ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
"Antonio Ruberti"
CONSIGLIO NAZIONALE DELLE RICERCHE

P. Bertolazzi, C. Guerra, G. Liuzzi

PREDICTING PROTEIN-LIGAND AND PROTEIN-PEPTIDE INTERFACES

R. 2, March 2013

Paola Bertolazzi – Consiglio Nazionale delle Ricerche, Istituto di Analisi dei Sistemi ed Informatica "A. Ruberti", Viale Manzoni 30, 00185 Rome, Italy.
paola.bertolazzi@iasi.cnr.it.

Concettina Guerra – College of Computing, Georgia Institute of Technology, Atlanta, GA, USA. guerra@cc.gatech.edu.

Giampaolo Liuzzi – Consiglio Nazionale delle Ricerche, Istituto di Analisi dei Sistemi ed Informatica "A. Ruberti", Viale Manzoni 30, 00185 Rome, Italy.
giampaolo.liuzzi@iasi.cnr.it.

This work has been partially supported by the FLAGSHIP "InterOmics" project (PB.P05) funded by the Italian MIUR and CNR organizations

ISSN: 1128–3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.cnr.it
URL: http://www.iasi.cnr.it

Abstract

The paper deals with the identification of binding sites and concentrates on interactions involving small interfaces. In particular we focus our attention on two major interface types, namely protein-ligand and protein-peptide interfaces. As concerns protein-ligand binding site prediction, we classify the more interesting methods and approaches into four main categories: (a) shape-based methods, (b) alignment-based methods, (c) graph-theoretic approaches and (d) machine learning methods. Class (a) encompasses those methods which employ, in some way, geometric information about the protein surface. Methods falling into class (b) address the prediction problem as an alignment problem, i.e. finding protein-ligand atom pairs that occupy spatially equivalent positions. Graph theoretic approaches, class (c), are mainly based on the definition of a particular graph, known as the protein contact graph, and then apply some sophisticated methods from graph theory to discover subgraphs or score similarities for uncovering functional sites. The last class (d) contains those methods that are based on the learn-from-examples paradigm and that are able to take advantage of the large amount of data available on known protein-ligand pairs.

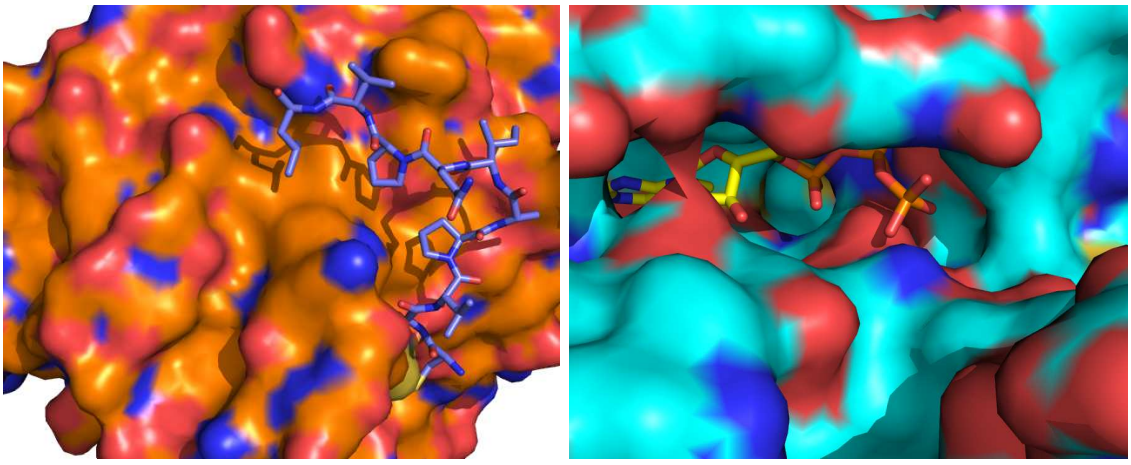
As for protein-peptide interfaces, due to the often disordered nature of the regions involved in binding, shape similarity is no longer a determining factor. Then, in geometry-based methods, geometry is accounted for by providing the relative position of the atoms surrounding the peptide residues in known structures. Finally, also for protein-peptide interfaces, we present a classification of some successful machine learning methods. Indeed, they can be categorized in the way adopted to construct the learning examples. In particular, we envisage three main methods: distance functions, structure and potentials and structure alignment.

1. Introduction

Inferring protein function from structure when the sequence is not conserved is a challenging problem in drug design for which many methods have been developed over the past few years. Sequence based methods for inferring function can be successfully applied when an overall sequence identity of 30% exists between an unknown protein and a characterized one. Unfortunately for about 50% of unannotated molecules (see [67]) no known protein can be found with such degree similarity.

Proteins perform their function by interacting with other molecules, either proteins or ligands. Although a large number of protein structures are available in the Protein Data Bank (PDB), relatively fewer structures of proteins in complex with ligands or peptides exist. The experiments used to derive such structures are difficult especially for large complex sizes. Computational methods can aid and complement experimental techniques for fast identification of putative sites for experimental validation. One approach to derive the structure of a complex is "docking" where a molecule is docked to a target protein to verify the their geometrical and chemical fitting. An alternative approach to derive the function of uncharacterized protein structures is binding site recognition. This strategy is based on the principle that, if a surface region of one protein is similar to that of the binding site of another protein with known function, the function of the one protein can be inferred and its interaction with the known molecule predicted.

In this review, we deal with the identification of promising binding sites and focus on the interactions involving small interfaces which typically occur in protein-ligand or protein-peptide binding (see e.g. Figure 1). We will not be concerned with large protein-protein interactions as those occurring between protein domains within the same structure or within large stable complexes. Small interactions are generally transient and reversible; by contrast, large interactions tend to be permanent.



(a) (b)

Figure 1: Docking position of: (a) dipeptide-chymotrypsin with protein 1ab9 and (b) ligand Adenosine-5'-triphosphate (ATP) with protein 1atp.

Predicting protein-ligand interaction has been the subject of many studies over the two decades. Among the most common studied ligands are ATP, NAD, HEME. They exhibits different degree of conformational variability when binding to a protein. For instance, ATP shows many different

shapes, from an extended conformation to a compact one. Ligands such as steroids tend to be less flexible in their structure. The protein-peptide binding prediction has been less extensively studied. Protein-peptide interactions often involve disordered peptides which may undergo a disorder-to-order transition upon binding. An example is that of a kinase protein interacting with a short peptide of another protein. The low specificity of these small molecules, due to the shape variability of the ligands and the disorder of peptides, renders the prediction problem very difficult.

Finding the interface region of a target protein is conceptually similar when a small ligand or a short peptide is considered. The great majority of methods for binding site recognition and prediction hinge on a fundamental assumption, namely that the properties of a protein region in general, or binding site more in particular, depend upon the geometry and/or physico-chemical property of some or all the residues in a more or less local neighborhood of the site of interest. Although geometry plays a role in the computational methods used in both cases of ligands and short peptides, the way it is used somewhat differs. In protein-ligand binding site prediction, the shape of the binding site is important for recognition since it is often conserved for different proteins bound to the same ligand. Thus a number of approaches, referred to as shape-based, characterize ligand binding sites using geometric descriptors, either local or global, to determine shape similarity with known proteins. Other approaches compare surface regions by finding the largest number of atoms/ residues in the two regions that are spatially similar. These latter approaches fall into two categories: alignment based (where typically the transformation that best superimposes two structures is determined) and graph theoretic approaches. In the case of protein-peptide binding site, due to the often disordered nature of the regions involved, shape similarity is no longer a determining factor. Instead geometry is accounted for by providing the relative position of the atoms surrounding the peptide residues in known structures.

A class of methods that have been employed in both protein-ligand and protein-peptide prediction with a reasonable degree of success is machine learning methods. The learn-from-examples paradigm is particularly fit for this problem. Indeed, learning machines are naturally able to take advantage and exploit the large amount of available data on known protein-ligand and protein-peptide bindings represented by diverse databases of annotated proteins.

In the following, we characterize the different types of interfaces (section 2). Then section 3 and 4 present methods for the prediction of protein-ligand and protein-peptide interaction, respectively.

2. Types of interfaces

Protein-ligand

In biochemistry, the definition of ligands is rather general, it refers to various molecules that perform a variety of functions in association with proteins. Ligands include substrates, inhibitors, activators, and neurotransmitters. Knowledge of ligand-protein binding sites has important implications in functional prediction and in structure-based drug discovery.

Ligand conformation has been extensively analyzed focusing on certain ligands, such as ATP, NAD, HEME and steroids. It has been observed that the ligands exhibit significant conformational variability upon binding; for instance the conformation of ATP in different complexes may range from an extended one to a very compact one, with intermediate arrangements also possible [68]. Interestingly, some conformations are not even close to an energy minimum. This behavior renders the prediction problem difficult, because the corresponding binding sites differ

in shape, size and chemical composition. However, one feature that seems common to a variety of interactions is that they tend to occur in areas of the protein surfaces that correspond to pockets/cavities, often the largest one (see e.g. Figure 1(b)). The former consideration is at the base, and indeed motivates, the many algorithms for protein-ligand binding site prediction. Almost all of the algorithms proposed in the literature in recent years search for pockets/cavities on the protein surface, see e.g. the recent survey paper [36].
Many databases exist on protein-ligand complexes. The database described in [7] has about 10,000 protein-ligand crystal structures. All biologically relevant ligands are annotated, and experimental chemical binding-affinity data is reported when available. Databases for biologically relevant ligand-protein interactions derived from solved structures from the Protein Data Bank (PDB) are presented in [79], [38],[76] (PSMD), they have a different degree of redundancy and manual update and maintenance. The Pocketome [53] is an encyclopedia of conformational ensembles of all druggable binding sites that can be identified experimentally from co-crystal structures in the PDB. It contains 2,227 entries in total, at least 943 entries from mammals. The database of protein-chemical structural interactions [38] includes all existing 3D structures of complexes of proteins with low molecular weight ligands. The LigFam database [25] is a comprehensive collection of a high-quality manually curated protein ligand interactions and functional information. The database also contains structure-guided alignments (identifies important conserved interactions across families of homologous proteins), ligand conformations (crucial for drug-design), SNP, mutation, and disease information (identifies important variations in the protein-ligand complex that could affect function as shown by mutation studies). The database [37] contains 4 million similar pairs of binding sites obtained with a method known as Pocket Similarity Search using Multiple-Sketches (PoSSuM); it includes all the discovered pairs with annotations of various types (e.g., CATH, SCOP, EC number, Gene ontology).

Protein-peptide

The interactions of proteins or their domains with short peptides are ubiquitous and play a major role in cellular function, particularly in signaling and regulatory processes. Their implication in human diseases and cancer draws the interest of many research groups. Furthermore, peptides are considered as potential drug candidates and synthetic peptides have been designed and marketed for a variety of diseases [46, 74].
According to [52], protein-peptide interactions could explain up to 15-40% of the interactome. Peptide sequences are generally at the interfaces, in terms of linear motifs present at the binding regions, here we will only review the work done on structural characterization and its use in the prediction of protein-peptide and protein-ligand binding sites.
unspecific in isolation and not well preserved throughout evolution. As for their structures, the peptides tend to be flexible and reside in disordered regions of proteins and gain structure upon binding. On the other hand, little conformational change of the unbound proteins upon binding to the peptide is observed [63] when comparing them to the same proteins bound to peptides.
Much work on the prediction of protein-peptide interaction has focused on short peptides binding well characterized protein domains such as SH2, SH3, PDZ, and PTB. Such domains differ in their binding specificity, i.e. in their ability to distinguish between different peptides. Highly specific domains only interact with few peptides, generally characterized by specific sequence motifs, while domains with low specificity may, under different circumstances, bind to a variety of peptides, especially when the domain appears in combination with other domains. A description of the binding specificity of domains and how this information can be used in modeling and in

the prediction of interactions is in [26].

The more challenging problem of identifying the binding site of a generic peptide without any restriction on the protein family/domain has been addressed only recently once a significant number of structures of complexes between proteins and short-peptides have started to accumulate. A recent structural database of protein-peptide complexes, PeptiDB, contains 103 high-resolution, non-redundant complexes of proteins with short (5-15Å long) peptides. The DOMINO database of protein-peptide interactions currently contains 200 peptide binding domains, with 10,800 interactions in human. Other structural databases include PEPX [73], PhosphoELM [27], Minimotif Miner [48], SCANSITE [56], PhosphoMotif Finder [1], MHCPEP [32], and PepBank [65].

Based on these databases, a characterization of protein-peptide interfaces could be obtained in terms of variety of geometric and physico-chemical properties. One such property is the average solvent-accessible surface area that is buried upon binding (ASA) [63]. This area is generally small, for instance half the size as in protein-protein interactions. The shape of the interface tend to be more planar (see e.g. Figure 1(a)) than that of a protein-ligand interface, even when the binding occurs in a pocket on the surface of the protein, often the largest pocket. Peptides display interfaces that are better packed than protein-protein interfaces and contain significantly more hydrogen bonds, mainly those involving the peptide backbone [47].

3. Protein-Ligand Binding Site Prediction

There is a long and rich list of papers dealing with the identification of binding sites of ligands and their classification. Early work in this area dates back to the late 90' and beginning of 00's and often represented the extension of work done for the comparison of proteins folds and their alignment. For instance, geometric hashing techniques were initially applied to the comparison of entire protein structures and later adapted to the case of binding site recognition [66]. Similarly, graph-theoretic approaches were used in both instances of the comparison problem. As a further example, spherical harmonics representation of molecules provided the basis for a comparison of globular proteins as well as of binding sites. Obviously the adaptation was generally far from trivial, as the reduction of the data to be compared was counter-balanced by the difficulty in dealing with the conformational flexibility of the interacting molecules. Starting from a somewhat limited repertoire of computational methods, a number of strategies have capitalized on their integration in conjunction with the use of data from different sources to obtain more robust and faster solutions. For instance, some approaches combine sequence and structure information [6, 10, 11, 80] while others integrate structural information with physico-chemical properties [34, 39, 49, 66, 40]. To improve the performances of these methods, when the structure of the ligand is known, docking and scoring functions can be used to calculate the binding affinity for a specific ligand. Some of these techniques have been made available over the web [4, 2,].

The structural datasets of proteins as well as the set of ligands used differ in the numerous papers published over the years making a fair comparison between the different approaches somewhat complicated. As surveys exist on earlier works [41, 31, 78], especially on surface pocket identification, rather than trying to summarize all the work in this field, we will concentrate on few approaches.

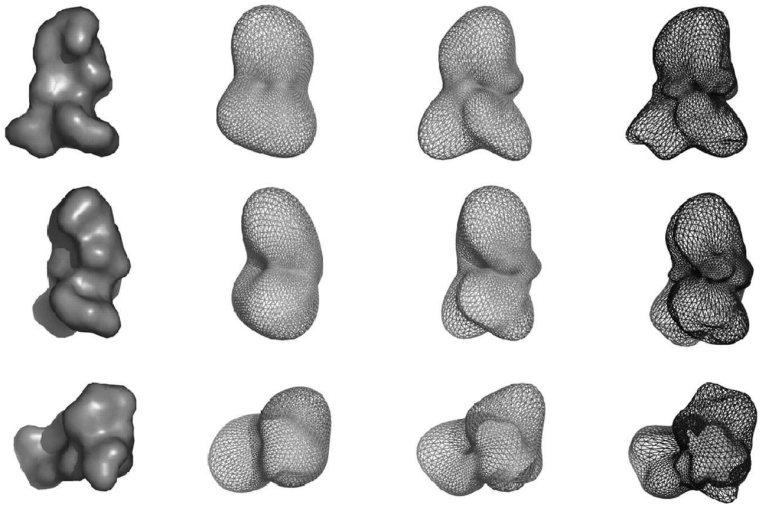


Figure 2: Examples of surface representation by spherical harmonics. The left column shows different views of a predicted binding pocket in 1b14. The second column shows the reconstructed 3D shape using spherical harmonics of order 4. The third column shows the reconstruction up to order 6. The fourth column shows the reconstructed model up to 14th order.

3.1. Shape-based methods

It has long been recognized that shape plays an important role in binding site recognition. These methods employ geometric information about the protein surface sometimes enriched by physico-chemical description. Among the most popular shape-based methods are geometric hashing [55, 24] which uses a set of distance constraints to detect structural motifs without any assumption on the substructure of the functional site. Some other widely used methods include the Template Search and Superposition (TESS) method [75], JESS [6], the Fuzzy Functional Forms (FFF) method [23] and Q-SiteFinder [42].

In the following we concentrate on approaches to protein binding site comparison and localization using spherical harmonics [18, 43, 62] and spin images [12].

Spherical harmonics have long been used as global shape descriptors to represent closed surfaces (see e.g. Figure 2). In biochemistry and computational biology they have been applied to the modeling, visualization and comparison of globular proteins.

Molecular shapes are approximated as functions defined on the unit sphere by representing each surface point by its spherical coordinates (r, θ, ϕ) and giving the radial coordinate r as a function $f(\theta, \phi)$ of the two angle coordinates θ and ϕ. Spherical harmonics are sets of differentiable, complex functions $Y_m^l(\theta, \phi)$ of two variables, θ and ϕ, indexed by two integers, l and m. They form a complete set of orthonormal functions and thus any function f of θ and ϕ can be expanded as $f(\theta, \phi) = \sum_{l \in N, m \in [-l,l]} c_m^l Y_m^l(\theta, \phi)$, where c_m^l are the spherical harmonic coefficients. The expansion coefficients describe the shape at different levels of resolutions, from coarse to fine corresponding to low and high coefficients, respectively.

As global shape descriptors, spherical harmonics require some adaptation to be applied to the problem of binding site recognition. For instance, in one of the earliest approach to binding site recognition [15] a single sphere is placed at the center of the binding site, typically residing in a surface cavity, and then "inflated" until it approximates the shape of the cavity. Then the

comparison of two binding sites is performed by comparing two spherical functions as in the problem of comparing two complete molecular surfaces [62].
In [43] protein shapes are classified based on the similarity of the expansion coefficients of the spherical harmonic representation, typically restricted to the low-order ones. The ℓ_2 distance is chosen as a measure of dissimilarity in coefficient space. Basically the same idea was used in [50, 69] to compare and cluster protein binding sites. The comparison is performed on the aligned binding sites to avoid the determination of the rotation for the spherical harmonic representation at the expense of possible loss of accuracy.
In [18] a method to quickly identify promising binding sites, either in a protein cavity or on an entire protein surface, without explicitly aligning them is presented, i.e., without actually computing the optimal rotation that best overlaps two binding sites. To represent a given binding site, the notion of *Binding Ball*, a spherical description of a query binding site, is introduced. After creating the Binding Ball for a given query binding site, it is "rolled" over a protein's surface to be examined, and each position is efficiently scored, evaluating all possible rotations at that location simultaneously, by making use of a specific property of the Spherical Fourier Transform (SFT).
The problem of identifying regions of similarity on proteins has been addressed in [12] based on a spin-image representation of molecular surfaces. A protein is described by a collection of two-dimensional images, called spin images, each associated to a surface point. The spin image of a surface point P is a two-dimensional accumulator array that represents all the surface points in a reference frame defined by point P and its normal n. The matching strategy is based on the observation that surfaces with similar shape tend to have similar spin images, thus reducing a complex 3D matching problem into a 2D problem for which a simpler and more efficient solution exists.
Morphological as well as topological properties of protein surfaces are used in [16] to describe protein pockets for ligand positioning.

3.2. Alignment-based methods

The comparison of protein surfaces, cavities, or of binding sites has often been addressed as an alignment problem, i.e. finding atom pairs on two protein surfaces that occupy spatially equivalent positions. It can be formulated as follows: given two sets A and B of points, find two possibly large subsets A' of A and B' of B with high degree of *similarity*. There are various ways of defining the similarity between two point sets in 3D space leading to the proposal of different distance functions and associated algorithms; they include the root mean square distance (RMSD), the closest point distance [9], the bottleneck distance [21]. The alignment problem often involves the search for the isometric transformation which best superimposes two given protein structures.
Alignment-based methods may incorporate information about chemical properties of the matched atoms thereby improving the performance and the accuracy of the results.
An important aspect of the matching is the choice of a suitable surface representation; in the literature common ways of representing a surface are Connolly's representation [19], alpha-shapes [44] and pseudo-vertices [66]. In the simplest instance, the surface is represented as a cloud of points, each corresponding to a surface atom.
One way to solve the surface alignment problem is by using the Iterative Closest Point (ICP) algorithm [9], originally introduced for image registration. Briefly, ICP aligns (registers) two sets of points P and Q by iteratively alternating between registration and alignment steps. In the

registration step, for each point in P its closest point in the other set Q is determined, resulting in the subset Y of Q. An alignment is then obtained between P and Y by finding the rotation and translation that best superimpose P and Y. These two steps are repeated until the change in root mean square distance between P and Y is below a selected threshold.
In [8] the ICP is applied to the problem of detecting similarity in binding sites for classification purposes. It searches for the isometric transformation that best superimposes active regions of two structures and provides a list of matched atoms along with their RMSD. The ICP method is solved by a global optimization algorithm belonging to the class of controlled random search methods [14, 17, 58]. These methods, although heuristic in nature, are very efficient and reliable for the global minimization of nonlinear multivariate functions of several variables. The dissimilarity measure proposed is based on the solution to an *Asymmetric Assignment Problem* on a bipartite graph associated to the matching problem [8].
A Triangulation-based Iterative-closest-point for Protein Surface Alignment (TIPSA) was proposed in [22]. TIPSA seeks to determine the maximum number of atoms that can be superposed between two protein binding sites, with the constraint that any pair of matched atoms has a distance below a given threshold. The heuristics employed obtains a solution by expanding a seed that consists of similar tetrahedrons between two binding sites obtained from 3D Delaunay triangulation.

3.3. Graph-theoretic methods

Another class of methods involve graph-theoretic methods [3, 35, 39, 51, 64, 77]. They were first proposed in the context of computational chemistry and data mining, and are characterized by a completely different problem formulation and, consequently, algorithmic solutions with respect to the methods reviewed in the previous sections. They start by transforming the protein structures to graphs as follows. The protein structures are described by sets of 3D points (atoms, residues, or other interesting points) and their relationships (typically, proximity). They are represented by graphs where nodes correspond to points and edges connect related points. Nodes of the graph can have associated properties, either geometrical or physico-chemical. The graph structure is referred to as a *protein contact graph* when nodes correspond to residues and an undirected edge is present between two nodes if the corresponding residues are within a given distance from each other.
Graph theory offers several algorithms to establish relationships between structural data, such as graph and sub-graph isomorphism, maximal bipartite graph matching and clique detection. In structural bioinformatics, graph algorithms and statistical inference are used to find representative subgraphs or score similarities between neighborhoods of residues. They exploit graph similarities measures to discover functional sites (e.g. the graphlet kernel method proposed in [71] or the graph-based clique detection (GG) algorithm proposed in [20]).
The theoretical aspects of the optimal matching of binding sites are discussed in [64] where the general problem of finding the largest common subset of two or multiple sets of points is shown to be NP-hard. Therefore approximations are presented to efficiently address the problem of comparing two or multiple binding sites via a combination of geometric hashing and k-partite matching.
In [35] an approach is presented that uses maximal common subgraph comparison and harmonic shape image matching to detect locally similar regions between two molecular surfaces augmented with properties such as the electrostatic potential or lipophilicity. The complexity of the problem is reduced by a set of filters that implement various geometric and physico-chemical

heuristics.
Another way to solve the matching problem is to construct an auxiliary data structure, the *association graph*, where nodes represent pairs of candidate corresponding points on two structures which are selected based on their geometrical or chemical similarity. Edges connect consistent correspondences, where consistency is expressed either in terms of chemical similarity or similarity of Euclidean distances between pairs of corresponding points. Given the association graph, the matching problem is formulated as the problem of finding a maximal clique in it and solved with one of the several heuristics developed for this classical graph problem.
The association graph approach is used in [51] to detect nearly-optimal approximate solutions for the graph-matching problem. A two step-heuristic procedure that uses different levels of resolutions of the structure representations, from residues to atomic representation, allows a fast implementation of the maximal clique detection thus enabling the comparison of large binding sites.
In [71] a graph-based kernel method is proposed for annotating functional residues in protein structures represented by contact graphs. The method uses the graphlet representation of every vertex in a graph [59]. Briefly, graphlets are small connected non-isomorphic induced subgraphs of a graph. Their frequency is an important measure of graphs both at the local and global level. A similarity measure between two nodes is expressed as the inner product of their respective frequency vectors and is used in a supervised learning framework to classify protein residues. The method was applied to the identification of catalytic residues in proteins, as well as to the problem of predicting phosphorylation sites in protein structures.
The problem of multiple graph alignment for active sites characterization is also addressed in [77] using inexact graph-matching techniques. Optimized algorithms are presented for the efficient calculation of multiple graph alignments for the analysis of physico-chemical descriptors representing protein binding pockets.
The interesting results of the above graph-based methods is that in several cases it was possible to identify structural features that are characteristic for a given protein family and allow to discriminate among related families. In other words, for selected high-quality datasets of proteins from the PDB the proteins that bind similar ligands could be predicted and separated from those binding different ligands based only on local atomic similarities.

3.4. Machine learning methods

This rather wide and heterogeneous class contains those methods that adopt a machine learning technique for identification and prediction of binding site regions or residues. To this aim, almost all these methods employ information describing the surrounding environment of a single residue or position in the protein structure. Usually, they collect a set of structural, physico-chemical and evolutionary properties which are then encoded into a fixed-length vector. Then, it is possible to compose labeled or non-labeled sets of training vectors for supervised or unsupervised, respectively, machine learning approaches. The pioneering idea from which the methods belonging to this class have originated, can be glimpsed in the paper [5] where the idea emerged of characterizing protein sites by a set of geometry and physico-chemical properties of surrounding residues.
The approaches presented in the literature mainly differ in two fundamental aspects. First, in the definition of the structure of samples that populate the environment where the learning machine is embedded. In particular, the definition of the features that describe a sample, i.e. the input vectors. Second, in the definition of the test set that will be used to train the learning

machine and to assess the validity of the predictions.
In [30], for instance, a neural network is trained to predict the location of binding sites in enzymes by employing structure and sequence information. More in particular, solvent accessibility, type of secondary structure, depth and cleft where the residue lies in, conservation score [72] and residue type are used as inputs to the neural networks.
In [13] a different yet powerful learning machine is used, namely random forest classifiers, to predict small ligand binding sites. The features used to define the input vector to the learning machine are evolutionary conservation, median solvent accessible surface area, counts of nearby residue pairs and statistically significant clustering of residue types in the site. The proposed method, i.e. SiteFinder, is able to accurately predict the binding site both for small molecules and metal ions.

4. Protein-Peptide interface prediction

Until very recently, protein-peptide docking and protein-peptide interaction recognition were regarded as subproblems of the well-known protein-protein docking problem. Indeed, a widely used technique to solve the protein-peptide interaction problem was that of applying protein-protein docking, thus considering the peptide as a protein itself. However, this approach has limited applications for peptides longer than 4 residues largely due the high degree of flexibility that has to be considered when docking typical peptides of 5-10 residues to a protein.
Many approaches for protein-peptide binding site recognition and prediction rely on some prior knowledge of the type of peptide binding to a domain and often require further knowledge of the peptide binding site on the protein. Hence, these approaches are generally only effective for finding new variants of known peptides, and cannot directly uncover new protein-peptide interaction types.

4.1. Geometry-based methods

Describing the binding environment S-PSSM

Given a dataset of non-redundant protein structures in complex with peptide segments, their binding sites provided the basis for a spatial characterization of peptide residues to be used for the prediction of binding sites of uncharacterized proteins. In [57] the binding environment was described by creating spatial position specific scoring matrices, called S-PSSM (which are an extension of PSSM [29]), for each of the 20 amino acids and the three phosphorylated variants. For each amino acid type, the structures of proteins interacting with all peptides containing that amino acid were analyzed. They were all superimposed and used to build a grid of occupancy for selected atom types, indicating the atom's preference to be in a particular spatial position relative to the considered amino acid type. Once built, the S-PSSMs were used for the prediction of candidate binding sites on the surface of proteins for a newly discovered peptide. More precisely, given a target peptide, each amino acid belonging to it was in turn considered and the positions on the surface of a protein best matching its corresponding S-PSSM matrices identified. The output of this process was a set of residue positions spread over the surface of the protein each corresponding to an amino acid of the type present in the peptide. Further analysis, imposing geometric constraints based on distances of such positions, was used to select and link among the candidate sites those consistent with all residues of the target peptide. This method was benchmarked on a relatively large set of unbound proteins showing good accuracy. A web

server based on this method and using an aggregate scoring measure for each amino acid type is available [70].
S-PSSM can also be used to describe binding patterns (as in [61]). Then, the problem of identification can be approached by constructing a probabilistic model of the likelihood of generating the residue sequences of the binding site. The probabilistic neural network is then trained (compute values of the parameters of the model) by using the Gibbs sampling algorithm [45] by using as training set a set of 285 interactions between 28 SH3 proteins and 143 SH3 binding partners.

4.2. Machine learning-based methods

Describing the binding by distance functions

A different and reportedly better approach can be that of learning peptide-peptide distance functions [33], indeed following the observation that peptides that bind to the same protein are "similar" to one another, and different from non-binding peptides. The distance function thus learned can then be used to compute a protein-peptide binding affinity function, that is the affinity of a given peptide to a protein of a certain family. More in particular, given a dataset of binding and non-binding peptides from an entire protein family, the method first extracts, for each protein, positive and negative equivalence constraints. Then, a single peptide-peptide distance function is learned using the DistBoost algorithm [32] which is a semi-supervised approach for clustering and learning. Finally, the protein-peptide affinity function is computed by using the peptide-peptide distance function. The reported results show that this method on the MHCPEP dataset outperforms two competing algorithms using PSSM and adopted in [60].

Describing the binding by structure and potentials

For the PDZ and SH3 domains, an approach based on a committee of learning machines is proposed in [81]. In particular, protein-peptide interaction is described by three types of data: (a) an orthogonal dot product encoding, that is the tensor product between the characteristic vectors representing the interface and peptide residue, respectively; (b) matrix of structurally interacting potentials, that is the interaction energies between the naturally occurring aminoacid residues; (c) the coding of physico-chemical properties, in which each aminoacid was represented by five physico-chemical properties. More precisely, each property is classified into five groups. For one property, one aminoacid is assigned to class $1, \ldots, 5$ and pairing of two aminoacids, w.r.t that property, gives rise to 15 combinations which are coded using 15 binaries. Since each aminoacid has 5 properties, we have as much as $15 \times 5 = 75$ combinations coded with 75 binaries. Then the interaction between m protein interface residues and n peptide residues is coded with $m \times n \times 75$ binaries. These three data types are used to train a committee of learning machines, two SVMs and a Probabilistic Neural Network (PNN), and the final prediction is given by the consensus of the three separate predictions.

Describing the binding by structure alignment

A further possibility to describe the binding between a protein and a peptide consists in employing a suitable similarity measure between peptides and protein binding pockets. More in particular, in [28] a so-called Generic String (GS) kernel is introduced and a Kernel Ridge Regression, i.e. a particular learning machine, is defined to the aim of predicting the protein-peptide binding affinity energy. The GS kernel definition takes advantage of a sequence-independent structure alignment heuristic which considers proteins secondary structure. In [28] three datasets, namely

PEPX [73], a dataset proposed in [54] for binding prediction in Major Histocompatibility Complexes of class II (MHC-II) and a benchmark for Quantitative Structure Affinity Modeling [82], are used to validate the results and show superiority of the approach with respect to other alternatives in the literature.

5. Conclusions

In structural bioinformatics, among the possible ways to infer protein function, one which benefits from computational techniques is the prediction of interaction among proteins and other molecules. Here we have considered interactions involving small molecules and reviewed methods for binding site recognition. We have not tried to review the huge amount of work done in this area but have concentrated on some approaches by separating them in few categories. Prediction methods generally rely on prior knowledge of protein binding sites accumulated over the years in the PDB. They can be effective in uncovering binding sites on proteins of unknown function similar to those of other proteins. Although they are tolerant of some variation in topology and physico-chemical composition, they cannot lead to the discovery of completely new interface patterns, assuming there are still some to be discovered.

References

[1] R. Amanchy, B. Periaswamy, S. Mathivanan, R. Reddy, S.G. Tattikota, and A. Pandey. A compendium of curated phosphorylation-based substrate and binding motifs. *Nature Biotechnology*, 25(3):285–286, 2007.

[2] M. Jambonand O. Andrieu, C. Combet, G. Deleage, F. Delfaud, and C. Geourjon. The sumo server: 3d search for protein functional sites. *Bioinformatics*, 21(20):3929–3930, 2005.

[3] P.J. Artymiuk, R.V. Spriggs, and P. Willett. Graph theoretic methods for the analysis of structural relationships in biological macromolecules. *Journal of the American Society for Information Science and Technology*, 56(5):518–528, 2005.

[4] G. Ausiello, P.F. Gherardini, P. Marcatili, A. Tramontano, A. Via, and M. Helmer-Citterich. Funclust: a web server for the identification of structural motifs in a set of non-homologous protein structures. *BMC Bioinformatics*, 9(Suppl. 2):S2, 2008. doi:10.1186/1471-2105-9-S2-S2.

[5] S.C. Bagley and R.B. Altman. Characterizing the microenvironment surrounding protein sites. *Protein Science*, 4(4):622–635, 1995.

[6] J.A. Barker and J.M. Thornton. An algorithm for constraint-based structural template matching: application to 3d templates with statistical analysis. *Bioinformatics*, 19(13):1644–1649, 2003. DOI: 10.1093/bioinformatics/btg226.

[7] M.L. Benson, R.D. Smith, N.A. Khazanov, B. Dimcheff, J. Beaver, P. Dresslar, J. Nerothin, and H.A. Carlson. Binding moad, a high-quality protein-ligand database. *Nucleic Acids Research*, 36(Database issue):D674–678, 2008. doi: 10.1186/1471-2105-11-488.

[8] P. Bertolazzi, C. Guerra, and G. Liuzzi. A global optimization algorithm for protein surface alignment. *BMC Bioinformatics*, 11:488, 2010. doi: 10.1186/1471-2105-11-488.

[9] P.J. Besl and N.D. McKay. A method for registration of 3-d shapes. *IEEE Trans. on Pattern Analysis and Mach. Intelligence*, 14:239–255, 1992.

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

[11] T.A. Binkowski, A. Joachimiak, and J. Liang. Protein surface analysis for function annotation in high-throughput structural genomics pipeline. *Protein Science*, 14(12):2972–2981, 2005.

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

[13] A.J. Bordner. Predicting small ligand binding sites in proteins using backbone structure. *Bioinformatics*, 24(24):2865–2871, 2008.

[14] P. Brachetti, M. De Felice Ciccoli, G. Di Pillo, and S. Lucidi. A new version of the Price's algorithm for global optimization. *Journal of Global Optimization*, 10:165–184, 1997.

[15] W. Cai, X. Shao, and B. Maigret. Protein ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. *Journal of Molecular Graphics and Modelling*, 20:313–328, 2002.

[16] V. Cantoni, A. Gaggia, R. Gatti, and L. Lombardi. Geometrical constraints for ligand positioning. *Proceedings of Bioinformatics - BIOSTEC* ., pages 26–29, 2011.

[17] L. Cirio, S. Lucidi, F. Parasiliti, and M. Villani. A global optimization approach for the synchronous motors design by finite element analysis. *Journal of Applied Electromagnetics and Mechanics*, 16:13–27, 2002.

[18] M. Comin, F. Dellaert, and C. Guerra. Binding balls: Fast detection of binding sites using a property of spherical fourier transform. *Journal of Computational Biology*, 16(11):1577–1591, 2009.

[19] M.L. Connolly. Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16:548–558, 1983.

[20] H. Deng, G. Chen, W. Yang, and J.J. Yang. Predicting calcium- binding sites in proteins - a graph theory and geometry approach. *Proteins*, 64(1):34–42, 2006.

[21] A. Efrat, A. Itai, and M.J. Katz. Geometry helps in bottleneck matching and related problems. *Algorithmica*, 31:1–28, 2001.

[22] L. Ellingson and J. Zhang. Protein surface matching by combining local and global geometric information. *PLoS ONE*, 7(7):e40540, 2012. doi:10.1371/journal.pone.0040540.

[23] J.S. Fetrow and J. Skolnick. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and t1 ribonucleases. *Journal of Molecular Biology*, 281(5):949–968, 1998.

[24] D. Fischer, H.J. Wolfson, S.L. Lin, and R. Nussinov. Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. *Protein Science*, 3(5):769–778, 1994.

[25] Georgetown University Medical Center. Innovation center for biomedical informatics. (URL)http://icbi.georgetown.edu/biomedical/drug-discovery/ligand/.

[26] D. Gfeller. Uncovering new aspects of protein interactions through analysis of specificity landscapes in peptide recognition domains. *FEBS Letters*, 586(17):2764–2772, 2012.

[27] T. Gibson, F. Diella, H. Dinkel, K. Gould, C. Gemünd, C. Chica, S. Cameron, and N. Blom. Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins.
(URL)http://phospho.elm.eu.org/.

[28] S. Giguère, M.M., F. Laviolette, A. Drouin, and J. Corbeil. Learning a peptide-protein binding affinity predictor with kernel ridge regression. *BMC Bioinformatics*, 14(82), 2013. doi:10.1186/1471-2105-14-82.

[29] M. Gribskov, A.D. McLachlan, and D. Eisenberg. Profile analysis: Detection of distantly related proteins. *PNAS, Proceedings of the National Academy of Sciences*, 84:4355–4358, 1987.

[30] A. Gutteridge, G.J. Bartlett, and J.M. Thornton. Using a neural network and spatial clustering to predict the location of active sites in enzymes. *Journal of Molecular Biology*, 330(4):719–734, 2003.

[31] S. Henrich, M. H. Outi Salo-Ahen, B. Huang, F.F. Rippmann, G. Cruciani, and R.C. Wade. Computational approaches to identifying and characterizing protein binding sites for ligand design. *Journal Molecular Recognition*, 2009.

[32] T. Hertz, A. Bar-Hillel, and D. Weinshall. Learning distance functions for image retrieval. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR04)*, 2004. Washington DC (USA).

[33] T. Hertz and C. Yanover. Pepdist: A new framework for protein-peptide binding prediction based on learning peptide distance functions. *BMC Bioinformatics*, 7(S3), 2006. doi:10.1186/1471-2105-7-S1-S3.

[34] C. Hofbauer and A. Aszódi. Sh2 binding site comparison: a new application of the surfcomp method. *Journal of Chemical Information and Modeling*, 45(2):414–421, 2005. DOI: 10.1021/ci0497049.

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

[36] B. Huang. Identification of pockets on protein surface to predict protein-ligand binding sites. *Focus on Structural Biology*, 8:25–39, 2013.

[37] J. Ito, Y. Tabei, K. Shimizu, K. Tsuda, and K. Tomii. Possum: a database of similar protein-ligand binding and putative pockets. *Nucleic Acids Research*, 40(Database issue):D541–548, 2012. http://possum.cbrc.jp/PoSSuM/database.html.

[38] O.V. Kalinina, O. Wichmann, G. Apic, and R.B. Russell. Combinations of protein-chemical complex structures reveal new targets for established drugs. *PLoS Computational Biology*, 7(5):e1002043, 2011.

[39] N. Kinoshita, J. Furui, and H. Nakamura. Identification of protein functions from a molecular surface database, ef-site. *J. Struct. Funct. Genomics*, 2:9–22, 2001.

[40] D. Kuhn, N. Weskamp, S. Schmitt, E.H. Hullermeier, and G. Klebe. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *Journal of Molecular Biology*, 359:1023–1044, 2006.

[41] A.T.R. Laurie and R.M. Jackson. Methods for the prediction of protein ligand binding sites for structure-based drug design and virtual ligand screening. *Current Protein Peptide Science*, 21(9):1908–1916, 2005.

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

[43] S.E. Leicester, J.L. Finney, and R.P. Bywater. A quantitative representation of molecular surface shape. i: Theory and development of the method. *Journal of Mathematical Chemistry*, 16:315–341, 1994.

[44] J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and caviteis: measurement of binding site geometry and implications for ligand design. *Protein Science*, 7:1884–1897, 1998.

[45] J.S. Liu, A.F. Neuwald, and C.E. Lawrence. Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *Journal of the American Statistical Association*, 90:1156–1170, 1995.

[46] A. Loffet. Peptides as drugs: Is there a market? *Journal of Peptide Science*, 8(1):1–7, 2002. doi: 10.1002/psc.366.

[47] N. London, D. Movshovitz-Attias, and O. Schueler-Furman. The structural basis of peptide-protein binding strategies. *Structure*, 18(2):188–199, 2010. doi: 10.1016/j.str.2009.11.012.

[48] T. Mi, J.C. Merlin, S. Deverasetty, M.R. Gryk, T.J. Bill, A.W. Brooks, L.Y. Lee, V. Rathnayake, C.A. Ross, D.P. Sargeant, C.L. Strong, P. Watts, S. Rajasekaran, and M.R. Schiller. Minimotif miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. *Nucleic Acids Research*, 40(D1):D252–D260, 2012. doi:10.1093/nar/gkr1189.

[49] R. Minai, Y. Matsuo, H. Onuki, and H. Hirota. Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interaction. *Proteins: Structure, Function, and Bioinformatics*, 72:367–381, 2008.

[50] R.J. Morris, R.J. Najmanovich, A. Kahraman, and J.M. Thornton. Real spherical harmonic expansion coefficients as 3d shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, 21:2347–2355, 2005.

[51] R. Najmanovich, N. Kurbatova, and J. Thornton. Detection of 3d atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics*, 24(16):105–111, 2008.

[52] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. De Masi, T.J. Gibson, J. Lewis, L. Serrano, and R.B. Russell. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biology*, 3(12):e405, 2005. doi: 10.1371/journal.pbio.0030405.

[53] G. Nicola, C.A. Smith, and R. Abagyan. New method for the assessment of all drug-like pockets across a structural genome. *Journal of Computational Biology*, 15(3):231–240, 2008.

[54] M. Nielsen, C. Lundegaard, T. Blicher, B. Peters, A. Sette, S. Justesen, S. Buus, and O. Lund. Quantitative predictions of peptide binding to any hla-dr molecule of known sequence: Netmhciipan. *PLoS Computational Biology*, 4(7):e1000107, 2008. http://dx.plos.org/10.1371

[55] R. Nussinov and H.J. Wolfson. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *PNAS, Proceedings of the National Academy of Sciences*, 88(23):10495–10499, 1991.

[56] J.C. Obenauer, L.C. Cantley, and M.B. Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Research*, 31(13):3635–3641, 2003.

[57] E. Petsalaki, A. Stark, E. García-Urdiales, and R.B. Russell. Accurate prediction of peptide binding sites on protein surfaces. *PLoS Computational Biology*, 5(3):e1000335, 2009. doi: 10.1371/journal.pcbi.1000335.

[58] W.L. Price. A controlled random search procedure for global optimization. In L Dixon and G Szego, editors, *Towards Global Optimization 2*. North-Holland, Amsterdam, 1978.

[59] N. Przulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007. doi: 10.1093/bioinformatics/btl301.

[60] P.A. Reche, J.P. Glutting, H. Zhang, and E.L. Reinher. Enhancement to the rankpep resource for the prediction of peptide binding to mhc molecules using profiles. *Immunogenetics*, 56(6):405–419, 2004.

[61] D.J. Reiss and B. Schwikowski. Predicting protein-peptide interactions via a network-based motif sampler. *Bioinformatics*, 20(S1):i274–i282, 2006.

[62] D.W. Ritchie and G.J.L. Kemp. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *Journal of Computational Chemistry*, 20(4):383–395, 1999.

[63] Rosetta design group. Macromolecular modeling blog^{TM}. (URL)http://rosettadesigngroup.com/blog/742/the-structural-basis-of-peptide-protein-binding-strategies/.

[64] M. Shatsky, A. Shulman-Peleg, R. Nussinov, and H. Wolfson. The multiple common point set problem and its application to molecule binding pattern detection. *Journal of Computational Biology*, 13(2):407–428, 2006.

[65] T. Shtatland, D. Guettler, M. Kossodo, M. Pivovarov, and R. Weissleder. Pepbank–a database of peptides based on sequence text mining and public peptide data sources. *BMC Bioinformatics*, 8:280, 2007.

[66] A. Shulman-Peleg, R. Nussinov, and H.J. Wolfson. Recognition of functional sites in protein structures. *Journal of molecular biology*, 339(3):607–633, 2004.

[67] J. Skolnick and M. Brylinski. Findsite: a combined evolution/structure-based approach to protein function prediction. *Briefings in Bioinformatics*, 10(4):378–391, 2009. doi: 10.1093/bib/bbp017.

[68] G.R. Stockwell and J.M. Thornton. Conformational diversity of ligands bound to proteins. *Journal of Molecular Biology*, 356:928–944, 2006.

[69] G.R. Stockwell and J.M. Thornton. Conformational diversity of ligands bound to proteins. *Journal of Molecular Biology*, 356:928–944, 2006.

[70] L.G. Trabuco, S. Lise, E. Petsalaki, and R.B. Russell. Pepsite: prediction of peptide-binding sites from protein surfaces. *Nucleic Acids Research*, 40(Web Server issue):W423–W427, 2012. doi: 10.1093/nar/gks398.

[71] V. Vacic, L. M. Iakoucheva, S. Lonardi, and P. Radivojac. Graphlet kernels for prediction of functional residues in protein structures. *Journal of Computational Biology*, 17(1):55–72, 2010.

[72] W.S. Valdar. Scoring residue conservation. *Proteins*, 48(2):227–241, 2002.

[73] P. Vanhee, J. Reumers, F. Stricher, L. Baeten, L. Serrano, J. Schymkowitz, and F. Rousseau. Pepx: a structural database of non-redundant protein-peptide complexes. *Nucleic Acids Research*, 38(Database issue):D545–D551, 2010. doi: 10.1093/nar/gkp893. http://pepx.switchlab.org/.

[74] P. Vlieghe, V. Lisowski, J. Martinez, and M. Khrestchatisky. Synthetic therapeutic peptides: science and market. *Drug Discovery Today*, 15(1-2):40–56, 2010. doi: 10.1016/j.drudis.2009.10.009.

[75] A.C. Wallace, N. Borkakoti, and J.M. Thornton. Tess: a geometric hashing algorithm for deriving 3d coordinate templates for searching structural databases. application to enzyme active sites. *Protein Science*, 6:2308–2323, 1997.

[76] I. Wallach and R. Lilien. The protein-small-molecule database (psmdb), a non-redundant structural resource for the analysis of protein-ligand binding. *Bioinformatics*, 25(5):615–620, 2009. http://compbio.cs.toronto.edu/psmdb/.

[77] N. Weskamp, E. Hllermeier, D. Kuhn, and G. Klebe. Multiple graph alignment for the structural analysis of protein active sites. *ACM/IEEE Transactions on Computational Biology and Bioinformatics*, 4(2):310–320, 2007.

[78] F. Xin and P. Radivojac. Computational methods for identification of functional residues in protein structures. *Current Protein and Peptide Science*, 12:456–469, 2011.

[79] J. Yang, A. Roy, and Y. Zhang. Biolip: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Research*, 2012. doi: 10.1093/nar/gks966.

[80] H. Yao, D.M. Kristensen, I. Mihalek, M.E. Sowa, C. Shaw, M. Kimmel, L. Kavraki, and O. Lichtarge. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *Journal of molecular biology*, 326(1):255–261, 2003.

[81] L. Zhang, C. Shao, D. Zheng, and Y. Gao. An integrated machine learning system to computationally screen protein databases for protein binding peptide ligands. *Molecular & Cellular Proteomics*, 5:1224–1232, 2006.

[82] P. Zhou, X. Chen, Y. Wu, and Z. Shang. Gaussian process: an alternative approach for qsam modeling of peptides. *Amino Acids*, 38:199–212, 2010. http://dx.doi.org/10.1007/s00726-008-0228-1.