# Computational Methods for Identification of Functional Residues in Protein Structures

Fuxiao Xin and Predrag Radivojac*

*School of Informatics and Computing, Indiana University, Bloomington, Indiana, USA*

**Abstract:** The recent accumulation of experimentally determined protein 3D structures combined with our ability to computationally model structure from amino acid sequence has resulted in an increased importance of structure-based methods for protein function prediction. Two types of methods for function prediction have been proposed: those that can accurately predict overall biochemical or biological roles of a protein and those that predict its functional residues. Here, we review approaches used for the computational identification of functional residues in protein structures and summarize their applications to a wide variety of problems in functional proteomics, such as the prediction of catalytic residues, post-translational modifications, or nucleic acid-binding sites. We examine four different problems in order to perform a comparison between several recently proposed methods and, finally, conclude by identifying limitations and future challenges in this field.

**Keywords:** Comparative evaluation, functional residue prediction, functional site prediction, protein function prediction, protein structure, structure-based protein function prediction.

## 1. INTRODUCTION

The computational prediction of protein function from 3D structure can be carried out in two ways: by predicting the overall biochemical or biological roles of the molecule, or by identifying specific residues that are necessary for a particular function [1-2]. For example, based on structural similarity with a known DNA-binding protein, it can be hypothesized that a protein also binds DNA without identifying contact residues. On the other hand, the similarity of local structural neighborhoods between some residues in a protein of interest and the DNA-binding residues in other proteins can be used to probabilistically infer that these residues, and consequently the whole protein, are involved in DNA binding. Such similarities may be based on the geometry of the neighborhood, but can also include physicochemical properties and evolutionary conservation.

The global and local strategies of predicting function are complementary because structurally similar proteins can exhibit different functions [2-3] and also because local structural neighborhood methods may result in a large number of false positive predictions. However, in order to understand the molecular mechanism of a protein's biological function it is often necessary to predict functional residues. Such approaches can directly lead to the computational prediction of the molecular basis of disease since many pathologies arise as a consequence of an alteration of protein function [4]. Other applications include computationally aided protein engineering [5] or drug design [6].

*Address correspondence to this author at the School of Informatics and Computing, Indiana University, 901 E. 10th Street, Bloomington, IN 47408, USA; Tel: (812) 856-1851; Fax: (812) 856-1995;
E-mail: predrag@indiana.edu

Currently, the major repository of protein structures, Protein Data Bank (PDB), contains over 65,000 experimentally solved molecular structures, and its size has been increasing steadily in the recent years [7]. Most of the early deposited proteins were associated with at least some notion of function, even if the functional residues remained incompletely identified. However, the advent of structural genomics projects [8] resulted in many structures for which not only functional residues, but also the overall biological function of the molecule is unknown. Most of these proteins were selected as targets because of the low sequence similarity to proteins with solved structures or to provide structural insights into large, functionally diverse superfamilies [9]. Thus, a more complete coverage of the protein sequence/structure space coupled with the increased accuracy of structural modeling [10] have resulted in the growing importance of the protein function prediction from structure [2].

Here, we review recent advances in methodology used to predict functionally important residues from protein structures. We provide classification of such methods and performance comparisons between several tools. We conclude the article by discussing limitations and challenges in this field.

### 1.1. What is a Functional Residue?

The notion of a *functional residue* has been referred to extensively in the literature, but no clear definition has been proposed. Here, we generally consider a residue to be functional if it is necessary for a protein to carry out its biological role. Accordingly, a mutation of a functional residue may lead to an altered function of the entire molecule that can further produce phenotypic effects in the cell. Such functional alterations can be caused by changes in protein stability or dynamics, in which case a protein may adopt a different structure or, more generally, significantly change its

1389-2037/11 $58.00+.00 

probability distribution over the space of 3D conformations. Alternatively, a mutation of a functional residue may not result in observable changes of the protein structure or dynamics but still impact the overall function of the molecule. These situations typically involve residues that directly participate in the chemistry of reactions with a substrate or provide required binding specificity.

A *functional site* is referred to here as a more general concept that includes one or more functional residues that collectively provide desired functionality. Such sites include surface pockets or patches that provide interfaces with ligands or macromolecular partners, catalytic triads for enzymatic activity, etc. When the context permits, we refer to functional residues and functional sites interchangeably. Examples of functional sites are shown in Fig. (**1**).

## 2. METHODS FOR PREDICTING FUNCTIONAL RESIDUES IN PROTEIN STRUCTURES

Methods designed to computationally identify functional sites from protein structures primarily consider local structural neighborhoods surrounding the sites of interest. The goal is to capture particular geometry, physicochemical properties and/or patterns of evolutionary conservation of some or all residues in a local neighborhood that may be signatures of functional sites. Subsequently, some form of pattern similarity is employed and optimized to best discriminate between functional and non-functional residues.

We distinguish three basic strategies according to which computational methods attempt to describe and identify functional sites, Fig. (**2**). For example, a functional site can be described by a set of distance constraints between pairs of residues (or atoms). We refer to such strategy as *template-based* and illustrate it in Fig. (**3A**). Similarly, a residue neighborhood may be characterized by various geometric, physicochemical or evolutionary properties (or features) that can be extracted from the local structure and then used by a machine learning algorithm. We refer to this approach as *residue microenvironment-based,* Fig. (**3B**). The third strategy is represented by approaches designed to identify local neighborhoods that belong to particular classes of higher-order shapes, such as pockets, clefts or surface patches. These methods are used to predict larger interface regions between a protein and its partners, e.g. ligands, other proteins, or nucleic acids. We refer to these methods as *residue macroenvironment-based methods,* Fig. (**3C**). One separate class of methods, based on its distinct problem formulations and algorithmic solutions, is represented by the *graph-theoretic* approaches. Such algorithms start by transforming protein structures to graphs and then employ combinatorial algorithms with statistical inference to find representative subgraphs or score similarities between neighborhoods of residues, Fig. (**3D**). Finally, many methods involve *structural post-processing,* where spatial proximity is imposed on a set of the potential functional residues that were predicted using either the above-mentioned structure-based or some sequence-based methods. Usually structural post-processing is used for filtering isolated predictions that are considered to be false positives. We extensively discuss all five approaches in the following subsections.

We note that these strategies are not mutually exclusive, as illustrated in Fig. (**2**), and in fact, a number of proposed algorithms combine two or more of them. Rather, the classification is provided to emphasize major differences among methods and to facilitate discussion of strengths and limitations of each methodology.

### 2.1. Template-based Methods

Template-based methods define and construct local structural motifs or patterns that characterize functional sites. Most of the early methodology can be traced back to the fields of chemistry and computer science in which various algorithms were proposed to find interesting patterns from a

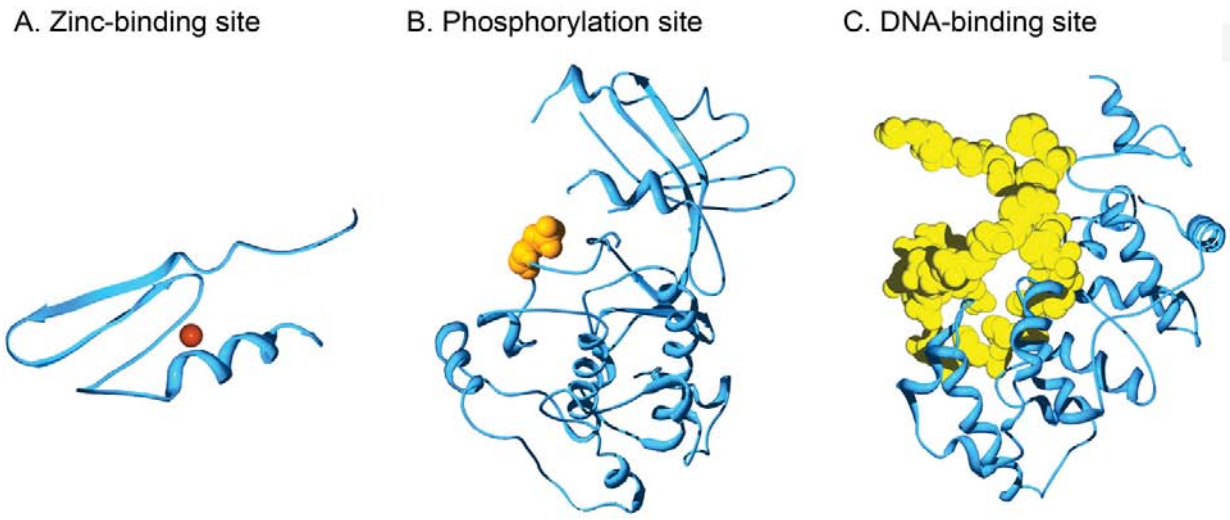


**Fig. (1).** Examples of problems related to the prediction of functional sites. **A**: zinc-binding in 1ncs with the Zn ion shown in red; **B**: phosphorylation site S474 in 2bva with the phosphate group colored in orange; C: DNA-binding residues in 1orn with the DNA nucleotides shown in yellow. In panels A and C, functional residues are the residues involved in contact with the co-factor or the nucleic-acid, depending on a particular definition of the contact (usually it is a pre-specified distance cutoff). In panel B, the phosphorylation site is the residue being phosphorylated.

set of chemical 3D structures and search for those patterns in new structures [11-14]. The computational biology community has recognized the importance of local structural neighborhoods to the global function of a protein (e.g. catalytic triads in serine proteases), proposed more powerful algorithms, and associated approximate structural matches with quantitative scores such as P-values or posterior probabilities.

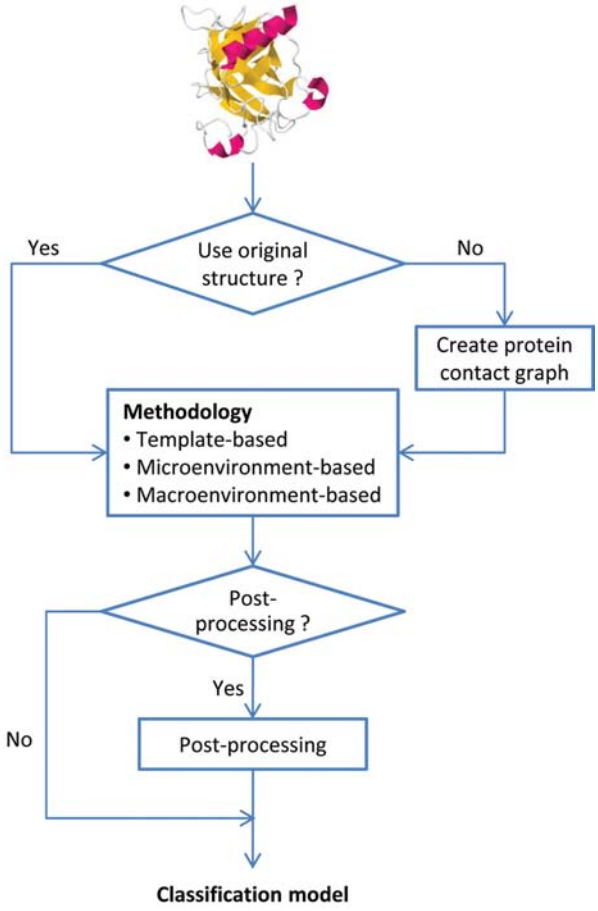


**Fig. (2).** Flowchart showing general methodologies for the prediction of functional residues from protein structures. Protein structures can be used directly or can be transformed to a graph representation. The subsequent core algorithm typically belongs to one of the three categories (template-based, microenvironment-based or macroenvironment-based) based on how the structural neighborhood is used and what types of geometric shapes the algorithm is trying to identify. Post-processing can be applied to further boost the performance.

One pioneering idea came from computer vision by the use of geometric hashing [15-16]. Geometric hashing can be used to detect structural motifs without assumptions on the substructure of the functional sites and has been shown to recognize active sites in enzymes. The TEmplate Search and Superposition (TESS) method is also based on geometric hashing: it defines templates by using a reference frame based on amino acid side chains and relative positons of atoms in the vicinity, limited by a user-specified cuttoff distance [17]. JESS is a faster and more flexible version of TESS and was applied to discover templates of enzyme active sites [18], ligand-binding sites, DNA-binding sites and also to predict global protein function [19].

A different group of template-based approaches define structural templates using sets of interresidue or interatomic distances over a set of functional residues (typically spatially close). Gregory *et al.* defined templates as three (Cα, Cβ) atoms and their interatomic distances. Then, they combined them with hydrophobicity and other requirements to search for new metal-binding sites [20]. Similar methods were proposed by Wallace *et al.* [21] and Russell [22] where root mean square distance (rmsd) constraints were imposed on the pattern comprised of the side-chain atoms of the functional residues. Another template-based approach used to identify active sites is the Fuzzy Functional Forms (FFFs) [23]. FFF is a set of geometric descriptors (distances) between key functional residues with a degree of tolerance obtained by analyzing a number of structurally and functionally similar proteins. An improved version of the algorithm augments the standard FFF results with a scoring function based on evolutionary information [24].

More recent methods have further generalized the notion of a template by considering physicochemical properties of a local structural neighborhood. For example, Shulman-Peleg *et al.* define templates via triangles of physicochemical properties and use hierarchical scoring to find functional sites [25]. Similarly, Innis *et al.* identify regions in protein structures with high density of functional groups such as hydroxyl and carboxylate group from conserved residues, because such regions were deemed functionally relevant [26].

PHUNCTIONER represents a distinct class within template-based methods [27]. It starts with a structural alignment within a superfamily and then finds evolutionarily conserved residues corresponding to the subset of proteins annotated with the same Gene Ontology (GO) terms [28], but excludes residues conserved in the entire family. Position-specific scoring matrices are built for those GO-conserved residues, such that the functional residues can be predicted in a new protein after it is structurally aligned to the training proteins. This tool was primarily constructed to predict GO terms from structures of highly divergent proteins, but it can also be seen as a predictor of functional residues. Its application is limited to proteins with global structural similarity.

Most template-based methods are coupled with an algorithm for the identification of templates from a set of protein structures (although this step may be manual or semi-automated, e.g., [23]), an algorithm to assess the statistical significance of a match, a database of known templates, and a web tool that can identify patterns in a database of structures or search a structure against a database of patterns. For example, SuMo uses triangles of chemical groups to represent protein structures, instead of backbone information, and provides a web service to search for ligand binding sites [29-30]. PROCAT provides a database of derived 3D templates of enzyme active sites using the TESS algorithm [17, 21]. SPASM aims to find matches of a given motif in a database of structures, while RIGOR can search a protein structure against a database of structural motifs [31]. PDBSiteScan scans a query structure for functional sites using all annotated sites in PDB as templates [32]. Each template is composed of three atoms (N, Cα, C) from the functional residue

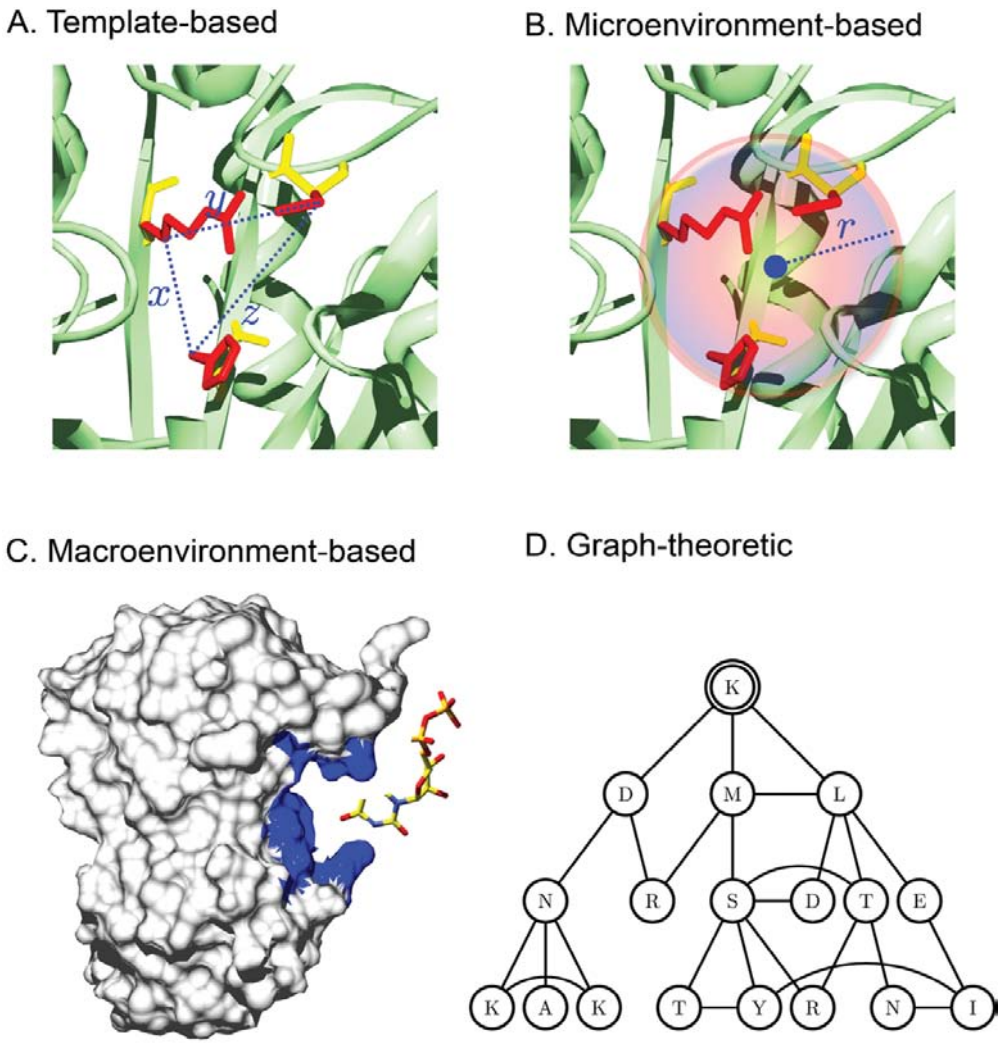


**Fig. (3).** Illustration of the (**A**) template-, (**B**) microenvironment-, (**C**) macroenvironment- and (**D**) graph-theoretic approach in describing and identifying functional residues/sites. **A**: a catalytic triad described by a set of 3 distances (*x*, *y*, *z*); **B**: a residue microenvironment described by radius *r*. Residue microenvironment may be centered on a particular residue, but also elsewhere in space; **C**: ligand-binding residues highlighted in blue (protein-protein and protein-nucleic acid binding sites will have relatively flat interfaces); **D**: a graph representation of a part of protein structure, where nodes represent amino acids and edges suggest that distances between two amino acids are below some threshold value. Double-circled node is the functional residue of interest.

itself as well as other neighboring residues within 5Å. It can identify active sites, post-translationally modified sites, binding sites, etc. GASPS uses a genetic algorithm strategy to create 3D templates consisting of 3-10 conserved residues in a protein family to identify family members from the background [33]. Finally, PINTS provides a web service to search a given pattern against a database of structures, and given a structure, provides a search against a pattern database. The pattern database includes ligand-binding sites, SITE annotations in PDB files, surface residues and conserved residues. The output matches are assigned P-values to indicate the statistical significance that differentiates true functional matches from the background [34-35].

In summary, template-based methods provide a natural way of identifying functional residues and show promising performance in modeling active sites and metal-binding sites (see summary of methods in Table **1**). However, various aspects related to assessing the significance of a match remain unsolved; for example, it is unclear how to build the background distribution [36]. Template-based methods usually do not exploit the power of machine learning.

### 2.2. Residue Microenvironment-based Methods

The main signature of residue microenvironment-based methods is the focus on a single residue or position in the structure and its surrounding environment. Usually, a set of structural, physicochemical and evolutionary properties are collected and encoded into a fixed-length vector. Sets of functional (positive) and non-functional (negative) residues are then incorporated into supervised machine learning approaches.

Although not the first, FEATURE is the flagship of the residue microenvironment-based methods [37-42]. First proposed in 1995 by Bagley *et al.* [37], FEATURE models the neighborhoods of the sites of interest using concentric

**Table 1. List of the Representative Methods in Each Category and a Brief Description of the Methodology. T-Template-based, μE-Microenvironment-based, ME-Macroenvironment-based**

| Category | Method | Method Description |
|---|---|---|
| T | FFFs [23]; Gregory *et al.* [20]; Russell *et al*. [22]; Wallace *et al.* [21]; SiteEngine [25]; Innis *et al.* [26]; PHUNCTIONER [27]; SuMo [29]; GASPS [33] ; | Template derivation and application |
| | TESS [17]; JESS [18]; SPASM and RIGOR[31]; | Template searching algorithms |
| | SuMo [30]; PROCAT [17, 21]; PDBSiteScan [32]; PINTS [34]; ProFunc [19, 94]; | Web servers |
| | ASSAM [110]; DRESPAT [103]; Huan *et al.* [106-108] | Graph-based, subgraph-isomorphism; Frequent subgraph mining |
| μE | FEATURE [111]; POOL [58] | Machine learning, Naïve Bayes |
| | Zvelebil and Sternberg [46]; Ota *et al.* [47] | Human expert knowledge base |
| | Stawiski *et al.* [112]; Ahmad *et al.* [50-51]; Gutteridge *et al.* [60] | Machine learning, Neural network |
| | Panchenko *et al.* [48]; THEMATICS [55] Wei *et al.* [56]; | Scoring function |
| | Bhardwaj *et al.* [49]; Kuznetsov *et al.* [52]; Bhardwaj *et al*. [53]; THEMATICS-SVM [57]; Youn *et al.* [59]; Petrova and Wu [62] | Machine learning, SVMs |
| | Structure kernel [66]; Graphlet kernel [65]; | Machine learning, SVMs with custom kernels |
| | Tang *et al.* [54] | Machine learning, genetic algorithm integrated neural network |
| | ResBoost [63]; | Machine learning, decision tree and boosting |
| | DISCERN [64]; | Machine learning, logistic regression |
| | GG [105]; | Graph-based, clique detection |
| | Amitai *et al*. [104]; | Graph-based, scoring function with network features |
| ME | SURFNET [69]; PocketPicker [113]; Xie and Bourne [70]; Q-SiteFinder [71]; SCREEN [74]; | Pocket detection |
| | LIGSITEcsc [75]; SURFNET-ConSurf [76]; C*onCavity* [77]; SitePredict [78]; | Pocket detection, evolutionary conservation and physicochemical properties |

spheres within which it enumerates various properties, including atom/residue type, atom/residue physicochemical properties, chemical groups and secondary structure information. A set of non-functional sites as controls are also collected from the structures and a naive Bayes classifier is constructed. FEATURE was applied to predict calcium binding sites, disulfide bond-forming sites, enzyme active sites, ATP-binding sites, zinc binding sites etc. WebFEATURE [40] provides a web service for those models and seqFeature [42] creates functional site libraries from PROSITE patterns. To allow quick database search for similar environments, S-BLEST [43], a variant of FEATURE, uses Manhattan distance while the approach by Yoon *et al.* [44] uses a weighted Hamming distance. For a more detailed overview of the FEATURE framework and its applications, we refer readers to [45].

As early as in 1988, Zvelebil and Sternberg used spherical neighborhoods to extract features such as secondary structure information, B-factors, residue separation, relative solvent accessibility, and electrostatic interactions to characterize catalytic residues [46]. More recently, Ota *et al*. used a rule-based approach with conservation, destabilizing potential and solvent accessibility information for catalytic residue prediction [47]. Panchenko *et al.* predicted functional sites by first looking for conserved residues in a multiple sequence alignment, and then scored each conserved residue using the averaged conservation score of residues in its structural neighborhood. This method also takes into account residue solvent accessibility [48]. Bhardwaj *et al*. predicted DNA-binding sites using support vector machines (SVMs) with features such as solvent accessibility, local residue composition, net charge and electrostatic potentials [49]. Other methods for predicting DNA-binding residues also use various sequence and structure properties [50-53]. The most frequently addressed problem, however, is that of catalytic residue prediction, where a number of residue microenvironment-based methods have been developed [47, 54-64].

Kernel-based methods have also been proposed to identify functional sites [65-66]. Such methods need not define the feature space explicitly; rather, a similarity function is defined over all pairs of data points with the kernel property providing a guarantee that objects (here, structural neighborhoods) can be mapped to some high-dimensional, but generally unknown, feature space [67]. Xin *et al.* introduced a structure kernel method that uses oriented spherical microenvironments divided into cells of non-uniform volume [66]. A product kernel was then defined to incorporate geometric, chemical and evolutionary determinants in the similarity (kernel) function between two structural neighborhoods. The structure kernel was used to predict catalytic residues and to identify situations when mutations lead to human disease via the loss or gain of catalytic activity [66].

Residue microenvironment-based strategies provide a very promising avenue for identifying functional residues, predominantly because of the general machine learning framework and straightforward applicability to different problems. Such methods can automatically identify features relevant for the task, but may result in classification that is not easily described by a small set of human-interpretable rules.

### 2.3. Residue Macroenvironment-based Methods

Most methods discussed in this paper focus on the prediction of enzyme active sites, co-factor binding sites, or post-translational modification sites, where a relatively compact local structural region is involved. However, a large group of algorithms and tools have been developed to identify particular classes of larger structural neighborhoods, e.g. surface patches, pockets, cavities or clefts, which provide interfaces to ligands or macromolecular partners. These methods are highly valuable because protein-protein interactions lie at the center of almost every cellular process and protein-DNA binding is essential for genetic activities. Similarly, accurate identification of ligand-binding sites is valuable in the context of structure-based drug design. Residue macroenvironment-based methods have been reviewed recently, thus we provide only a brief summary and refer authors to relevant publications where appropriate.

Early methods for ligand binding site prediction focused on detecting structural pockets because previous studies had shown that binding sites are usually found in largest pockets [68]. Such methods may be based on surface geometry, specifically the shape and size of the pocket (SURFNET [69] and the approach by Xie and Bourne [70]), or based on the interaction energy between a probe molecule and the protein (e.g. Q-Site Finder uses a methyl group as probe [71]). Those methods are summarized in a recent review [72]. Similar protein surfaces suggest similar binding activity, which is an idea behind the method by Binkowski *et al.* [73]. Since most proteins have multiple pockets and predominantly one ligand-binding site [74], pocket detection algorithms alone cannot distinguish true sites from false positives. Evolutionary conservation [75-77] and physicochemical properties [78] are typically included to improve the prediction accuracy. Methods based on global structure similarity to transfer ligand-binding sites have also shown high accuracy [79-80].

All of the above-mentioned methods predict ligand-binding sites in a protein without assuming its binding partner. Docking and scoring functions can be used to calculate the binding affinity and direction for a specific ligand but the knowledge of the ligand structure is required. Typically, pocket detection is used as the first step to reduce the search space and to speed up the process compared to blind docking [81]. Flexible docking that considers protein dynamics and conformational changes upon binding improves prediction performance. However, these methods are computationally expensive [82]. A recent study provides extensive performance comparisons of the docking algorithms and scoring functions [83].

Although shape complementarity is also necessary for protein-protein binding [84], unlike ligand binding sites, protein-protein binding sites have a relatively flat surface [85]. This makes geometry based methods involving pocket detection for ligand binding less effective at predicting protein-protein binding. However, methods such as docking and homology-/threading-based binding site transfer can still be applied (for details regarding those methods, we refer readers to [86]). Features extracted from 3D structures, such as solvent accessible surface area, B-factors and electrostatic potentials, together with various sequence and physicochemical properties, can be fed into machine learning approaches for the prediction of protein interaction residues. For more detailed reviews, we refer readers to [87-88]. We note that there exist several categories of protein-protein interaction interfaces [89-90], thus treating them differently could result in better prediction performance. For example, while it is commonly believed that protein-protein interface residues are more conserved than other surface residues to evolutionarily preserve the interaction, it has been shown that antibody-binding residues (B-cell epitopes) are significantly less conserved than other surface residues [91]. Also, currently all known antibodies interacting with antigens use similar structural regions, so antigen binding sites can be predicted using structural alignments of antibodies [92].

Protein-nucleic acid binding site/residue prediction models use similar strategies as protein-protein interaction models, but can further incorporate electrostatic potentials since positively charged surface patches suggest possible nucleic acid-binding sites [93]. Helix-turn-helix motif detection as well as DNA binding site templates can also be used to predict protein-DNA binding residues [94].

### 2.4. Structural Post-Processing

Conservation during evolution and spatial proximity are usually considered common properties of functional residues; thus a number of methods were developed to take advantage of these properties. Structural post-processing refers to those methods that use spatial clustering to group functional residues into functional sites and remove isolated residues that are likely to be false positive predictions. Evolutionary Trace (ET), first introduced by Lichtarge *et al.* in 1996 [95], is one of these methods. It starts with clustering proteins into different sequence identity groups. The conserved residues within each group are then retained and mapped to protein structures. Several ET variants were pro-

posed subsequently: a weighted ET that identifies the variability of a position in a multiple sequence alignment [96]; ConSurf that incorporates the physicochemical properties of the replaced amino acids in a multiple sequence alignment [97]; and 3D cluster analysis that calculates a conservation score for each residue based on the conservation of residues in its structural neighborhood [98]. Madabushi *et al.* account for gaps in the multiple sequence alignment and also report the number of clusters identified and the size of the largest cluster to automate the output of the ET results [99], while Yao *et al.* quantitatively assess the significance of trace clusters compared to functional sites [100]. Aloy *et al.* also map conserved residues to structures, and then use spatial clustering to define functional sites [101]. Finally, Gutteridge *et al.* use a similar strategy to filter catalytic residue prediction results and improve the prediction accuracy [60].

### 2.5. Graph-Theoretic Approaches

Based on the types of structural patterns they search for, graph-theoretic approaches can be used in any of the three main methodological groups (template, residue microenvironment, residue macroenvironment). However, these approaches represent a special category based on the distinct problem formulations and algorithmic approaches. Instead of using atomic coordinates directly, graph-theoretic methods start with transforming protein structures into graphs and then exploit various motif finders and graph similarity measures, combined with machine learning, to discover functional sites. Representative graph similarity measures involve subgraph enumeration, subgraph isomorphism, or identification of frequent subgraphs, although other measures, e.g. random walk-based scoring, can be applied as well.

Different methods often use different node and edge representations. Nodes may represent atoms or groups of atoms (residues), while edges typically reflect distances between atoms. Here too, the earliest applications to molecular 3D structures came from chemistry and computer science. For example, in an early work Brint and Willett [14] proposed a graph-theoretic approach to identifying maximal common substructures in a set of molecules and later adapted the technique in order to find 3D motifs of amino acid side chains in proteins structures using subgraph isomorphisms [102]. Most of the early methods readily benefitted from extensive work in graph theory.

More recent research includes DRESPAT where patterns are defined as complete subgraphs with three to six nodes [103]. Amitai *et al.* also created a protein structure graph in which the edges included backbone peptide bonds as well as the side chain non-covalent bonds. Their analysis confirmed earlier work that active sites have high network centrality and low relative solvent accessibility [104]. Deng *et al.* used a graph method to predict calcium binding sites where oxygen atoms in a protein structure represent nodes and edges indicate that the two oxygen atoms are within a predefined distance. Oxygen clusters, which refer to those cliques with more than 4 nodes, were selected. Distance between nodes and their geometric center is calculated and only those cliques within a distance range are considered calcium binding sites [105].

In a series of papers, Huan *et al.* used data mining concepts to extract spatial motifs in protein structure families [106-108]. The rationale was to identify frequent subgraphs in a protein family in an attempt to identify groups of residues responsible for the common function. Note that these methods relax the constraints of the maximum common subgraph in a set of graphs by using the ideas of frequent itemset mining [109].

Finally, the graphlet kernel method first creates a protein structure graph based on the Cα atoms and their distances [65]. It then enumerates oriented labeled graphlets of different sizes with a pivot point representing the residue of interest. The similarity between two vertices of interest was defined as an inner product between their graphlet counts. This technique does not result in identification of a template *per se*; rather it enumerates all templates in a structural neighborhood and uses them to define a similarity (kernel) function. The graphlet kernel method has been applied to the problem of catalytic residue prediction and identification of phosphorylation sites [65].

Graph-theoretic methods usually lead to the most principled approaches, based on the formalisms and elegance of graph algorithms. However, they suffer from the inability to model actual residue positions (instead, edges are constructed when two residues are closer than a predefined distance) and the spatial orientation of structural neighborhoods.

## 3. PERFORMANCE COMPARISONS BETWEEN MACHINE LEARNING METHODS

Machine learning principles are frequently used in predicting functional residues from protein structures. While different approaches have been introduced and evaluated on different datasets, most are not limited to a specific problem. Even when applied to the same problem, different data sets and evaluation strategies are often used to assess the performance, thus making comparisons across different studies hard to interpret and generally inconclusive. In order to gain insight into relative strengths and weaknesses of different methods, here we aim to assess their performance by training them on multiple data sets and evaluating them using the same protocol. This strategy is reasonably unbiased, but poses a limitation on the number of methods that can be tested because it requires availability of the training algorithms (or source code) instead of the pre-trained predictors. Thus, most web servers which only predict a particular type of functional residues and methods without software releases could not be evaluated.

We compare four published methods according to their ability to model functional residues: FEATURE [45], our in-house implementation of the method by Gutteridge *et al.* [60] referred to as GBT (based on the initials of the authors' last names), as well as the graphlet kernel [65], and structure kernel [66] methods. Structural post-processing has not been used for the GBT method, because it is equally applicable to the other three approaches.

The four problems were selected such that each method has already been evaluated on one or more data sets in the original publication. We used the default parameters for all methods, but emphasize that in practice an additional parameter optimization process could result in a different performance.

### 3.1. Data Sets

All methods were applied to problems involving (1) zinc-binding residues, (2) DNA-binding residues, (3) catalytic residues and (4) phosphorylation sites. The Zinc-binding data set was constructed based on the work by Ebert *et al.* [110] who considered a residue zinc-binding if it had N, O, or S atoms (referred to as the coordinating atoms) within 3Å of the zinc ion. The DNA-binding data was constructed by Yan *et al.* [111] and then mapped to protein structures using the PDB atom-seqres correspondence maps in the Astral compendium [112]. The catalytic residue data set was constructed from the Catalytic Site Atlas v.2.2.12 [113], but only proteins with literature-supported catalytic residues were included. Finally, the phosphorylation data set was constructed by Vacic *et al.* [65]. The original negative data sets were constructed to include all the non-positive residues in the protein chains that contain at least one positive (functional) residue. Redundant proteins were removed using the Astral40 database, except for the phosphorylation data in which Astral40 filtering removes too much data. Instead, Astral95 filtering was used for the phosphorylation data set. For each problem, a balanced data set was created by including all positive (functional) and an equal-sized random sample of the negative (non-functional) instances. The data sets are summarized in Table **2** and provided in Supplementary Materials.

### 3.2. Evaluation Protocol

The methods were evaluated using the per-chain 10-fold cross-validation on exactly the same training and test data. The set of chains was initially split into ten non-overlapping partitions. In each of the steps of cross-validation, all residues from the $1/10^{th}$ of the test chains were used for testing, whereas the residues from the remaining chains were used to construct a classifier. This procedure was adopted to best emulate a realistic scenario in which a completely new chain is presented to the classifier. The performance was evaluated by estimating the area under the ROC curves (AUC). The ROC curve is a plot of the true positive rate (sensitivity) as a function of the false positive rate (1 – specificity). A perfect classifier will have *AUC* = 1 and a classifier that provides outputs uniformly randomly will have *AUC* = 0.5.

### 3.3. Performance Comparisons

AUC estimates for each method on each data set are shown in Table **3** and Fig. (**4**). The results indicate, as expected, that each method indeed can be successfully applied to various problems even if it was not originally designed for it. However, the performance accuracies among them varied. Structure kernel achieved the highest accuracy on three of the four data sets: zinc-binding residues, catalytic residues, and phosphorylation sites. FEATURE, on the other hand, was the best-performing model on the DNA-binding sites. The success of the structure kernel might be due to the oriented neighborhoods and explicit encoding of the evolutionary conservation in the similarity function. Its good performance suggests that, in principle, it is possible to create a unified classifier for most types of functional sites. Although the GBT method also uses evolutionary conservation, it collects features that are specifically important for catalytic residues (via an earlier feature analysis by Bartlett *et al.*[114]). Thus, it resulted in a generally weaker performance on the other types of problems. The graphlet kernel was originally designed to predict function on specifically selected amino acids (e.g., phosphorylation on serine); however, here we applied it on a combination of all potential functional residues which resulted in a decreased performance.

Interestingly, despite not accounting for evolutionary conservation, FEATURE was the best classifier on the DNA-binding residues (note that unlike the other three approaches, FEATURE incorporates counts of atoms and chemical groups), implying that undirected residue neighborhoods and relatively simpler patterns are sufficient to model DNA-binding. This is not surprising since protein-DNA interac-

**Table 2. The Size of Four Data Sets Used in Comparisons. Each Data Set Contains the Same Number of Positive and Negative Examples**

| | Zn-Binding | DNA-Binding | Catalytic | Phosphorylation |
|---|---|---|---|---|
| No. protein chains | 445 | 102 | 314 | 679 |
| No. positive examples | 1,420 | 2,921 | 988 | 1,157 |

**Table 3. The Performance of Four Methods on Each of the Four Data Sets, Measured by the Area Under the ROC Curve (AUC). The Highest Performance for Each Data Set is Highlighted in Bold**

| | | Zn-Binding | DNA-Binding | Catalytic | Phosphorylation |
|---|---|---|---|---|---|
| AUC | FEATURE | 0.767 | **0.824** | 0.754 | 0.620 |
| | GBT | 0.713 | 0.713 | 0.814 | 0.559 |
| | Graphlet kernel | 0.758 | 0.691 | 0.732 | 0.688 |
| | Structure kernel | **0.808** | 0.800 | **0.839** | **0.711** |

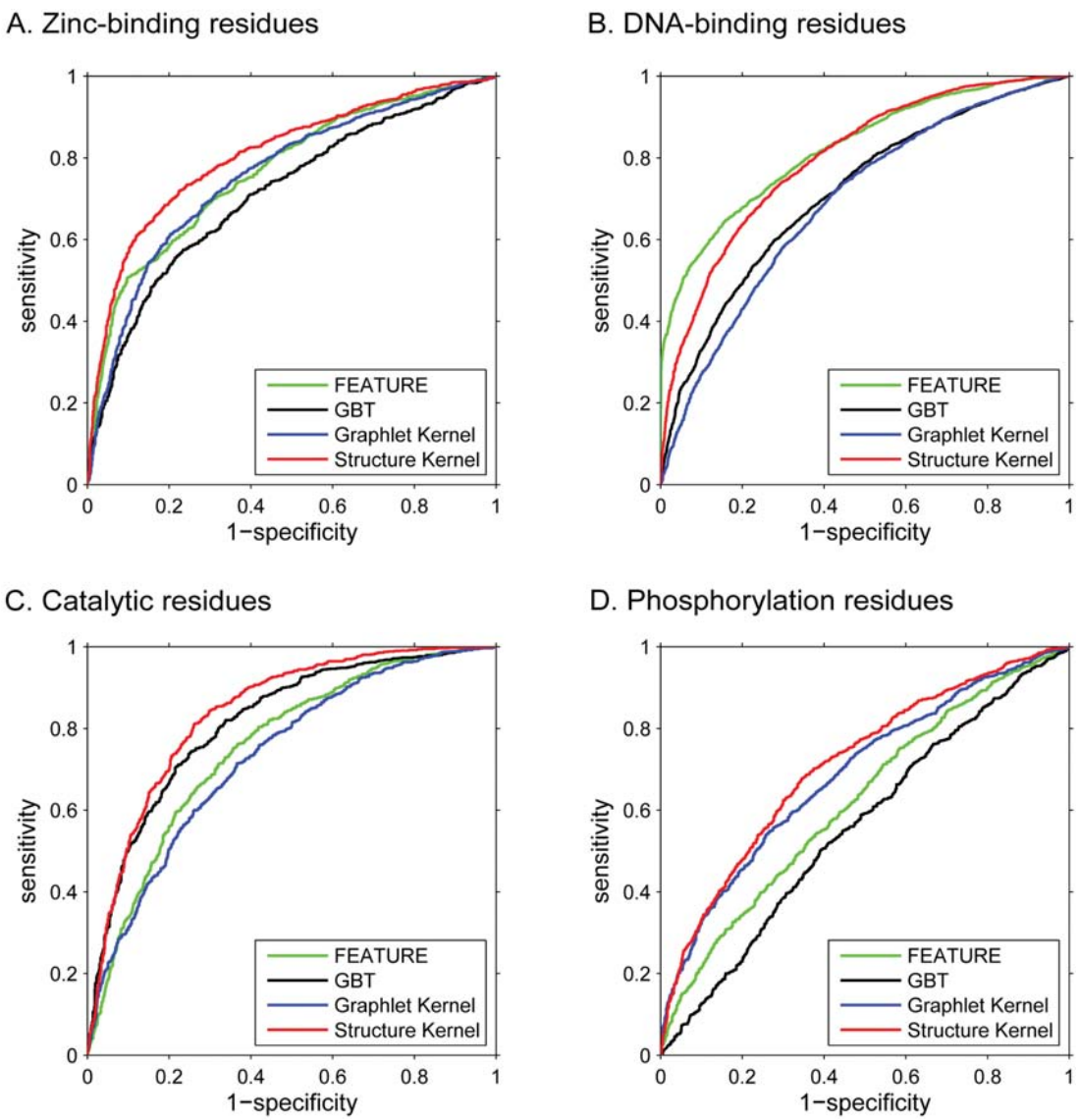


**Fig. (4).** ROC curves of the four methods on each of the data sets.

tions are mostly based on electrostatic attractions between the negatively charged DNA and positively charged protein surface patches [93]. It is noteworthy that all four methods were less accurate on phosphorylation sites, although the phosphorylation data set retained sequences with up to 95% sequence identity. A possible reason for such performance might be that the phosphorylation sites are frequently located in the loop regions, which have much higher degrees of motion compared to the rest of protein structure. Thus, the identification of phosphorylation sites may require use of larger local neighborhoods which were not captured by the methods with their respective default settings.

## 4. FUTURE CHALLENGES

As reviewed in this article, a plethora of methods have been proposed in the field of functional residue prediction from protein structure. We classified such methods into template-based, residue microenvironment- and residue-macroenvironment-based, depending on the types of functional sites they aim to identify. While great strides have been achieved, the protein function prediction and, in particular, prediction of functional residues, remains challenging for several reasons.

First, protein structures are static models reflecting a number of experimental artifacts such as crystal contacts, concentration-driven oligomerization, as well as the conditions of the experiment. It has been previously argued that high-resolution structural models are needed for accurate functional assignments [115]. Similarly, temperature, pH, and salt concentration have been shown to affect protein structure or lack thereof [116]. Even under the same experimental conditions, proteins show intrinsic hierarchical dynamics at different time scales [117]. The presence of ligands, other proteins, nucleic-acids, or post-translational modifications in experimentally determined structures may also result in structures different from that of an unmodified monomeric molecule and subsequently introduce noise in the training data.

Computational methods for predicting protein structure show great promise in modeling apo molecules [118], hence it is to be expected that a combination of structure and function prediction will become more prominent in the future. In addition, methods that introduce structural dynamics have started to emerge. Glazer *et al.* used molecular dynamics to model conformational flexibility of a protein structure in order to predict functional sites [119-120]. Liu *et al.* used *de novo* loop modeling to predict the structure and dynamics of the loop region and then applied FEATURE to predict calcium-binding sites [121]. At this time, these methods are slow and, because of the small number of examples tested, not well characterized with respect to the classification performance. Nevertheless, they represent the most promising recent trend in the prediction of functional sites.

Second, an inherent problem in determining protein function from structure is the existence of functionally relevant disordered regions [122-123]. Such regions cannot be characterized by the time-invariant atomic coordinates and thus are missing from structural models in PDB. Disordered regions, however, may be preferred in signaling, especially when post-translational modifications such as phosphorylation [124-126], methylation [127], and ubiquitination [128] are involved. Alternatively, some regions may be structured in PDB but require local unfolding prior to the post-translational modification. Structural features from these sites lead to biased models and disrupt the overall accuracy.

Third, numerically quantifying the significance of a matched pattern is non-trivial. In template-based approaches, a matched pattern may be associated with a P-value that the pattern could be found by chance in a structural database of a given size. However, determining P-values is problematic, (1) because the empirical null distribution may not be sufficient to accurately calculate low P-values, (2) because of the stringency of the multiple hypothesis testing correction, or (3) because of the inability to accurately model the null distribution analytically [129]. Noble has discussed approaches to assigning accurate P-values and suggested use of the false discovery rate (fdr) to determine the significance of a match [129]. In supervised approaches, the challenge is in incorporating the proper class priors in order to predict the posterior probability that the residue of interest is functional. One example of this is the prediction of catalytic residues where in one of the most influential studies the ratio of positive vs. negative sites in training was selected to be 1 : 6, effectively setting the class prior of the catalytic residue to 1/7 = 0.143 [60]. These training set class priors have been followed by other studies [54, 59, 66], although in part to enable fair comparisons among methods. While this enabled the model to learn the concept, such predictions typically result in a high fdr (also referred to as over-prediction) because the actual class prior in the structures of enzymes is about 0.009 (ratio 1 : 114). Even worse, if the predictor is also applied to non-enzymes (~70% of all functionally characterized proteins), even a model with 95% accuracy may not be practically useful. Methods for estimating class priors have been proposed [130-132] as well as the algorithms that partition inhomogeneous data sets and separately estimate class priors on each partition [133]. Prediction of functional residues is frequently equivalent to the prediction of rare events, resulting in class-imbalanced data sets. Various training methods have been introduced to address class-imbalance [134], in part because the classification costs are unknown for the cost-sensitive learning scenarios. A more general problem, learning from biased data, has been addressed in statistical and machine learning communities as well [135-137], but so far we are not aware of applications and impact on protein function prediction.

The fourth major challenge stems from the very definition of protein function as well as the biological and environmental context in which experimental assays designed to determine function were carried out. For example, in characterizing catalytic residues, there are discrepancies in their definitions. The Catalytic Site Atlas generally classifies residues as catalytic if they are involved in the chemistry of catalysis, suggesting that residues involved in substrate binding, residues supporting the geometry of active sites, or residues involved in co-factor binding are not, despite being necessary for the catalysis. This definition, however, has not been consistently followed in the literature, potentially leading to confusion in not only defining catalytic residues, but also in applicability of the methods designed to predict them. In addition, function determination by experimentalists is tied to a particular organism, tissue or a particular set of environmental conditions [138]. A residue may be phosphorylatable, but because a particular kinase is not expressed in the organism or tissue, this phosphorylation event may not occur. Similarly, a phosphorylation event may not be detectable because of the low abundance of a post-translationally modified form of the protein that may be outside of the dynamic range of the instrument used to determine post-translational modifications (affinity enrichment strategies are available only for some post-translational modifications, such as phosphorylation). A positive prediction on such residues will lead to overestimation of the false positive rate. It is well-known that biological databases contain errors in deposited protein functions [139-140]. While similar error estimates have not been provided for functional residues, it is clear that numerous problems exist.

In summary, despite the achieved success and decades of active research, the accurate prediction of functional residues from protein structures is still an open problem. We anticipate that the area will further grow, most likely in the direction of incorporating protein dynamics with statistical and machine learning approaches and towards biomedical applications. When combined with more standardized function definitions and accurate database annotations, more sophisticated and powerful algorithms can and will emerge. The future of the field is bright.

## ACKNOWLEDGEMENTS

We thank Wyatt Clark and Daniel Schrider for proofreading the article. This work was supported by the National Science Foundation grant DBI-0644017 to PR.

## SUPPLEMENTARY MATERIAL

Two files per each data set used in Section 3 are provided. The .pos and .neg files list functional and non-functional residues in protein structures, respectively.

Supplementary material is available on the publishers Web site along with the published article.

## REFERENCES


[1] Rost, B.; Liu, J.; Nair, R.; Wrzeszczynski, K.O.; Ofran, Y. Automatic prediction of protein function. *Cell Mol. Life Sci.*, **2003**, *60* (12), 2637-2650.

[2] Lee, D.; Redfern, O.; Orengo, C. Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.*, **2007**, *8* (12), 995-1005.

[3] Laskowski, R.A.; Thornton, J.M. Understanding the molecular machinery of genetics through 3D structures. *Nat. Rev. Genet.*, **2008**, *9* (2), 141-151.

[4] Wang, Z.; Moult, J. SNPs, protein structure, and disease. *Hum. Mutat.*, **2001**, *17* (4), 263-270.

[5] Baker, D. An exciting but challenging road ahead for computational enzyme design. *Protein Sci.*, **2010**, *19* (10), 1817-1819.

[6] Tramontano, A. The ten most wanted solutions in protein bioinformatics. CRC Press: **2005**.

[7] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.*, **2000**, *28* (1), 235-242.

[8] Burley, S. An overview of structural genomics. *Nat. Struct. Biol.*, **2000**, *Struc. Genomic Supplement*, 932-934.

[9] Dessailly, B.H.; Nair, R.; Jaroszewski, L.; Fajardo, J.E.; Kouranov, A.; Lee, D.; Fiser, A.; Godzik, A.; Rost, B.; Orengo, C. PSI-2: structural genomics to cover protein domain family space. *Structure*, **2009**, *17* (6), 869-881.

[10] Kryshtafovych, A.; Venclovas, C.; Fidelis, K.; Moult, J. Progress over the first decade of CASP experiments. *Proteins*, **2005**, *61 Suppl 7*, 225-236.

[11] Cone, M.; Venkataraghven, R.; McLafferty, F. Molecular structure comparison program for the identification of maximal common substructures. *J. Am. Chem. Soc.*, **1977**, *99* (23), 7668-7671.

[12] Lesk, A. Detection of three-dimensional patterns of atoms in chemical structures. *Comm. ACM*, **1979**, *22* (4), 224.

[13] Crandell, C.; Smith, D. Computer-assisted examination of compounds for common three-dimensional substructures. *J. Chem. Inform. Comp. Sci.*, **1983**, *23* (4), 186-197.

[14] Brint, A.; Willett, P. Algorithms for the identification of three-dimensional maximal common substructures. *J. Chem. Inform. Comp. Sci.*, **1987**, *27* (4), 152-158.

[15] Nussinov, R.; Wolfson, H.J. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Nat. Acad. Sci. USA*, **1991**, *88* (23), 10495-10499.

[16] Fischer, D.; Wolfson, H.; Lin, S.L.; Nussinov, R. 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 Sci.*, **1994**, *3* (5), 769-778.

[17] Wallace, A.C.; Borkakoti, N.; Thornton, J.M. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.*, **1997**, *6* (11), 2308-2323.

[18] Barker, J.A.; Thornton, J.M. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics*, **2003**, *19* (13), 1644-1649.

[19] Laskowski, R.A.; Watson, J.D.; Thornton, J.M. Protein function prediction using local 3D templates. *J. Mol. Biol.*, **2005**, *351* (3), 614-626.

[20] Gregory, D.S.; Martin, A.C.; Cheetham, J.C.; Rees, A.R. The prediction and characterization of metal binding sites in proteins. *Protein Eng.*, **1993**, *6* (1), 29-35.

[21] Wallace, A.C.; Laskowski, R.A.; Thornton, J.M. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.*, **1996**, *5* (6), 1001-1013.

[22] Russell, R. B. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.*, **1998**, *279* (5), 1211-1227.

[23] Fetrow, J.S.; Skolnick, J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.*, **1998**, *281* (5), 949-968.

[24] Cammer, S.A.; Hoffman, B.T.; Speir, J.A.; Canady, M.A.; Nelson, M.R.; Knutson, S.; Gallina, M.; Baxter, S. M.; Fetrow, J.S. Structure-based active site profiles for genome analysis and functional family subclassification. *J. Mol. Biol.*, **2003**, *334* (3), 387-401.

[25] Shulman-Peleg, A.; Nussinov, R.; Wolfson, H.J. Recognition of functional sites in protein structures. *J. Mol. Biol.*, **2004**, *339* (3), 607-633.

[26] Innis, C.A.; Anand, A.P.; Sowdhamini, R. Prediction of functional sites in proteins using conserved functional group analysis. *J. Mol. Biol.*, **2004**, *337* (4), 1053-1068.

[27] Pazos, F.; Sternberg, M. J.E. Automated prediction of protein function and detection of functional sites from structure. *Proc. Nat. Acad. Sci. USA*, **2004**, *101* (41), 14754-14759.

[28] Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; Harris, M.A.; Hill, D.P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J.C.; Richardson, J.E.; Ringwald, M.; Rubin, G.M.; Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.*, **2000**, *25* (1), 25-29.

[29] Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins*, **2003**, *52* (2), 137-145.

[30] Jambon, M.; Andrieu, O.; Combet, C.; Deleage, G.; Delfaud, F.; Geourjon, C. The SuMo server: 3D search for protein functional sites. *Bioinformatics*, **2005**, *21* (20), 3929-3930.

[31] Kleywegt, G.J. Recognition of spatial motifs in protein structures. *J. Mol. Biol.*, **1999**, *285* (4), 1887-1897.

[32] Ivanisenko, V.A.; Pintus, S.S.; Grigorovich, D.A.; Kolchanov, N.A. PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucleic Acids Res.*, **2004**, *32* (Web Server issue), W549-W554.

[33] Polacco, B.J.; Babbitt, P.C. Automated discovery of 3D motifs for protein function annotation. *Bioinformatics*, **2006**, *22* (6), 723-730.

[34] Stark, A.; Russell, R.B. Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res.*, **2003**, *31* (13), 3341-3344.

[35] Stark, A.; Sunyaev, S.; Russell, R.B. A model for statistical significance of local similarities in structure. *J . Mol. Biol.*, **2003**, *326* (5), 1307-1316.

[36] Gherardini, P.F.; Helmer-Citterich, M. Structure-based function prediction: approaches and applications. *Brief Funct. Genomic Proteomic*, **2008**, *7* (4), 291-302.

[37] Bagley, S.C.; Altman, R.B. Characterizing the microenvironment surrounding protein sites. *Protein Sci.*, **1995**, *4* (4), 622-635.

[38] Bagley, S.C.; Wei, L.; Cheng, C.; Altman, R.B. Characterizing oriented protein structural sites using biochemical properties. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **1995**, *3*, 12-20.

[39] Wei, L.; Altman, R.B. Recognizing protein binding sites using statistical descriptions of their 3D environments. *Pac. Symp. Biocomput.*, **1998**, 497-508.

[40] Liang, M.P.; Banatao, D.R.; Klein, T.E.; Brutlag, D.L.; Altman, R.B. WebFEATURE: An interactive web tool for identifying and visualizing functional sites on macromolecular structures. *Nucleic Acids Res.*, **2003**, *31* (13), 3324-3327.

[41] Wei, L.; Altman, R.B. Recognizing complex, asymmetric functional sites in protein structures using a Bayesian scoring function. *J. Bioinform. Comput. Biol.*, **2003**, *1* (1), 119-138.

[42] Wu, S.; Liang, M.P.; Altman, R.B. The SeqFEATURE library of 3D functional site models: comparison to existing methods and applications to protein function annotation. *Genome Biol.*, **2008**, *9* (1), R8.

[43] Mooney, S.D.; Liang, M. H.-P.; DeConde, R.; Altman, R.B. Structural characterization of proteins using residue environments. *Proteins*, **2005**, *61* (4), 741-747.

[44] Yoon, S.; Ebert, J.C.; Chung, E.-Y.; De Micheli, G.; Altman, R.B. Clustering protein environments for function prediction: finding PROSITE motifs in 3D. *BMC Bioinform.*, **2007**, *8 Suppl 4*, S10.

[45] Halperin, I.; Glazer, D.S.; Wu, S.; Altman, R.B. The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. *BMC Genomics*, **2008**, *9 Suppl 2*, S2.

[46] Zvelebil, M.J.; Sternberg, M.J. Analysis and prediction of the location of catalytic residues in enzymes. *Protein Eng.*, **1988**, *2* (2), 127-138.

[47] Ota, M.; Kinoshita, K.; Nishikawa, K. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol.*, **2003**, *327* (5), 1053-1064.

[48] Panchenko, A.R.; Kondrashov, F.; Bryant, S. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.*, **2004**, *13* (4), 884-892.

[49] Bhardwaj, N.; Langlois, R.; Zhao, G.; Lu, H. Structure Based Prediction of Binding Residues on DNA-binding Proteins. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, **2005**, *3*, 2611-2614.

[50] Ahmad, S.; Gromiha, M.M.; Sarai, A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* **2004**, *20* (4), 477-486.

[51] Ahmad, S.; Sarai, A. PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinform.*, **2005**, *6*, 33.

[52] Kuznetsov, I.B.; Gou, Z.; Li, R.; Hwang, S. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins*, **2006**, *64* (1), 19-27.

[53] Bhardwaj, N.; Lu, H. Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions. *FEBS Lett.*, **2007**, *581* (5), 1058-1066.

[54] Tang, Y.-R.; Sheng, Z.-Y.; Chen, Y.-Z.; Zhang, Z. An improved prediction of catalytic residues in enzyme structures. *Protein Eng. Des. Sel.*, **2008**, *21* (5), 295-302.

[55] Ondrechen, M.J.; Clifton, J.G.; Ringe, D. THEMATICS: a simple computational predictor of enzyme function from structure. *Proc. Nat. Acad. Sci. USA*, **2001**, *98* (22), 12473-12478.

[56] Wei, Y.; Ko, J.; Murga, L.F.; Ondrechen, M.J. Selective prediction of interaction sites in protein structures with THEMATICS. *BMC Bioinform.*, **2007**, *8*, 119.

[57] Tong, W.; Williams, R.J.; Wei, Y.; Murga, L.F.; Ko, J.; Ondrechen, M.J. Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines. *Protein Sci.*, **2008**, *17* (2), 333-341.

[58] Tong, W.; Wei, Y.; Murga, L.F.; Ondrechen, M.J.; Williams, R.J. Partial order optimum likelihood (POOL): maximum likelihood prediction of protein active site residues using 3D Structure and sequence properties. *PLoS Comput. Biol.*, **2009**, *5* (1), e1000266.

[59] Youn, E.; Peters, B.; Radivojac, P.; Mooney, S.D. Evaluation of features for catalytic residue prediction in novel folds. *Protein Sci.*, **2007**, *16* (2), 216-226.

[60] Gutteridge, A.; Bartlett, G.J.; Thornton, J.M. Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.*, **2003**, *330* (4), 719-734.

[61] Torrance, J.W.; Bartlett, G.J.; Porter, C.T.; Thornton, J.M. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *J. Mol. Biol.*, **2005**, *347* (3), 565-581.

[62] Petrova, N.V.; Wu, C.H. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinform.*, **2006**, *7*, 312.

[63] Alterovitz, R.; Arvey, A.; Sankararaman, S.; Dallett, C.; Freund, Y.; Sjolander, K. ResBoost: characterizing and predicting catalytic residues in enzymes. *BMC Bioinform.*, **2009**, *10*, 197.

[64] Sankararaman, S.; Sha, F.; Kirsch, J.F.; Jordan, M.I.; Sjolander, K. Active site prediction using evolutionary and structural information. *Bioinformatics*, **2010**, *26* (5), 617-624.

[65] Vacic, V.; Iakoucheva, L.M.; Lonardi, S.; Radivojac, P. Graphlet kernels for prediction of functional residues in protein structures. *J. Comput. Biol.*, **2010**, *17* (1), 55-72.

[66] Xin, F.; Myers, S.; Li, Y.F.; Cooper, D.N.; Mooney, S.D.; Radivojac, P. Structure-based kernels for the prediction of catalytic residues and their involvement in human inherited disease. *Bioinformatics*, **2010**, *26* (16), 1975-1982.

[67] Ben-Hur, A.; Ong, C.S.; Sonnenburg, S.; Scholkopf, B.; Ratsch, G. Support vector machines and kernels for computational biology. *PLoS Comput. Biol.*, **2008**, *4* (10), e1000173.

[68] Laskowski, R.A.; Luscombe, N.M.; Swindells, M.B.; Thornton, J.M. Protein clefts in molecular recognition and function. *Protein Sci.*, **1996**, *5* (12), 2438-2452.

[69] Laskowski, R.A. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **1995**, *13* (5), 323-30, 307-308.

[70] Xie, L.; Bourne, P.E. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinform.*, **2007**, *8 Suppl 4*, S9.

[71] Laurie, A.T.R.; Jackson, R.M. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinform.*, **2005**, *21* (9), 1908-1916.

[72] Laurie, A.T.R.; Jackson, R.M. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Curr. Protein Pept. Sci.*, **2006**, *7* (5), 395-406.

[73] Binkowski, T.A.; Joachimiak, A.; Liang, J. Protein surface analysis for function annotation in high-throughput structural genomics pipeline. *Protein Sci.*, **2005**, *14* (12), 2972-2981.

[74] Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins* **2006**, *63* (4), 892-906.

[75] Huang, B.; Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.*, **2006**, *6*, 19.

[76] Glaser, F.; Morris, R.J.; Najmanovich, R.J.; Laskowski, R.A.; Thornton, J. M. A method for localizing ligand binding pockets in protein structures. *Proteins*, **2006**, *62* (2), 479-488.

[77] Capra, J.A.; Laskowski, R.A.; Thornton, J.M.; Singh, M.; Funkhouser, T.A. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, **2009**, *5* (12), e1000585.

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

[79] Brylinski, M.; Skolnick, J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Nat. Acad. Sci. USA*, **2008**, *105* (1), 129-134.

[80] Wass, M.N.; Sternberg, M.J.E. Prediction of ligand binding sites using homologous structures and conservation at CASP8. *Proteins*, **2009**, *77 Suppl 9*, 147-151.

[81] Hetenyi, C.; van der Spoel, D. Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci.*, **2002**, *11* (7), 1729-1737.

[82] Ritchie, D.W. Recent progress and future directions in protein-protein docking. *Curr. Protein Pept. Sci.*, **2008**, *9* (1), 1-15.

[83] Warren, G.L.; Andrews, C.W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M.H.; Lindvall, M.; Nevins, N.; Semus, S.F.; Senger, S.; Tedesco, G.; Wall, I.D.; Woolven, J.M.; Peishoff, C.E.; Head, M.S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.*, **2006**, *49* (20), 5912-5931.

[84] Lawrence, M.C.; Colman, P.M. Shape complementarity at protein/protein interfaces. *J. Mol. Biol.*,**1993**, *234* (4), 946-950.

[85] Jones, S.; Thornton, J.M. Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.*, **1997**, *272* (1), 121-132.

[86] Szilagyi, A.; Grimm, V.; Arakaki, A.K.; Skolnick, J. Prediction of physical protein-protein interactions. *Phys. Biol.*,**2005**, *2* (2), S1-S16.

[87] Zhou, H.-X.; Qin, S. Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, **2007**, *23* (17), 2203-2209.

[88] Ezkurdia, I.; Bartoli, L.; Fariselli, P.; Casadio, R.; Valencia, A.; Tress, M.L. Progress and challenges in predicting protein-protein interaction sites. *Brief Bioinform.*, **2009**, *10* (3), 233-246.

[89] Ofran, Y.; Rost, B. Analysing six types of protein-protein interfaces. *J. Mol. Biol.*, **2003**, *325* (2), 377-387.

[90] Vacic, V.; Oldfield, C.J.; Mohan, A.; Radivojac, P.; Cortese, M.S.; Uversky, V.N.; Dunker, A.K. Characterization of molecular recognition features, MoRFs, and their binding partners. *J. Proteome. Res.*, **2007**, *6* (6), 2351-2366.

[91] Ponomarenko, J.V.; Bourne, P.E. Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC Struct. Biol.*, **2007**, *7*, 64.

[92] Ofran, Y.; Schlessinger, A.; Rost, B. Automated identification of complementarity determining regions (CDRs) reveals peculiar characteristics of CDRs and B cell epitopes. *J. Immunol.*, **2008**, *181* (9), 6230-6235.

[93] Stawiski, E.W.; Gregoret, L.M.; Mandel-Gutfreund, Y. Annotating nucleic acid-binding function based on protein structure. *J. Mol. Biol.*,**2003**, *326* (4), 1065-1079.

[94] Laskowski, R.A.; Watson, J.D.; Thornton, J.M. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, **2005**, *33* (Web Server issue), W89-W93.

[95] Lichtarge, O.; Bourne, H.R.; Cohen, F.E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*,**1996**, *257*, 342-358.

[96] Landgraf, R.; Fischer, D.; Eisenberg, D. Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng.*, **1999**, *12* (11), 943-951.

[97] Armon, A.; Graur, D.; Ben-Tal, N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.*, **2001**, *307* (1), 447-463.

[98] Landgraf, R.; Xenarios, I.; Eisenberg, D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, **2001**, *307* (5), 1487-1502.

[99] Madabushi, S.; Yao, H.; Marsh, M.; Kristensen, D.M.; Philippi, A.; Sowa, M.E.; Lichtarge, O. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.*, **2002**, *316* (1), 139-154.

[100] Yao, H.; Kristensen, D.M.; Mihalek, I.; Sowa, M. E.; Shaw, C.; Kimmel, M.; Kavraki, L.; Lichtarge, O. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.*, **2003**, *326* (1), 255-261.

[101] Aloy, P.; Querol, E.; Aviles, F.X.; Sternberg, M.J. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.*, **2001**, *311* (2), 395-408.

[102] Artymiuk, P.J.; Poirrette, A.R.; Grindley, H.M.; Rice, D.W.; Willett, P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.*, **1994**, *243* (2), 327-344.

[103] Wangikar, P.P.; Tendulkar, A.V.; Ramya, S.; Mali, D.N.; Sarawagi, S. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J. Mol. Biol.*, **2003**, *326* (3), 955-978.

[104] Amitai, G.; Shemesh, A.; Sitbon, E.; Shklar, M.; Netanely, D.; Venger, I.; Pietrokovski, S. Network analysis of protein structures identifies functional residues.*J. Mol. Biol.*, **2004**, *344* (4), 1135-1146.

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

[106] Huan, J.; Wang, W.; Washington, A.; Prins, J.; Shah, R.; Tropsha, A. Accurate classification of protein structural families using coherent subgraph analysis. *Pac. Symp. Biocomput.*, **2004,** 411-422.

[107] Huan, J.; Bandyopadhyay, D.; Wang, W.; Snoeyink, J.; Prins, J.; Tropsha, A. Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. *J. Comput. Biol.*, **2005**, *12* (6), 657-671.

[108] Bandyopadhyay, D.; Huan, J.; Liu, J.; Prins, J.; Snoeyink, J.; Wang, W.; Tropsha, A. Structure-based function inference using protein family-specific fingerprints. *Prot. Sci.*, **2006**, *15* (6), 1537-1543.

[109] Agrawal, R.; Imielinski, T.; Swami, A. *Mining associations between sets of items in massive databases*, Proc. of the ACM-SIGMOD 1993 Int'l. Conf. on Mgmt. of Data, Washington, D.C., Washington, D.C., **1993**, 207-216.

[110] Ebert, J.C.; Altman, R.B. Robust recognition of zinc binding sites in proteins. *Prot. Sci.*, **2008**, *17* (1), 54-65.

[111] Yan, C.; Terribilini, M.; Wu, F.; Jernigan, R.L.; Dobbs, D.; Honavar, V. Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinform.*, **2006**, *7*, 262.

[112] Brenner, S.E.; Koehl, P.; Levitt, M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **2000**, *28* (1), 254-256.

[113] Porter, C.T.; Bartlett, G.J.; Thornton, J.M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **2004**, *32* (Database issue), D129-D133.

[114] Bartlett, G.J.; Porter, C.T.; Borkakoti, N.; Thornton, J.M. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **2002**, *324* (1), 105-121.

[115] Skolnick, J.; Fetrow, J.S.; Kolinski, A. Structural genomics and its importance for gene function analysis. *Nat. Biotechnol.*, **2000**, *18* (3), 283-287.

[116] Mohan, A.; Uversky, V.N.; Radivojac, P. Influence of sequence changes and environment on intrinsically disordered proteins. *PLoS Comput. Biol.*, **2009**, *5* (9), e1000497.

[117] Henzler-Wildman, K.; Kern, D. Dynamic personalities of proteins. *Nature*, **2007**, *450* (7172), 964-72.

[118] Baker, D. A surprising simplicity to protein folding. *Nature*, **2000**, *405* (6782), 39-42.

[119] Glazer, D.S.; Radmer, R.J.; Altman, R.B. Improving structure-based function prediction using molecular dynamics. *Structure*, **2009**, *17* (7), 919-929.

[120] Glazer, D.S.; Radmer, R.J.; Altman, R.B. Combining molecular dynamics and machine learning to improve protein function recognition. *Pac. Symp. Biocomput.*, **2008,** 332-343.

[121] Liu, T.; Altman, R.B. Prediction of calcium-binding sites by combining loop-modeling with machine learning. *BMC Struct. Biol.*, **2009**, *9*, 72.

[122] Dunker, A.K.; Brown, C.J.; Lawson, J.D.; Iakoucheva, L.M.; Obradovic, Z. Intrinsic disorder and protein function. *Biochemistry*, **2002**, *41* (21), 6573-6582.

[123] Radivojac, P.; Iakoucheva, L.M.; Oldfield, C.J.; Obradovic, Z.; Uversky, V.N.; Dunker, A.K. Intrinsic disorder and functional proteomics. *Biophys. J.*, **2007**, *92* (5), 1439-1456.

[124] Iakoucheva, L.M.; Radivojac, P.; Brown, C.J.; O'Connor, T.R.; Sikes, J.G.; Obradovic, Z.; Dunker, A.K. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, **2004**, *32* (3), 1037-1049.

[125] Collins, M.O.; Yu, L.; Campuzano, I.; Grant, S.G.N.; Choudhary, J.S. Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. *Mol. Cell. Proteo.s*, **2008**, *7* (7), 1331-1348.

[126] Gsponer, J.; Futschik, M.E.; Teichmann, S.A.; Babu, M.M. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science*, **2008**, *322* (5906), 1365-1368.

[127] Daily, K.M.; Radivojac, P.; Dunker, A.K. *Intrinsic disorder and protein modifications: building an SVM predictor for methylation*, IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), San Diego, California, U.S.A., November **2005**; San Diego, California, U.S.A., **2005**, 475-481.

[128] Radivojac, P.; Vacic, V.; Haynes, C.; Cocklin, R.R.; Mohan, A.; Heyen, J.W.; Goebl, M.G.; Iakoucheva, L. M. Identification, analysis, and prediction of protein ubiquitination sites. *Proteins*, **2010**, *78* (2), 365-680.

[129] Noble, W.S. How does multiple testing correction work? *Nat. Biotechnol.*, **2009**, *27* (12), 1135-1137.

[130] Latinne, P.; Saerens, M.; Decaestecker, C. *Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: evidence from a multi-class problem in remote sensing, Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, Williamstown, MA, USA, June 28 - July 1; Brodley, C. E.; Danyluk, A. P. Eds. Morgan Kaufmann: Williamstown, MA, USA, **2001**, 298-305.

[131] Vucetic, S.; Obradovic, Z. *Classification on data with biased class distribution*, Proceedings of the 12th European Conference on Machine Learning (ECML 2001), Freiburg, Germany, September 5-7; Freiburg, Germany, **2001**, 527-538.

[132] Saerens, M.; Latinne, P.; Decaestecker, C. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural. Comput.*, **2002**, *14* (1), 21-41.

[133] Radivojac, P.; Chawla, N.V.; Dunker, A.K.; Obradovic, Z. Classification and knowledge discovery in protein databases. *J. Biomed. Inform.*, **2004**, *37* (4), 224-239.

[134] Chawla, N.; Cieslak, D.; Hall, L.; Joshi, A. Automatically countering imbalance and its empirical relationship to cost. *Data Mining Knowl. Dis.*, **2008**, *17* (2), 225-252.

[135] Heckman, J. Sample selection bias as a specification error. *Econometrica: J. Econo. Soc.*, **1979**, *47* (1), 153-161.

[136] Zadrozny, B. *Learning and evaluating classifiers under sample selection bias*, Twenty-First International Conference on Machine Learning (ICML 2004), Banff, Alberta, Canada July 4-8; Banff, Alberta, Canada **2004**.

[137] Cortes, C.; Mohri, M.; Riley, M.; Rostamizadeh, A. *Sample Selection Bias Correction Theory*, Springer-Verlag: **2008**, 53.

[138] Thornton, J.M.; Todd, A.E.; Milburn, D.; Borkakoti, N.; Orengo, C.A. From structure to function: approaches and limitations. *Nat. Struct. Biol.*, **2000**, *7 Suppl*, 991-994.

[139] Brenner, S.E. Errors in genome annotation. *Trends Genet.*, **1999**, *15* (4), 132-133.

[140] Schnoes, A.M.; Brown, S.D.; Dodevski, I.; Babbitt, P.C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, **2009**, *5* (12), e1000605.

**Received: April 01, 2011** **Revised: April 01, 2011** **Accepted: May 04, 2011**