, function, and interaction with other macromolecules. Coupled to the JPD, STING's relational database (STING_RDB) [5] contains hundreds of protein descriptors calculated for all structures deposited in the PDB such as electrostatic potential, contacts energy, density, hydrophobicity, and many more.

Electrostatics has been shown to have a fundamental role in regulating interactions between biological macromolecules, such as proteins and nucleic acids [6, 7]. Among other aspects, the electrostatic contribution to the (de)solvation process has proved to be remarkably important in many biological phenomena. For many applications, that contribution is modeled as a dielectric linear response to the electric field generated by the charge borne by the biomolecular system. Consistent with this model, the Poisson–Boltzmann equation (PBE) has proved to be able to provide quantitative estimates of the electrostatic interaction energy of biomolecules [8].

In addition to more than 1,300 other descriptors, Sting's Java Protein Dossier and relational database (STING_RDB) encompasses also the description of some electrostatic features, providing the numerical value of the mean electrostatic potential (EP) at each residue and at some relevant atoms, as well as the potential over the molecular surface. In STING, the EP value is calculated on a per atom basis and then reported for all eligible PDB files in a residue by residue fashion. Four precalculated categories are shown: (1) EP at the alpha carbon atom, (2) EP value at the last heavy atom of any residue side chain (LHA), (3) average EP value over all amino acid atoms, and (4) EP value averaged over the patch of the molecular surface that is attributable to that particular amino acid. The complete description of calculations employed in order to solve the Poisson–Boltzmann equation for biomolecules is given in [9].

In this chapter, we analyze the quantitative and qualitative assessment of the role that electrostatic potential has on protein structure–function relationship and, in particular, its role in defining nano-environment characteristics of functional protein districts. Protein districts considered in this analysis are: protein–protein interfaces

(PPI), catalytic site residues (CSR), binding site residues (BSR or interface-forming residues: IFR), and secondary structure elements (SSE).

12.2 Nano-Environment Characteristics for Specific Protein Districts

Having the entire set of amino acid residue properties previously calculated and stored in the STING_RDB [5], we are ready to obtain a description of the nano-environment of protein districts as complete as it is currently possible. Mostly because of the fact that the BlueStar STING offers easy access to a very rich repository of protein characteristics, the STING platform [4, 10–12] has already been used for predicting enzyme class [13], protein–ligand analysis [14, 15], protein mutant analysis [16, 17], protein–protein interaction pattern analysis [18] as well as in research linked to some specific biological problems [19, 20].

We can explore the properties of a nano-environment using a simple method that is both self-explanatory and intuitive. To understand it, imagine that we can insert an imaginary probe anywhere inside a protein structure and obtain back a report describing characteristics of the environment in which the probe was embedded. Obviously, we cannot physically do this in the real world, and therefore the probe needs to be substituted with the calculation of values, metrics, and forces we desire to quantify at each particular point/site. This approach somewhat resembles, but with a different focus, that of the GRID method for the calculation of molecular interaction fields in drug design.

The advantage of this approach is that any amino acid residue, or any of its side or main chain atoms, could serve as the center for the probe and from that particular point, the interplay of all forces might be estimated, cataloged, and stored into an appropriate database—in our case the STING_RDB. Once stored, the attribute values could be mapped back to the protein structure for visual inspection or used in statistical/numerical analysis.

Our assumption is that any specific environment is fine-tuned for its function and therefore can be identified, parameterized, and classified accordingly. If one were to consider, for example, protein contact interfaces, it could be expected that such specific areas of the protein, occupying part of its surface, have characteristics sufficiently different from the ones built by amino acid residues found at non-interacting surface areas. In fact, we consider such assumption being in line with the biological requirements for performing a specific function; the function in this example being communication with a very specific partner protein. So, this protein district or functional region, as we name it, is described by precise attributes and their values, making possible not only to distinguish it from the rest of the protein structure but also predicting the district coordinates in other proteins that have not been characterized chemically/biologically.

Similarly, the nano-environment within which CSR are acting must be very specific for each protein family (or if we would like to be more precise, subfamilies defined up to and including the third digit in the EC number {for example: 3.4.21.x}). Such peculiarity of nano-environments for CSR is also intuitively expected because similar, or better, the same chemical reactions need specific conditions to operate on diverse substrates. The identification and classification of the CSR nano-environment provides a fundamental tool for predicting the enzyme class of those proteins whose structure has been deciphered but for which no experimental data exists to identify their biological function and activity. As it is well known, each year more and more protein structures with no known function are deposited in the PDB [21] creating a very strong demand for computationally based enzyme classification methods.

In addition to interface and CSR nano-environments, we will also address here the environment of binding residues and of secondary structure elements, because those environments are also expected to be very specific and, therefore, potentially useful both for classification and prediction purposes.

In all cases of protein district nano-environments, the electrostatic potential plays a crucial role and its relevance needs to be contrasted with other protein structure attributes/properties.

Procedures

Proper procedures for data collection and analysis had to be designed in order to maximize the volume of data, eliminate redundancy, and to ensure we could operate with independent protein structure descriptors. Some of the data preparation procedures we have used are briefly described here.

In order to ensure a proper analysis, we needed carefully designed data sets to collect protein structures that could provide useful information on relevant nano-environment characteristics, (such as the electrostatic potential or surface hydrophobicity index (SHI) in some specific protein structure districts).

For the nano-environment analysis of catalytic site districts, members of protein families and subfamilies differing among themselves only at the fourth EC number (x.y.z.*) were assembled in datamarts, which were additionally filtered with regard to their sequence similarity. The sequence similarity threshold used in this case was 40 %. Properties of the active site were then checked against those of the rest of the protein to identify significant variations that could clearly distinguish the nano-environment.

To analyze protein–protein interactions, we first identified in the PDB all protein–protein complexes and then we added several filters to select the most informative ones. These filters were defined in eight consecutive layers as described below.

This work started in 2010 and was divided in a number of projects, executed by members of our lab. All results presented in this chapter were collected before December 2012 and data completely analyzed before June 2013. The version of the PDB that we have used for the initial dataset selection contained protein molecules available until November 8, 2010. We downloaded from the PDB ftp site [22, 23] a total of 165.720 chains in 68.997 PDB files. This initial ensemble of structures was used as the starting point, providing the original material for the subsequent restraint

guided selection that would eventually result in the final working dataset, which we will refer to here as the "DS95" data set.

The first filtering layer consisted in selecting only those structures obtained by X-ray diffraction (NMR structures were not considered). The second layer consisted in using only PDB files that contained only protein chains (i.e., protein–DNA and protein–RNA complexes were excluded from the analysis). The third layer consisted in using PDB files (asymmetric units) that contained exactly the number of chains that EBI PISA version 1.18 [24] indicated as the correct oligomeric state. The fourth layer selected only PDB files with at least two protein chains. The fifth kept only structures with X-ray resolution better or equal to 3 Å. The sixth layer actually consisted of two subfilters: the first one eliminated all PDB files containing protein chains with less than 50 amino acids, and the second one excluded all complexes having an interface with an area of less than 200 $Å^2$ (as calculated by the SurfV program [25]). The seventh layer eliminated all PDB files containing incomplete proteins: for example, the ligand-binding domain from the AMPA subtype Glutamate receptor (263 residues *per* chain) is available in 3KGC in its dimeric form, but it does not correspond to the real, full protein length nor does it represent its real oligomeric state; in this case, a better representation of the complex is available in PDB entry 3KG2, which contains the full-length AMPA subtype Glutamate receptor as an homo-tetramer having 823 residues in each chain. We decided to remove structures of incomplete proteins using sequence information from UniProtKB [26]. Sequences were retrieved from UniProtKB in FASTA format and the relevant details retrieved from the sequence header. The PDBSWS database—PDB/UniProt Mapping was used to relate identifiers of the UniProtKB to their counterparts in PDB [27]. The eighth and last filtering layers consisted in removing sequence redundancy: this was done using PDB clusters [28], specifically Cluster_95 [29]. The resulting final data set, subsequently denominated DS95, ended up containing a total of 6931 non-redundant chains from 6192 PDB files. The above-described multilayer procedure was mostly automatized (except for some manual inspections [to be described below]), providing necessary robustness in application and fast results when demand for repetitive filtering was identified.

An additional feature was considered during data analysis (although not as a selective step): we annotated PDB structures that contained chains factually proven to belong to membrane proteins. This information was derived from the PDB TM [30] and MPtopo [31] databases, and helped us identify 119 distinct chains (from 65 PDB files) corresponding to membrane spanning proteins.

It is important to mention that, in spite of such a rigorous selection, we were still able to identify in DS95 some protein chains that were actually fragments, and also some structures where the oligomeric state was different between PISA and PDB. Additional manual curation was required to eliminate those PDB entries as well.

In addition to DS95, we also prepared DS100, DS70, and DS30, using the corresponding clusters provided by [29]. The respective numbers of chains and PDB entries making up the protein complexes for the mentioned datasets were: DS100–9009 chains from 8082 PDB entries; DS95–6931 chains from 6132 PDB entries; DS70–6368 chains from 5743 PDB entries, and finally, DS30–4605 chains from 4219 PDB entries. The reason for building four data sets was to have as complete

information as possible on how the data would change by successive elimination of similar (sequence-wise) proteins. As it turns out, DS95 proved to be in many ways the most representative dataset for our goals and was used to analyze protein–protein interfaces, including considerations about the hydrophobic effect being the principal driving force for protein binding, with the electrostatic interactions providing complementary binding energy.

The nano-environment of secondary structure elements was studied using a different approach, with a dataset consisting of various datamarts. In this case, we first created datamarts containing proteins with: (A) only alpha helical elements (with turns but no beta pleated sheets present), (B) only beta sheets present (with turns but no alpha helices present), (C) both alpha helices and beta sheets present (as well as turns), and (D) no regular secondary structure elements (unstructured or partially structured proteins). The definition of SSE was established by requiring a consensus between the definitions provided by the Stride and DSSP algorithms and the definition provided in the PDB file itself.

Our goal here was to single out any significant variation in the average values of structure/physical/chemical properties that were specific of each nano-environment. To this end, we compared the same attribute values between the investigated nano-environment and the rest of the protein structure, looking for any significant variation that would be clearly in evidence. When this variation occurred simultaneously for a number of descriptors, a "composite signal" would in fact be assembled, being characteristic of only one given SSE type.

12.2.1 Protein Function and Catalytic Site Residues

Enzymes perform their biological role using some specific amino acids known as catalytic site residues (CSR). Thus, the function and taxonomy of a particular enzyme can be obtained indirectly through the differentiation of its CSR from the rest of the protein amino acids, followed by a comparison of their observed properties with known and cataloged evidence about CSRs in other enzyme families.

We hypothesized that the catalytic reactions performed by enzymes must depend on the physicochemical properties of the nano-environment around the CSR. Based on this conjecture, we have proposed a method for the characterization and prediction of CSR using structural protein descriptors from STING_RDB. In particular, this database provides helpful information about the physicochemical properties shared by CSR for a variety of enzyme families.

The goal of our investigation was to characterize the common elements of the nano-environment surrounding the CSR's (based on their physicochemical properties) by identifying, analyzing, and finally presenting comprehensible rules for selecting only the CSR, extracted from STING_RDB of structural protein descriptors.

The enzyme structures available from PDB were separated according to their EC numbers and their CSR were labeled according to the annotation found in the

Catalytic Site Atlas [32]. Sequence redundancy (up to 40 % maximum) of proteins was identified, and enzymes above the threshold were removed from further consideration. Then, the STING's protein structural descriptors were extracted for all amino acids of all selected enzymes. Next, attributes were selected using an adapted evolutionary algorithm called GARIPPER [33] and protein structure descriptors stored in STING_RDB so that they could be delivered as an input to the rule induction algorithm RIPPER [34]. In this way, we were able to obtain "human comprehensible" rule sets for CSR's selection for enzymes belonging to different EC numbers. Sequence conservation [35, 36] parameters were excluded from analysis in order to obtain a fully physicochemical characterization of the CSR's nano-environment. Due to the unbalanced distribution of the two classes of amino acid residues (CSR and non-CSR), some modifications were introduced in GARIPPER to allow for more robust processing. We added techniques for preprocessing data, using under and over sampling methods, into the evolutionary algorithm to achieve a proper selection of the suitable ratio between CSR and non-CSR samples in the training dataset. That modified version of GARIPPER was named as GARIPPEROUS (GARIPPER Over and Under Sampled).

What we noticed immediately before starting the large-scale examination of CSR nano-environments was that a CSR could be specified (selected/separated) uniquely and simply through a set of selection rules based on a list of physicochemical parameters and of corresponding values. Surprisingly, we observed that imposing value constrains on only a few attributes could eliminate all amino acids in a protein but the CSR. Using initially a manual approach, we tested 25 different protein families and all of them gave positive results in terms of obtaining a simple and reduced set of rules for separating CSR from all the other ones. Such sets of rules usually contained from 2 to 7 attributes and the corresponding ranges for their numerical values. In fact, once applied to a single representative of the enzyme subfamily (defined with the first three digits of its corresponding EC number), the filtering procedure would also identify specifically the CSR in other members of that same subfamily (with few exceptions). This fact coupled to the observation that there are definite and precise differences among enzyme subfamilies, prompted us to suggest that it would be possible to build a table of CSR nano-environment characteristics specific for each enzyme family, and that these tables could be later assembled into what we named "the periodic table of enzymes." The name is intended to suggest the specificity of the description used for each enzyme family, albeit there is no expected periodicity in the encountered descriptors or in their numerical values.

In Fig. 12.1, panels a, b, and c, the selection of CSR is illustrated based on applied structural and physicochemical parameter constraints. The procedure reveals how the ensemble of amino acid residues remaining on the visual display of the STING's ^{Java} Protein Dossier becomes smaller as additional parameter constraints are added to the list of previously established ones. At the final stage, only the CSR are shown, and the SELECT procedure is completed. The selected parameters and their numerical values used to obtain a comprehensive (yet minimalistic) description of the nano-environment for a given enzyme function (serine protease: elastase—1PPF)

are presented in Table 12.1. Application of the same (or very similar) constraints to the structure of another member of the same subfamily would bring forward the corresponding (in spatial position and amino acid type) CSR (or a slightly broadened ensemble of residues which includes the CSR) even if the sequences of the two examined proteins were quite dissimilar. This observation shows that the nano-environment of CSR is mostly preserved within the subfamily and is therefore describable by a very similar syntax. As mentioned above, so far we have not found two different subfamilies of enzymes having the same constraints for the parameters describing their CSR nano-environment nor have we found an enzyme subfamily without a corresponding constraint set for filtering its corresponding CSR.

After successfully testing our approach manually, we resorted to an automated machine learning approach. For the automated machine learning approach, we opted

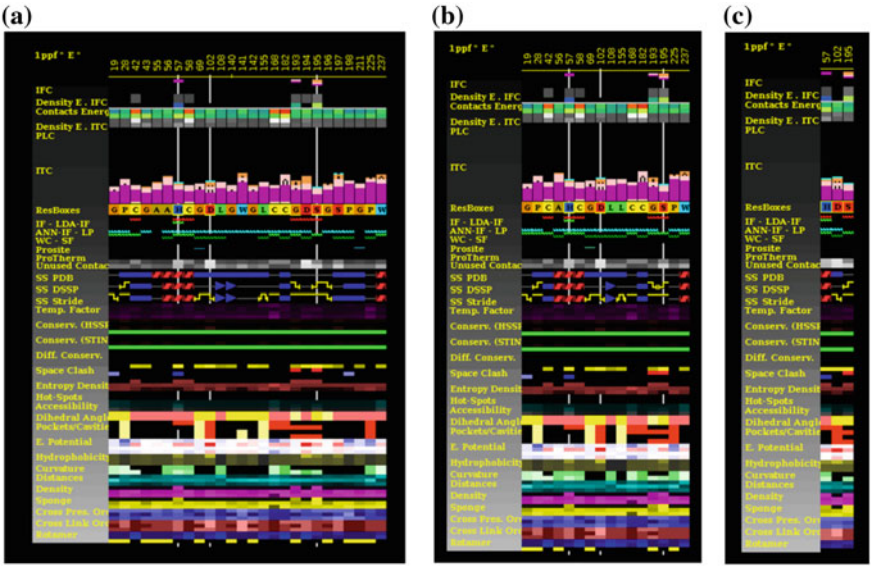

Fig. 12.1 Illustration of the "SELECT" procedure, available under Java Protein Dossier of BlueStar STING. In panel **a** the structure of human leukocyte elastase (EC#: 3.4.21.37—a serine protease) enzyme: 1PPF, contains a total of 218 amino acid residues in the E chain. The CSR ensemble is constituted by the following three residues: His_57, Asp_102 and Ser_195 (sometimes GLy_193 and Gly_196 are also included in the ensemble). In order to eliminate all other amino acid residues of the E chain, we applied a sequence of only three constraints: the first was conservation, measured in relative entropy (RE), with the RE values being restricted to less than or equal to 7 (indicating well-conserved residues). This first constraint eliminated most of the residues as they did not comply with imposed conditions. Only 16 residues (less than 10 % of the initial number) remained, including the CSRs. In panel **b** the second filter was imposed by selecting electrostatic potential, calculated at the surface of the protein, which is created by individual residues, and the range of values for this parameter was selected to be higher than –2 and below 300 kT/e. By applying these filters, only 4 residues complied, including the three which belong to the CSR ensemble. Panel **c** the last filter was the "Number of unused Contacts," being set to higher than 240 (implying a high potential of the CSR to create contacts with spatial neighbors)

Table 12.1 Amino acid residue parameters and their value ranges for section of CSR in 1ppf.pdb

Structure property	Range of values for the property
Conservation (HSSP): relative entropy	≥7
Electrostatic potential at surface	[−2; 300]
Unused contacts	≤240

for using the empirical cumulative distribution functions (ECDF) for EP descriptors averaged over nearest spatial neighbors (Weighted Neighbor Averages—WNA), as described in [37]. The plots shown in Fig. 12.2 indicate the probability of finding

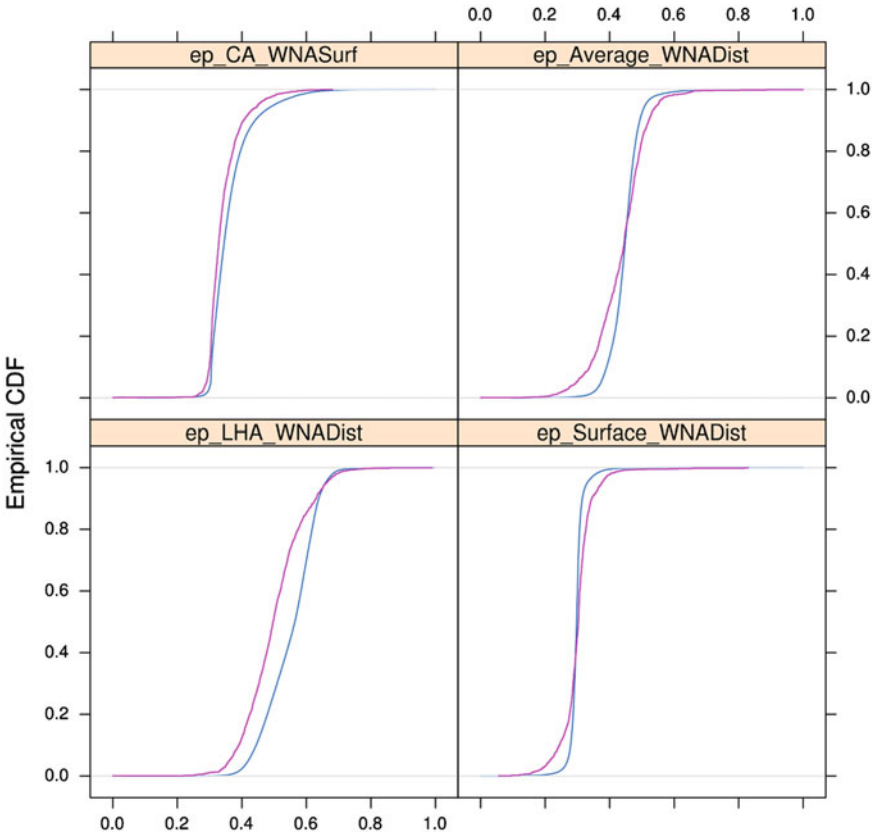

Fig. 12.2 The empirical cumulative distribution functions (ECDF) of electrostatic potential descriptors averaged over nearest spatial neighbors (Weighted Neighbor Averages—WNA) for two ensembles: CSR and non-CSR. Each subplot depicts the difference between the ECDF's of the catalytic residues (*red*) and non-catalytic residues (*blue*). The maximal distance between two curves corresponds to the Komolgorov–Smirnov statistics [38]

a value which is equal or lower than the EP value (x-axis normalized to [0, 1]) in two ensembles (CSR and non-CSR). The existence of a significant ($P < 0.01$) difference between two classes (for catalytic and non-catalytic residues) flags the EP attribute as a descriptor capable of distinguishing between the two. The EP calculated at the last heavy atom (LHA) of the residue side chain has the greatest distance between two distributions and therefore the highest potential for CSR versus non-CSR discrimination.

The machine learning approach was favored in our work as the above-described manual one was not sufficiently robust to be carried out for a large volume of enzymes containing numerous members of certain subfamilies. Even small variations in a set of parameters and/or modifications in range delimiters for their numerical values might create problems, which are insurmountable for the manual approach. Consequently, machine learning was employed and all sequence-wise, non-redundant members of enzyme subfamilies were analyzed aiming to obtain general sets of rules for the description of their CSR nano-environments. As in the manual approach described above, the EP continued being one of the relevant constraints, but in spite of the predictive power it has for distinguishing CSR from non-CSR (particularly in case of EP calculated around last heavy atom (LHA) in the side chain—see Fig. 12.2), it was actually found missing as the top ranked attribute in the final list of constraints selected for filtration in machine learning approaches.

12.2.2 Enzyme Specificity and Binding Site

Enzymes belonging to the same super family of proteins, in general, operate on a variety of substrates and are inhibited by a wide selection of inhibitors. In this part of our work, the main objective was to expand the scope of studies that consider only the catalytic site amino acids while analyzing enzyme specificity and, instead, include a wider category, which we have named the interface-forming residues (IFR). We wanted to identify those IFRs (characterized primarily by their decreased accessibility to solvent after docking of different types of inhibitors to, in this case study, subclasses of serine proteases) and then create a table (matrix) of all amino acid positions at the interface as well as their respective occupancies and characteristics. Our goal was to establish a platform for analysis of the relationship between IFR characteristics (their nano-environment) and binding properties/specificity for bimolecular complexes.

As a result of that effort, we have proposed a novel method for describing binding properties and delineating the specificity of serine proteases by compiling an exhaustive table of interface-forming residues (IFR) for serine proteases and their inhibitors. As the Protein Data Bank (PDB) does not contain all the data that our analysis required, an in silico approach was designed for building the corresponding complexes. The IFRs were obtained by “rigid body docking” among 70 structurally aligned, sequence-wise non-redundant, serine protease structures with three inhibitors: bovine pancreatic trypsin inhibitor (BPTI), ecotine, and ovomucoid third

domain inhibitor. Then, we created a table (matrix) of all amino acid positions at the interface and their respective occupancy. We also developed a new computational protocol for predicting IFRs for those complexes, which were not deciphered experimentally so far, achieving accuracy of at least 97 %. Details of those experiments are described in [39].

In the context of this book chapter, the conclusions that we reached regarding enzyme specificity were that the interfaces of serine proteases prefer polar (but including also glycine) residues (with some exceptions) (see Fig. 12.3). Thus, the IFR pocket of serine proteases is not formed by predominantly hydrophobic residues; it is a rather polar environment. The surfaces (not including interface areas) have a prevalence of charged residues. However, charged residues were found to be uniquely prevalent at the interfaces between the "miscellaneous-virus" subfamily of serine proteases and the three inhibitors. This prompted some speculations about how important this difference in IFR characteristics is for maintaining virulence of those organisms and significance of the electrostatic interaction in considering the molecular aspects of infectious processes.

Such description of the interface-forming residues (IFRs) provides a unique tool for both structure/function relationship analysis as well as a compilation of indicators detailing how the specificity of various serine proteases may have been achieved and/or could be altered. It also indicates that the interface-forming residues which also determine specificity of the serine protease subfamily cannot be presented in a canonical way but rather as a matrix of alternative populations of amino acids within respective nano-environments, occupying a variety of IFR positions. The descriptive level of the IFR nano-environment in this approach was somewhat coarser (in terms of amino acid residue type and position) than the level used in CSR nano-environment characterization, where physical and chemical descriptors are related to atoms in amino acid main and/or side chains. Nevertheless, the same assumption was applied and tested as in other nano-environment study cases, revealing a very similar positive output, giving us a more detailed knowledge on how enzymes fine-tune their specificity toward different target substrates/inhibitors based on nano-environment changes resulting from the complex interplay of forces generated by all surrounding and constitutive amino acid residues.

12.2.3 Physicochemical and Structural Description of Protein–Protein Interfaces

When considering protein–protein interactions, it is well known that they regulate most biological processes either within or outside cells. Protein–protein interactions are involved in gene expression regulation, metabolic pathways, immunologic response, etc. [40–42]. Proteins communicate with each other through a portion of their surfaces, being able to specifically recognize their partners even in a crowded environment within cells. In fact, macromolecules may interact with different partners by different binding modes, using for each occasion a different portion of their surface.

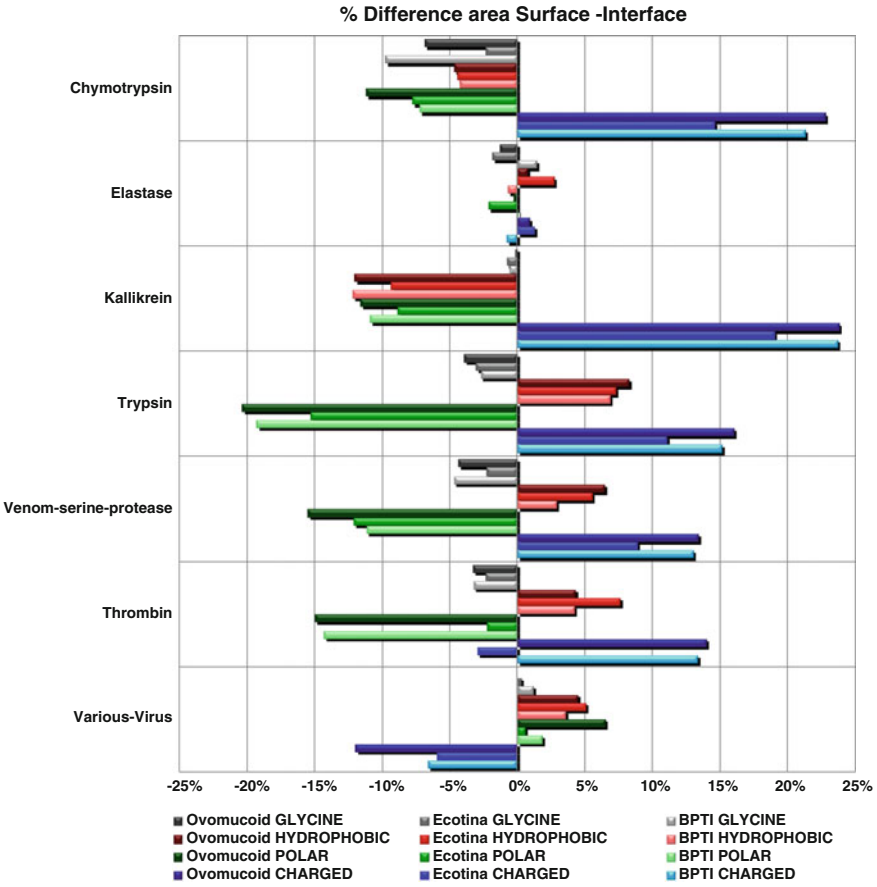

Fig. 12.3 Percentage difference in area occupied at: Surface-Interface. The nano-environment of serine protease interfaces seen through its amino acid composition: This figure presents the difference in occupancy percentage of total enzyme free surface and the IFR area for all 70 serine proteases bound to the inhibitor ecotine, BPTI, and ovomucoid third domain. The enzymes were classified into the following subfamilies: Chymotrypsin (4), Elastase (5), Kallikrein (4), Trypsin (9), Venom (2), Thrombin (5), and Miscellaneous-virus (2) {the number in parentheses representing the number of observed structures}. Average values of percent occupancy are presented for multimember subfamilies. *Bars on the right side* of the graph indicate that the residues are more frequently found at the surface than on the interface. *Bars on the left side* of the graph indicate that particular residue class is more frequently found at the interface than at the surface

In order to gain insight into the atomic details of the interactions between proteins, the knowledge of their three-dimensional structures is critical [43]. When enough structural information is gathered, complex biological processes may be understood in more detail, in particular, because the organism complexity is higher than the sum of the intricacies found in each individual component. The harmonious behavior of the many components inside cells accounts for its homeostasis. Each component of

such intricately coupled system could, in fact, be essential for a particular step of a given regulatory process, involving two protein partners acting in a cyclic way and resulting in coherent feedback [44]. Related to that, it is recognized that many health disorders are the result of protein–protein miscommunication at some level [45].

It is essential to be familiar with the fact that any study attempting to deal with protein–protein interactions will have to face the present lack of sufficient volume of curated data necessary for a consistent statistical analysis. However, this problem could be compensated today by modeling protein structures and their complexes. As stated in [43], it is unlikely to find a soluble protein that either lacks structural information available in public databases or that cannot be modeled by standard homology modeling techniques, such as Modeller [46], or threading algorithms, such as iTasser [47]. This statement can be confirmed by the number of new folding patterns in the Protein Data Bank [23]. At present, the last unique fold deposited in the PDB dates back to 2008.

Also, when it comes to protein–protein interactions, only about 15 % of the known protein structure complexes are so-called hetero-complexes (i.e., complexes composed by nonidentical proteins). This is due to the difficulty in obtaining the crystal state of hetero-complexes, especially in the case of transient complexes with low affinity.

This scenario has stimulated a continuous demand for computational structural biologists to develop tools which help increase the understanding of protein–protein associations by combining structural information on just a single protein with data coming from molecular biology and biophysical techniques, which usually have a lower resolution. Due to the great importance that functional protein networks represent to organisms homeostasis, the computational approaches to model those networks, predict protein interactions, and consequently, rationally design new drugs and agrochemicals represent a constantly increasing stimulus for the scientific community. Our objective, when using protein structure information and knowledge about their interfaces, is that we might be able to avoid a non-desired protein interaction to take place (eliminating side effects for drugs both in areas of human health and plant–pathogen interactions) [48–53]. For this, the understanding of the physicochemical and structural basis of protein–protein interfaces is mandatory. Also, the understanding of the basis of macromolecular recognition at the atomic level may be used to guide docking and molecular dynamics experiments, and also to assist in experimental design for site-directed mutagenesis to change specific area and volume constraints. On top of all this, it is very important to try to fully understand the driving force for protein–protein binding and in particular, which are its principal components.

The ability to predict whether two proteins would interact and the location for their interfaces is an open research topic. The international competition named *Critical Assessment of Predicted Interaction* [54] evaluates different methods for such a task. In the CAPRI, the monomeric structures of each protein–protein complex subunit are given and the multimeric structure is experimentally known but not released. The prediction is evaluated by counting correct interface contacts.

Many methods attempted to predict correctly interface-forming residues. Using the same test set composed of known protein structures (both isolated and in complex), Zhou and Qin [55] compared recently six methods accessible through their respective web servers: ProMate [56], PPI-Pred [57], PINUP [58], SPPIDER [59], cons-PPISP [60], and Meta-PPISP [61]. Each of these methods for predicting interfaces is using some structural and physicochemical properties of the interacting proteins, but only to a limited extent (among them: hydrophobicity, electrostatic potential, surface shape, solvent accessibility, hydrogen bonds established across the interacting proteins and space clashes).

All mentioned methods make use of the so-called *sequence conservation* attribute. Our work focused on designing an algorithm for classifying amino acid residues belonging to protein interfaces (separating them from those that do not), entirely excluding attributes that are not measured directly from the protein structure, such as sequence conservation.

To assess the potential of simple linear methods for prediction of interface-forming residues using physiochemical attributes only, a plot with the average values (divided by their respective standard deviations) for properties of interface and free surface amino acid residues was generated, based on a non-redundant dataset DS30 (see details in "Procedures" section). As shown in Fig. 12.4, a large number of parameters were analyzed with respect to their intrinsic capacity of differentiating those two residue ensembles, for all amino acid types. All the descriptors having their values away from zero are marked as most promising attributes for prediction purposes. Regarding the electrostatic potential, except for the EP@surf, the other three EP flavors are clearly capable of indicating which a. a. belong to the interface ensemble and which belong to the free protein surface.

Next, all descriptors may be linearly combined to develop an approach for predicting interface residues using linear discriminant analysis (LDA). The LDA uses the average and standard deviation values retrieved from a training dataset, for each attribute, for both interface and free surface residues. In the development of the STING-LDA predictor, the DS30 dataset was submitted to tenfold cross validation in order to check for possible training bias and the final predictor was built using the entire DS30. Any amino acid residue is then classified into interface or free surface ensemble following the maximum likelihood equations:

$$f_{\mathrm{IFR}} = \frac{1}{(2\pi)^{N/2}\,|\boldsymbol{\Sigma}_{\mathrm{IFR}}|^{1/2}} \exp\left[-\frac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{IFR}})'\,\boldsymbol{\Sigma}_{\mathrm{IFR}}^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{IFR}})\right]$$

$$f_{\mathrm{FSR}} = \frac{1}{(2\pi)^{N/2}\,|\boldsymbol{\Sigma}_{\mathrm{FSR}}|^{1/2}} \exp\left[-\frac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{FSR}})'\,\boldsymbol{\Sigma}_{\mathrm{FSR}}^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{FSR}})\right]$$

where IFR stands for interface-forming residues, FSR for free surface residue, x is the attributes vector for the amino acid residue being predicted, μ_{IFR} and μ_{FSR} are the vectors of attribute averages for each ensemble, and Σ_{IFR} and Σ_{FSR} are the vectors for attribute variances.

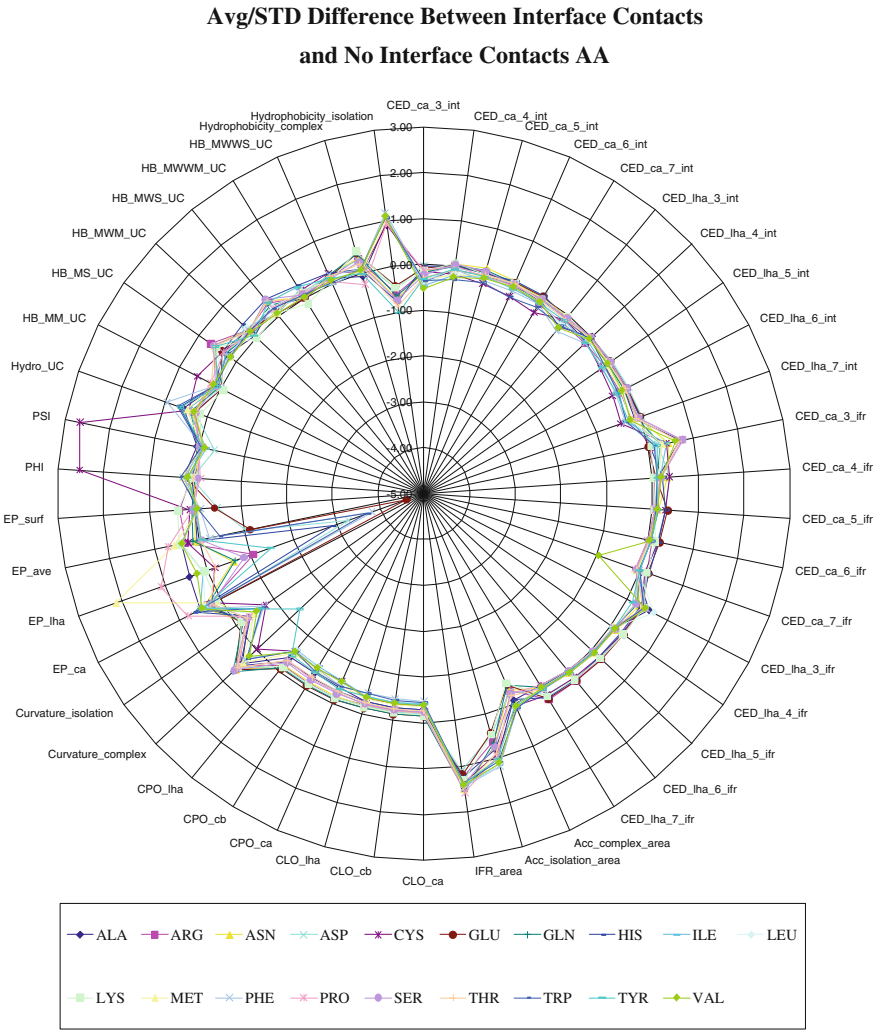

Fig. 12.4 Avg/STD difference between IFR and FSR. Radial plot for 46 different protein structure and physicochemical properties, presented for two ensembles—IFR and FSR and for all 20 amino acids. The values plotted are the attribute averages divided by their corresponding standard deviations, and were extracted from the BlueStar STING associated to the DS30 non-redundant dataset of protein–protein complexes. The values far from zero reveal high prediction power of the respective attribute. Full description of all attributes (and acronyms) can be found at: http://www.cbi.cnptia.embrapa.br/SMS/STINGm/help/MegaHelp_JPD.html

The STING-LDA gives the probability for individual amino acid residues to be located on the interface of protein–protein complexes. STING-LDA is currently implemented into the Java Protein Dossier (^{J}PD) module of BlueStar STING. The STING-LDA results on known protein–protein complexes show that high values of

the classification threshold (above or equal to 80 %) will return just a fraction of the true interface, but with high precision or reliability. In turn, when the classification threshold is reduced to a smaller value (under 40 or 30 %), the coverage of the interface predicted is higher, but with more uncertainty. It is up to the user's requirements that this classification threshold should be decided.

When comparing STING-LDA with the other methods, two outstanding points need to be emphasized: (A) all other methods use sequence conservation attributes while STING_LDA does not, guaranteeing that our method would still function for orphan structures, where other methods would fail; and (B) the performance of STING-LDA is higher than most other methods with the exception of Meta-PPISP and, for some classification thresholds, PINUP. The comparison was carried out following Zhou and Qin [55] work, where the precision is used to rank methods according to specific sensitivity (coverage) values. As one may clearly observe on Fig. 12.5, adding the WNA attributes to classifier increases performance of the predictor. However, adding conservation attribute to WNA attributes does not increase performance, indicating that certain plateau was reached. This means that all the

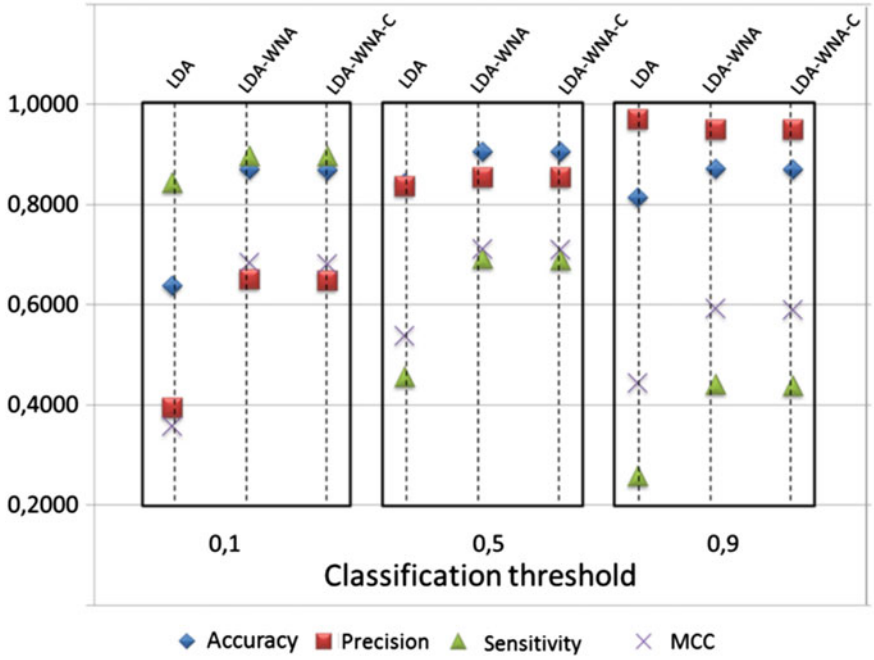

Fig. 12.5 IFR prediction performance dependence on cutoff values for the LDA classifier with conservation attributes and comparison with Sting-LDA-WNA. Classification with a cutoff of 0.5, the precision rate is always above 85 %, reaching more than 95 % with a cutoff of 0.9. The MCC rate is higher for a 0.5 cutoff; nevertheless, using a cutoff of 0.5 results in a similar MCC. When comparing the performance of the Sting-LDA_WNA with the Sting_WNA_Conservation classifiers, no difference is noted for the three selected cutoff values

necessary information for distinguishing IFRs from FSRs is present in the original descriptor set, (retrieved directly from a protein structure) if a sufficiently extensive list is used. The electrostatic potential properties (in one of the four available flavors) figured among the top 5 ranking attributes used by the STING-LDA_WNA for predicting protein interfaces. However, more appropriate insight into the real EP rank will be obtained only after understanding the main components and the principal driving forces that guide protein–protein binding, something we will discuss in the next section.

12.3 Hydrophobicity as the Major Driving Force of Protein–Protein Interactions and EP as a Crucial Complementary Alternative

To date, a quantitative assessment of the relevance of the hydrophobic effect as a determinant of protein–protein interactions remains an unmet goal. Quantifying it exactly, and then qualitatively analyzing possible exceptions, was never fully described in the literature in spite of the existence of a high volume of papers dealing with this issue.

Starting from the premise that the hydrophobic effect has a significant influence in almost all protein–protein associations [62–68], we decided to design a new approach that would define how to effectively measure the hydrophobicity of interfaces, and that would be capable of assessing precisely how important and wide spread is such contribution within the assembly of complexes in the known protein structure universe. To achieve this, we have defined a specific parameter associated/related to hydrophobicity: the surface hydrophobicity index (SHI). The principle considered here was that, if the hydrophobic effect is a driving force for protein oligomerization, the interface area should be slightly more hydrophobic than the remaining surface (here also referred to as free surface). This larger local hydrophobicity at the interface might be measured by a specific, well-described, and intuitive descriptor/parameter/index. Thus, the SHI of a given chain in isolation, which is to say the SHI of a protein chain not assembled into the complex with any other protein chain, is a value that considers hydrophobicity at the interface area plus the one at the remaining free surface area and it should be higher (more hydrophobic) than the SHI of a given chain in complex. In the latter case, we have a measure of hydrophobicity for only the free surface area for this particular protein chain(s) as the interface is not any more accessible to the solvent. Counting how many complexes obey such behavior in the datasets described in the "Procedure" section, one may have a very good idea of how often proteins use the surface hydrophobicity as a major driving force in order to create complex assemblies with other protein molecules. In other words, it is possible to precisely assess how important and wide spread the hydrophobic contribution is for the assembling of protein complexes in the known protein structure universe.

The three most cited hydropathy scales were used to construct three different SHI flavors: Kyte–Doolittle [69], Eisenberg [70], and Engelman [71]. All three SHI (hydropathy scales) flavors for all four data sets (DS100, DS95, DS50, and DS30) have shown a very similar behavior (albeit, not identical, as in fact was expected) regarding oligomerization and other derived indicators.

In this work, the ΔSHI (used interchangeably with: dSHI) was introduced and defined as the difference between SHI calculated for a selected chain, separated from any other one ($SHI_{isolation}$) and SHI calculated for the same chain but now assembled in a complex, as described in a corresponding PDB entry ($SHI_{complex}$). A positive Δ SHI indicates that the interface area is more hydrophobic than the remaining protein surface area.

Strong positive correlation was found to exist among the ΔSHI value and the ratio between corresponding interface size and the total surface area size (both for single chain proteins and complete protein/oligomeric complexes). This implies that as the size of the interface grows, so it does the area of hydrophobic residues that compose the selected interface, which, in turn, becomes buried during complex formation.

Slightly more than 91 % of all studied interfaces obey the rule: $\Delta SHI > 0$, and for interfaces of the most frequent size (>3,000 $Å^2$) in the DS95 set, this percentage rises to more than 98 %. Cases which do not obey the $\Delta SHI > 0$ rule were found to belong to three major classes: a) proteins having significantly smaller than the average interface sizes, b) membrane proteins, and c) some large oligomers from virus capsids. More importantly, a total of 99.9 % of the complexes where core residues are found to be part of the interface (85 % of the DS95 complexes), obey the $dSHI > 0$ or $dSHIp > 0$ or $dSHIcore > 0$, indicating clearly the high degree of occurrence of cases where hydrophobic effect is a major driving force in protein complex formation. The dSHIp corresponds to the SHI value calculated for protein conglomeration considered completely (as in capsids) and dSHIcore corresponds to the SHI value where the interface is identified with the region where amino acids have completely lost access to the solvent. In Table 12.2, we depicted how dSHI is behaving for those chains that have core residues and for those that do not.

In this part of our work, we describe how frequently proteins use the hydrophobic effect, assumed to be a major driving force that provides the energy necessary for establishing the protein complexes, and we also show how this influence varies with the size of the interface area. The intertwining of those two factors is also de-convoluted so that one could understand the influence of changing the profile of constituent amino acids in the function of the interface geometry and its chemical characteristics, (a typical example to illustrate such interdependency would be absence or presence of interface core). The density of internal and interchain contacts was also studied, yielding results that indicate a higher density of internal contacts among amino acids occupying the interface area when compared to the free surface area. The internal contact density profiles for small and large interfaces also offers a plausible explanation for compensative energy sources used instead of hydrophobic effect for protein complex formation in the case of proteins with much smaller than average interface sizes (where in fact the largest occurrence of deviation from $dSHI > 0$ rule was observed).

Table 12.2 The dSHI behavior for chains with and without core interface residues; all three hydrophobic scales are shown

	Kyte–Doolittle		Eisenberg		Engelman	
	Number	%	Number	%	Number	%
Chains without CORE						
dSHI > 0	684	65.52	732	70.11	713	68.30
dSHI = 0	26	2.49	16	1.53	46	4.41
dSHI < 0	334	31.99	296	28.35	285	27.30
Chains with CORE						
dSHI > 0	5521	93.78	5550	94.28	5557	94.39
dSHI = 0	29	0.49	20	0.34	55	0.93
dSHI < 0	337	5.72	317	5.38	275	4.67

From trends observed in Fig. 12.6, it is clear that the protein–protein interactions for most of the cases where the interface areas are close to its average value (or above it), predominantly use the hydrophobic force for binding. The average interface area is approximately 2,100 $Å^2$ (but the standard deviation is rather large). However, proteins that form smaller interfaces (below the value of an average interface area size), such as the case of serine proteases bound to their respective inhibitors, would have to employ alternative energy sources in order to compensate the deviation from the dSHI > 0 rule, most frequently finding it in electrostatic interactions. This point was confirmed by the presence of a higher density of charge–charge interactions and also of hydrogen bonding at those particular interface areas.

In the session dedicated to enzyme specificity, we outlined that the serine proteases, for example, have rather small interface areas (around 600$Å^2$) and at the same time a large portion of their IFRs belong to the ensemble of polar residues, indicating that the electrostatic potential and interactions generated from it could provide the missing energy source for stabilizing serine protease complexes.

12.4 Protein Folding and Elements of Secondary Structure

To understand the relationship between the amino acid sequence of a protein on one side and protein structure and function on the other, we proposed an in depth analysis of the nano-environment where the protein secondary structure elements (α-helix, β-sheet and turns) are inserted. The event that motivated such approach was the previous identification of the existence of certain "signals," i.e., a variation in the values of physical–chemical descriptors observed in the three-dimensional space where the secondary structure is inserted. Understanding how the elements

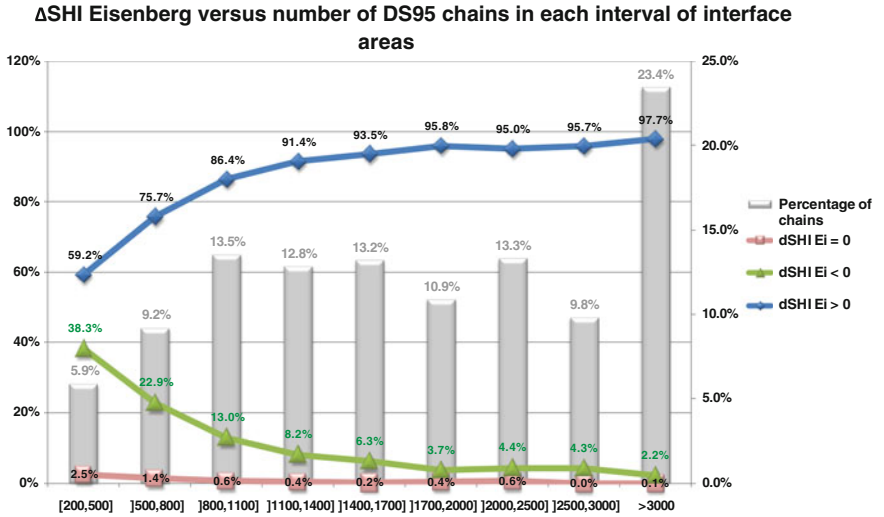

Fig. 12.6 ΔSHI Eisenberg versus number of DS95 chains in each interval of interface areas. Relationship between fraction of chains that obey ΔSHI > 0 rule (in % of D95 dataset) and size of their respective interface areas. The Eisenberg hydropathy scale was used to generate data (the Kyte–Doolittle and Engelman hydropathy scales were also employed yielding a data distribution following closely the ones presented in Fig. 12.5). Vertical bars show the percentage of DS95 chains having the size of the interface area in the designated range (*x-axis* shows the range for the interface area size in $Å^2$). On the *left* side, the y-axis shows the percent of protein chains in DS95 having ΔSHI > 0 (relevant for *blue*, *green*, and *pink* color curves) while on the *right* side of this plot, the *y-axis* indicates the percent of protein chains in DS95 having the interface size within the indicated span (relevant for the *gray* bars)

of secondary structure are formed paves the way to understanding how proteins assume their final structure and, hence, how they perform their function. In this work, we again used descriptors from the STING_RDB, a database unique in the world because it brings together in one place more than 1,300 descriptors (physicochemical and structural) of all amino acid residues, for each chain, of all structures deposited in the PDB (Protein Data Bank). The non-redundant structures from PDB, having corresponding structure/function descriptors stored in STING_RDB, were separated in different datamarts obeying strict selection rules as described above in "procedures." The structures contained in such datamarts had their secondary structure elements (of equal size) structurally aligned and then the physicochemical and structural attributes, describing the nano-environment where an element of secondary structure was located, were extracted and their averages calculated. This process was used to search for "signals" and was applied in order to the enhance signal to noise ratio (medium to high level noise is normally present in all biological measurements). We were able to identify a series of "signals" encountered in protein structural space and attributed to specific SSE types, but here we only present (Fig. 12.7) the EP signal for alpha helical and for beta pleated sheets. These signals

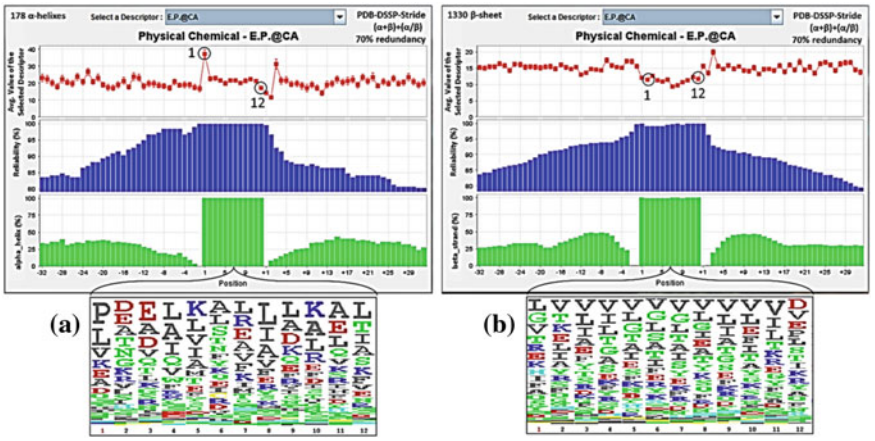

Fig. 12.7 Electrostatic potential calculated at alpha carbon (CA) atom of amino acid residues before, during and after the 12 amino acid residues long alpha helical structure (placed in the middle of the plot). The observed "signal," visualized here as a variation in average value of the EP@Ca, was obtained from 178 (**a**) structurally aligned alpha helices and 1,330 beta strands (**b**), encountered in proteins of ($\alpha + \beta$) and (α/β) type. As depicted on the top inset (*red line*), one can observe in (*A*) two peaks in EP@Ca average values: the first one occurring at the first amino acid residue of the SSE analyzed and a second one which occurs 3 residues after the C-terminal of the SSE studied and in (*B*) that the EP@Ca average value is clearly lower starting at the first amino acid residue before the N-terminal of the SSE analyzed and ending at the C-terminal of the SSE studied. The *blue* bar graph is showing the reliability of data in terms of how close is the number of structurally aligned structures at any position of the alignment to the optimal (maximum) value, which in the (*A*) case is equal to 178 and in the (*B*) case is equal to 1,330. The third graph, depicted in *green*, represents bars which indicate percentage of alpha helical/beta sheet (**a** and **b**, respectively) present at any point of the positional alignment (clearly, having a maximum value at the extension of SSE studied). At the lower part of this figure, one can observe the consensus sequence of the SSE structurally aligned. Comparing this consensus sequence to the Chou and Fasman propensity tables for alpha helices (in **a**) and beta strands (in **b**), one can see very high coincidence of amino acid types and ranking

prove correct the hypothesis that motivated this work and also show the importance of the EP parameter in constructing the appropriate nano-environment for each type of SSE.

The nano-environment of the SSE has shown that a composite "signal" is identifiable, containing a variation in average property values for accessibility, cross link order, cross presence order (the latter two properties related to packing and described in details in BlueStar STING manual), rotamer type and electrostatic potential calculated at the alpha carbon atom. Once again, the electrostatic potential is present as a major contributor to composing the appropriate nano-environment.

12.5 Case Study: Electrostatic Potential as a Possible Missing Clue in Considering Causes for Onset of Amyotrophic Lateral Sclerosis Disease in Patients with Mutated Superoxide Dismutase Enzyme

The amyotrophic lateral sclerosis disease belongs to a group of disorders known as *motor neuron diseases,* which are characterized by the gradual degeneration and death of motor neurones [72–75]. Approximately 10 % of the cases are genetically related and are inherited in an autosomal recessive manner, in which case the disease is named familial ALS or FALS. Only 20 % of FALS are directly linked to mutations found in superoxide dismutase (SOD1). To date, around 100 different mutations have been cataloged and structures reported. In the PDB (November 2013), there are 109 SOD1 structures from *homo sapiens*, 42 of them showing SOD1 with mutated residues. The SOD is a dimeric structure and its optimal functioning depends ultimately on how well two monomers are bound. Molecular dynamics studies have shown that the SOD1 mutants where the alanine (at position) 4 was substituted by valine (the most frequently found mutation in an aggressive form of FALS), is less stable in terms of maintaining its dimeric form and is destabilizing the metal-binding site [76], eventually leading to a misfolded enzyme state. Since it was already known that the SOD1 uses electrostatic attraction to achieve faster than diffusion limited substrate approach and recognition, exploring even further the electrostatic component for both stability- and substrate-related issues was somehow obvious and needed. Having precalculated the electrostatic potential values at crucial points of protein and/or mutant sites (amino acid residue atoms and surfaces), we used the BlueStar STING and its module MSSP (displaying aligned multiple structures single parameter) to compare wild-type and mutated structures. Our objective was to obtain more details on how a minor change such as a mutation of alanine in valine (close to the N-terminal of the SOD1) could cause onset of such a devastating disease and what is the role of electrostatic forces in this complex event.

The MSSP module displays the structurally aligned wild-type and mutant structures in the STING's structure window as well as the corresponding sequences (aligned following the structural alignment of the two chains) in the sequence window. In addition, the MSSP displays in a Cartesian plot the values of selected attributes aligned so as that the sequence of points corresponds to the alignment of the two structures. Any departure of values of selected properties of the two structures could be easily spotted and then analyzed. To get a comparison of the wild-type and mutated structures, we used first 1HL5 and its chain A (wild-type SOD1), and 1UXM, chain A (the SOD1 mutated structure at position: A4V), as shown in the upper panel of Fig. 12.8, in red and blue, respectively. Only the EP@surf values did present certain discrepancies among the two aligned structures, and in several regions, however, they were not very significant.

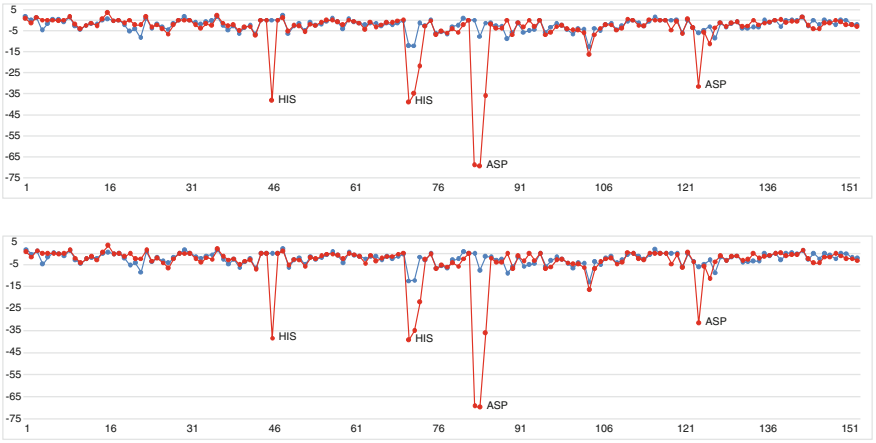

Fig. 12.8 The BlueStar Sting MSSP module output. Electrostatic potential calculated at the surface (EP@surf) of the nearest amino acid residue for 1HL5.pdb, *upper* panel (wild type, in *red*) and 1UXM.pdb (mutant, in *blue*) and 1SPD.pdb, *lower* panel (wild type, in *red*) and 1N19.pdb (mutated structure, in *blue*)

The second attempt, shown in the lower panel of Fig. 12.8, yielded more peculiar results; namely, we used the pdb structure 1SPD (shown in red) and its chain A (wild-type SOD1,) and 1N19 (shown in blue), chain A (the SOD mutated structure: A4V but also containing substitutions of its two free cysteine residues: C6A and C111S). The two cysteines were modified to avoid auto oxidation of their sulfur atoms. As one can easily observe, the mutated structure has a dramatic decrease of EP@surf at a number of positions, nevertheless, remote to the site of the mutated alanine (position number 4). A more thorough inspection of amino acid residues which suffered a great modification in value for their respective EEP@surf reveals that they are involved and/or very close to the metal-binding atoms (shown at Fig. 12.9).

One could clearly observe that the mutated structure is describable by the loss of a good portion of the electrostatic potential value at some residues located close to the metal ions. Whether or not this feature is related to the fact that the mutant structure has been reported to bind only 30% of metal ions as compared to the wild type, and also why we did not observe such behavior in other pairs of wild-type/mutant alignments, remains to be clarified. There are, however, a number of possible factors to be additionally considered in the analysis of this result such as the space group of the compared structures, monomer interfaces, contacts established among monomers, etc. In any case, the value of having the EP strength calculated and compared at specific sites in protein structures is undoubtedly high when considering structure/function relationship.

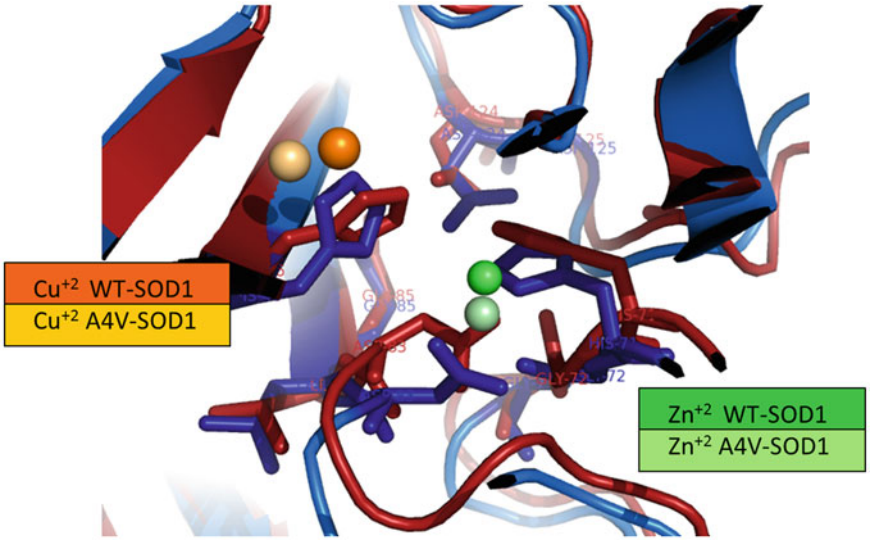

Fig. 12.9 Structural alignment of 1SPD_A (wild-type SOD1) and 1N19_A (A4V SOD1mutant) with emphasis on the Cu (*upper left*) and Zn (*lower right*) positions. Both the Cu and Zn atoms were displaced in the mutated structure, which could be a consequence of displacement of histidine residues at position 46 and 71 as well as aspartic acid at positions 83 and 124, exactly the ones which lost a good part of their electrostatic potential strength in the mutant

12.6 Conclusions

In this work, we purported the idea that biomolecules, and especially proteins, are especially engineered to realize a nano-environment suitable to their structural and functional properties. For instance, the specificity of enzymes is related to the composition and characteristics of substrate-binding residues. Such nano-environment allows very different substrates to bind and then be processed by the same set of CSR in different enzymes (belonging to the same family), undergoing exactly the same chemical transformation (normally described using the enzyme's EC nomenclature). Likewise, the building blocks of ordered protein structures—the secondary structure elements (SSE)—such as helical constructions and beta pleated sheets, are also inserted into very specific nano-environments which are defined both by the surrounding amino acid residues as well as by those of the SSE itself. For each SSE, there is an appropriate nano-environment which in turn would not be suitable for any other SSE type.

In this context, the electrostatic potential has proven to be a valuable asset for establishing the relationship between protein structure and function. This physicochemical property has been used for the past four decades as the single most important factor, especially when charged interactions were considered in the nano-universe of biological macromolecules. More recently, the EP has gained adequate space also in

comparative studies, which aim not only at describing biological events qualitatively but also at estimating them quantitatively.

Our studies were centered on the role of the EP in determining the function of protein districts, and on the relationship of structural properties (which includes EP) to the functional behavior of enzymes and proteins in general. Furthermore, we have established a road map for the analysis of the constitutional participation of different structural, physical, and chemical properties in composing complex "signals" which we described here as a perturbation in average values of composite attributes characterizing the vicinity of functional protein districts. As it was shown, all districts considered here (protein interfaces, catalytic sites, and secondary structure elements, as well as their slight variations), do include as a major constitutive component the electrostatic potential built by all participating and surrounding residues. To a different extent, EP was shown as a crucial element for protein specificity and interfacing and in the case of nano-environment characterization for CSR.

The single case study we present here opens a path for similar applications: we wanted to understand the intrinsic mechanistic and dynamical details crucial for explaining the onset of a particular disease, FALS.

Our future research perspectives revolve around the identification of the characteristics of the nano-environments specific for the protein–DNA and protein–drug interfaces with a wide spectrum of applications.

References

1. Neshich G, Mancini A, Yamagishi M, Kuser P, Fileto R, Baudet C, Pinto I, Montagner A, Palandrani J, Krauchenco J, Torres R, Souza S, Togawa R, Higa RH (2004) Java protein dossier: a novel web-based data visualization tool for comprehensive analysis of protein structure, Nucl Acids Res 32(Web Server issue):W595–W601
2. Neshich G, Togawa R, Mancini AL, Kuser PR, Yamagishi MEB, Pappas G Jr, Torres WV, Campos TF, Ferreira LL, Luna FM, Oliveira AG, Miura RT, Inoue MK, Horita LG, de Souza DF, Dominiquini F, Álvaro A (2003) STING millennium: a web based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. Nucl Acids Res 31(13):3386–3392
3. Neshich G, Borro LC, Higa R, Kuser P, Yamagishi M, Franco EH, Krauchenco J, Fileto R, Ribeiro A, Bezerra G, Velludo T, Jimenez T, Furukawa N, Teshima H, Kitajima K, Bava A (2005) Diamond STING server. Nucl Acids Res 33(Web Server issue):W29–35
4. Neshich G, Mancini AL, Yamagishi MEB, Kuser PR, Fileto R, Pinto IP, Palandrani JF, Krauchenco JN, Baudet C, Montagner AJ, Higa RH (2005) STING report: convenient web-based application for graphic and tabular presentations of protein sequence, structure and function descriptors from the STING database. Nucl Acids Res 33(Database Issue):D269–D274
5. Oliveira SRM, Almeida GV, Souza' KRR, Rodrigues DN, Kuser-Falcão PR, Yamagishi MEB, Santos EH, Vieira FD, Jardine JG, Neshich G (2007) STING_RDB: a relational database of structural parameters for protein analysis with support for data warehousing and data. Mining Genet Mol Res 6(4):911–922
6. Radic Z, Kirchhoff P, Quinn D, McCammon J et al (1997) Electrostatic influence on the kinetics of ligand. J Biol Chem 272

7. Sheinerman F, Norel R, Honig B (2000) Electrostatic aspects of protein-protein interactions. Curr Opin 10:153–159
8. Honig B, Nicholls A (1995) Classical electrostatics in biology and chemistry. Science 268:1144–1149
9. Rocchia W, Neshich G (2007) Electrostatic potential calculation for biomolecules— creating a database of pre-calculated values reported on a per residue basis for all PDB protein structures. Genet Mol Res 6(4):923–936
10. Togawa RC, Kuser PR, Higa RH, Yamagishi MEB, Mancini AL, Neshich G (2004) STING Millennium Suite: integrated software for extensive analyses of 3d structures of proteins and their complexes. BMC Bioinformatics 5(1):107
11. Mancini A, Higa R, Oliveira A et al (2004) STING contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. Bioinformatics 20(13):2145–2147
12. Neshich G, Mazoni I, Oliveira S, Yamagishi M, Kuser-Falcao P, Borro L, Morita D, Souza K, Almeida G, Rodrigues D et al (2006) The star STING server: a multiplatform environment for protein structure analysis. Genet Mol Res 5:717–722
13. Borro L et al (2006) Predicting enzyme class from protein structure using Bayesian classification. Genet Mol Res 5:193–202
14. Fernandez J, Hayashi M, Camargo A et al (2003) Structural basis of the lisinopril-binding specificity in N- and C-domains of human somatic ACE. Biochem Biophys Res Comm 308(2):219–226
15. de Freitas S, de Mello L, da Silva M et al (1997) Analysis of the black-eyed pea trypsin and chymotrypsin inhibitor alpha-chymotrypsin complex. FEBS Lett 409(2):121–127
16. Marcellino L, Neshich G, de Sa MG et al (1996) Modified 2S albumins with improved tryptophan content are correctly expressed in transgenic tobacco plants. FEBS Lett 385(3):154–158
17. Simoes M, Bahia D, Zerlotini A et al (2007) Single nucleotide polymorphisms identification in expressed genes of Schistosoma mansoni. Mol Biochem Parasitol 154(2):134–140
18. Melo R, Ribeiro C, Murray C et al (2007) Finding protein-protein interaction patterns by contact map matching. Genet Mol Res 6(4):946–963
19. Braghini C, Neshich I, Neshich G et al (2013) New mutation in the myocilin gene segregates with juvenile-onset open-angle glaucoma in a Brazilian family. Gene 523:50–57
20. Dias-Lopes C, Neshich I, Neshich G et al (2013) Identification of new sphingomyelinases D in pathogenic fungi and other pathogenic organisms. PLoS ONE 8(11)
21. Nadzirin N, Firdaus-Raih M (2012) Proteins of unknown function in the protein data bank (PDB): an inventory of true uncharacterized proteins and computational tools for their analysis. Int J Mol Sci 13(10):12761–12772
22. FTP site for PDB/RCSB [Online]. Available: ftp://ftp.wwpdb.org
23. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P (2000) The protein data bank. Nucl Acids Res 28:235–242
24. Henrick K, Krissinel E (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372:774–797
25. Sridharan S, Nicholls A, Honig B (1992) A new vertex algorithm to calculate solvent accessible surface areas. Biophys J 61:A174
26. UniProt Consortium (2009) The universal protein resource (UniProt) 2009. Nucleic Acid Res 37(Database issue):D169–D174
27. Martin AC (2005) Mapping PDB chains to UniProtKB entries. Bioinformatics 1;21 (23):4297–4301
28. Bourne P, Addess K, Bluhm W, Chen L, Deshpande N, Feng Z, Fleri W, Green R, Merino-Ott J, Townsend-Merino W, Weissig H, Westbrook J, Berman H (2004) The distribution and query systems of the RCSB protein data bank. Nucl Acids Res 1;32(Database issue):D223–D225
29. PDB, RCSB - PDB [Online]. Available: ftp://resources.rcsb.org/sequence/clusters/clusters95.txt
30. Tusnády G, Dosztányi Z, Simon I (2004) Transmembrane proteins in the protein data bank: identification and classification. Bioinformatics 20(17):2964–2972

31. Jayasinghe S, Hristova K, White SH (2001) A database of membrane protein topology. Protein Sci 10:455–458
32. Porter CT, Bartlett1 GJ, Thornton JM (2004) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucl Acids Res 32:D129–D133
33. Yang J, Tiyyagura A, Chen F, Honaver V (1999) Academia.edu, 1999. [Online]. Available:http://www.academia.edu/2791981/Feature_subset_selection_for_rule_induction_using_RIPPER
34. Cohen W (1995) Fast effective rule induction. Morgan, San Francisco
35. Higa R, Togawa R, Neshich G (2004) ConSSeq: a web-based application for analysis of amino acid conservation based on HSSP database and within context of structure. Bioinformatics 20(12):1983–1985
36. Higa R, Neshich G (2006) Building multiple sequence alignments with a flavor of HSSP alignments. Genet Mol Res 3(1):127–137
37. Porollo A, Meller J (2007) Prediction-based fingerprints of protein-protein interactions. Proteins 66(3):630–645
38. Justel A, Peña D, Zamar R (1997) A multivariate Kolmogorov-Smirnov test of goodness of fit. Stat Prob Lett 35(3):251–259
39. Ribeiro C, Togawa RC, Neshich IA, Mazoni I, Mancini AL, Minardi RCdM, Silveira CHd, Jardine JG, Santoro MM, Neshich G (2010) Analysis of binding properties and specificity through identification of the interface forming residues (IFR) for serine proteases in silico docked to different inhibitors. BMC Struct Biol 10:36
40. Xenarios I, Eisenberg D (2001) Protein interaction databases. Curr Opin Biotech 12:334–339
41. Ponsting l, Kabir T, Gorse D, Thornton J (2005) Morphological aspects of oligomeric protein structures. Prog Biophys Mol Biol 89:9–35
42. Reichmann D, Rahat O, Cohen M, Neuvirth H, Schreiber G (2007) The molecular architecture of protein-protein binding sites. Curr Opin Struct Biol 17:67–76
43. Alloy P, Russell R (2006) Structural systems biology: modelling protein interactions. Nature Rev Mol Cell Biol 7:188–197
44. Kitano H (2002) Computational systems biology. Nature 420(6912):206–210
45. Kastritis P, Bonvin A (2013) Molecular origins of binding affinity: seeking the Archimedean point. Curr Opin Struct Biol, pii: S0959-440X(13)00121-8. 19 July 2013, doi:10.1016/j.sbi.2013.07.001
46. Sali A, Blundell T (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815
47. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725–738
48. Lybrand T (1995) Ligand-protein docking and rational drug design. Curr Opin Struct Biol 5(2):224–228
49. Beeley L, Duckworthy D (1996) The impact of genomics on drug design. Drug Discov Today 7:474–480
50. Parrill A (1996) Evolutionary and genetic methods in drug design. Drug Discov Today 1(8):514–521
51. Wade R (1997) 'Flu' and structure-based drug design. Structure 5(9):1139–1144
52. Zsoldosa Z, Szaboa I, Szaboa Z, Johnson A (2003) Software tools for structure based rational drug design. J Mol Struct: Theochem 659–665, 666–667
53. Acharya C, Coop A, Polli J, MacKerell A Jr (2011) Recent advances in Ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. Curr Comput Aided Drug Des 7(1):10–22
54. Janin J, Wodak S (2007) The third CAPRI assessment meeting. Structure.15:755–759
55. Zhou H, Quin S (2007) Interaction-site prediction for protein complexes: a critical assessment. Bioinformatics 23(17):2203–2209
56. Neuvirth H, Raz R, Schreiber G (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol 338:181–199

57. Bradford J, Westhead D (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics 21:1487–1494
58. Liang S, Zhang C, Liu S, Zhou Y (2006) Protein binding site prediction using an empirical scoring function. Nucl Acids Res 34(13):3698–3707
59. Porollo A, Meller J (2007) Prediction-based fingerprints of protein-protein interactions. Proteins 66:630–645
60. Chen H, Zhou H-X (2005) Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. Proteins 61:21–35
61. Qin S, Zhou H-X (2007) Meta-PPISP: a meta web server for protein-protein interaction site prediction. Bioinformatics 23(24):3386–3387
62. Young L, Jernigan R, Covell D (1994) A role for surface hydrophobicity in protein-protein recognition. Protein Sci 3(5):717–729
63. Chothia C, Janin J (1975) Principles of protein-protein recognition. Nature 256:705–708
64. Tsai C, Lin S, Wolfson H, Nussinov R (1997) Studies of protein-protein interfaces; a statistical analysis of the hydrophobic effect. Protein Sci 6:53–64 [PubMed: 9007976]
65. Ben-Naim A (2006) On the driving forces for protein-protein association. J Chem Phys 125:024901–0249010
66. Argos et al (1988) An investigation of domain and subunit interfaces. Protein Eng 2:101–113
67. Hu Z, Ma B, Wolfson J, Nussinov R (2000) Proteins-structure function. Genetics 39:331–342
68. Jones S, Thornton J (1996) Proc Natl Acad Sci USA 93:13–20
69. Kyte J, Doolittle R (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132
70. Eisenberg D (1984) Three-dimensional structure of membrane and surface proteins. Annu Rev Biochem 53:595–623
71. Engelman D, Steitz T, Goldman A (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Biomol Struct 15:321–353
72. Deng H, Hentati A, Tainer J, Iqba lZ, Cayabyab A, Hung W, Getzoff E, Hu P, Herzfeldt B, Roos R et al (1993) Amyotrophic lateral sclerosis and structural defects in Cu, Zn superoxide dismutase. Science 261(5124):1047–1051
73. Yim H-S, Kang J-H, Chock PB, Stadtman ER, Yim MB (1997) A familial amyotrophic lateral sclerosis-associated A4V Cu, Zn-superoxide dismutase mutant has a lower Km for hydrogen peroxide. Correlation between clinical severity and the Km value. J Biol Chem 272(14):8861–8863
74. Cardoso R, Thayer M, DiDonato M, Lo T, Bruns C, Getzoff E, Tainer J (2002) Insights into Lou Gehrig's disease from the structure and instability of the A4V mutant of human Cu, Zn superoxide dismutase. J Mol Biol 324(2):247–256
75. DiDonato M, Craig L, Huff M, Thayer M, Cardoso R, Kassmann C, Lo T, Bruns C, Powers E, Kelly J, Getzoff E, Tainer J (2003) ALS mutants of human superoxide dismutase form fibrous aggregates via framework destabilization. J Mol Biol 332(3):601–615
76. Schmidlin T, Kennedy B, Daggett V (2009) Structural changes to monomeric CuZn superoxide dismutase caused by the familial amyotrophic lateral sclerosis-associated mutation A4V. Biophys J 97(6):1709–1718