Amino Acids (2008) 35:627–650
DOI 10.1007/s00726-008-0088-8

REVIEW ARTICLE

Bridging protein local structures and protein functions

Zhi-Ping Liu · Ling-Yun Wu · Yong Wang · Xiang-Sun Zhang · Luonan Chen

Received: 21 February 2008 / Accepted: 10 March 2008 / Published online: 18 April 2008

Abstract One of the major goals of molecular and evolutionary biology is to understand the functions of proteins by extracting functional information from protein sequences, structures and interactions. In this review, we summarize the repertoire of methods currently being applied and report recent progress in the field of in silico annotation of protein function based on the accumulation of vast amounts of sequence and structure data. In particular, we emphasize the newly developed structure-based methods, which are able to identify locally structural motifs and reveal their relationship with protein functions. These methods include computational tools to identify the structural motifs and reveal the strong relationship between these pre-computed local structures and protein functions. We also discuss remaining problems and possible directions for this exciting and challenging area.

Keywords Functional genomics · Functional motifs · Local structures · Protein function prediction

Introduction

DNA sequences can be called 'the blueprint of life', while proteins represent the fulfillment of this blueprint in terms of structures and functions. A fundamental goal of functional genomics research is to understand how proteins carry out functions in a living cell (Eisenberg et al. 2000; Brenner 2001; Goldsmith-Fischman and Honig 2003). In addition to experimental methods, computational methods have been extensively applied with the aim of developing hypotheses in terms of assigning specific functions to specific proteins and providing valuable biological insights. The basic rationale behind such research is that the gene sequence determines the amino acid sequence, and the amino acid sequence determines the protein structure, which, in turn, determines the protein function (Whisstock and Lesk 2003). Many proteins, even among those in the Protein Data Bank (PDB), have not yet been annotated, although we have succeeded in deriving their structures (Laskowski et al. 2003; Watson et al. 2005). We review here the in silico annotation methods currently used to determine protein function from protein local structures.

Generally speaking, proteins are the main catalysts, structure components, signal transfers and molecular machines in a biological organism. As such, they are the basic elements of functions. However, the definition of function means different things to different people since it is an evolving concept associated to an abundance of interpretations. In general, these functions can be described at many levels, ranging from the biochemical functions at the molecular level (e.g. catalytic or binding activities) to biological processes at the level of biomolecular cooperation (e.g. signal transduction or cellular physiological process) to the cellular components at the cell level of an organ (e.g. nucleus or rough endoplasmic) (Devos and Valencia 2000; Watson et al. 2005). Several schemes/tools/databases have been developed in recent decades for measuring protein functions in a systematic model with the aim of annotating the functions of proteins (Watson et al. 2005); these include EC (Barrett 1997), MIPS (Ruepp et al. 2004), GO (The Gene Ontology Consortium 2000; Camon

Z.-P. Liu · L.-Y. Wu · Y. Wang · X.-S. Zhang
Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, 100080 Beijing, China

L. Chen (✉)
Institute of Systems Biology, Shanghai University,
200444 Shanghai, China
e-mail: lnchen@staff.shu.edu.cn

Table 1 The classification schemes to define functions of proteins

Method	URL	Description
EC	http://www.chem.qmul.ac.uk/iubmb/enzyme/	The functional catalogue for enzyme. It provides four hierarchical level classes. For example, EC 1.1.1.163 represents cyclopentanol dehydrogenase
MIPS	http://mips.gsf.de/projects/funcat	The functional categories for yeast. It can be extended to other organisms of life. For example, 01.01.06.06.01.01 represents diaminopimelic acid pathway
GO	http://www.geneontology.org/	The systematic classification of proteins. It is species-independent and contains three relatively independent ontologies. For example, GO:0051635 represents bacterial cell surface binding (F)
KEGG	http://www.genome.jp/kegg/	Linking genomes to biological systems and also to environments by the processes of interaction and reaction mapping

MIPS, Munich Information Center for Protein Sequences; EC, Enzyme Commission; KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology

et al. 2004) and KEGG (Kanehisa and Goto 2000), as shown in Table 1.

Using the existing function annotations as 'gold standard' data, researchers have been able to develop many protein function annotation methods in recent years based on protein relationships. We summarize the existing function annotation methods in the framework of Fig. 1, which shows the basic tendency for the functional inference methodology—i.e. to explore sequence similarity, structure similarity, protein interaction and their integration. We briefly review these in the following list:

- Using sequence information**.** The methods in this category often utilize a BLAST, FASTA or PSI-BLAST score to detect the sequence similarity and annotate the functions to a target protein from its homologous protein (Whisstock and Lesk 2003; Watson et al. 2005). In the safe zone (Rost 1999) of sequence similarity, the sequence-based methods can provide putative annotations with high confidence (Wilson et al. 2000). A number of papers have tested the global performance between the relationship of the sequence similarity and function similarity. Shah and Hunter (1997) tested the sequence similarity among enzymes in many EC classes at various thresholds and concluded that the functional similarity could not be detected perfectly when the sequences are not similar enough. Wilson et al. (2000) and Devos and Valencia (2000) obtained similar results. Joshi and Xu (2007) presented a systematic analysis on the sequence–function relationships in four model organisms.
- Using structure information. Protein structures are more conserved than protein sequences (Orengo et al. 1999; Hou et al. 2005). A number of methods have been developed with the aim of assessing protein structure similarity (Kolodny et al. 2005); these can be grouped as coordinate-based [such as STRUCTAL (Gerstein and Levitt 1998), SAMO (Chen et al. 2006), TM-align (Zhang and Skolnick 2005) and ProSup (Lackner et al. 2000)], distance-matrix-based [such as DALI (Holm and Sander 1993), CE (Shindyalov and Bourne 1998), FATCAT (Ye and Godzik 2004), SSAP (Orengo and Taylor 1996)] and secondary-structure-based [such as VAST (Gibrat et al. 1996), SSM (Krissinel and Henrick 2004), LOCK (Singh and Brutlag 1997) and FAST (Zhu and Weng 2005)]. Classifying the proteins into different classes or families based on global structure similarity will assist researchers in determining the relationships among different proteins and provide a foundation of functional organization (Brenner 2001). SCOP (Murzin et al. 1995), CATH (Orengo et al. 1997) and FSSP (Holm and Sander 1996) comprehensively cluster all proteins with known structures. Based on

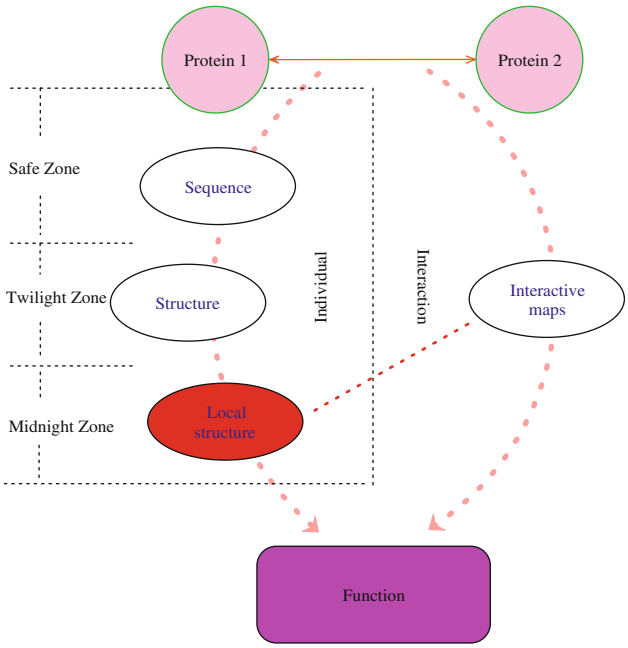

Fig. 1 Framework of existing function annotation methods. The *dotted line* links the individual methods with the interaction methods. The *safe zone* means that pairwise sequence identity is higher than 40%, the *twilight zone*, about 20–30%, the *midnight zone* below 20%

those clusters, the functional relationships among the proteins can be roughly detected.

- Using interactome information. Proteins always interact with other molecules to carry out their functions (Sharan et al. 2007). Information on protein–protein interactions or other interaction maps among molecules, such as DNA binding with protein, can be explored to annotate the protein functions from complexes and pathways of the biochemical processes. The network-based methods extend the functional inference from the single molecular level to a systematic level by considering interactions among genetic components and transferring functions among them (Vazquez et al. 2003; Barabasi and Oltvai 2004; Zhang et al. 2007). Sharan et al. (2007) cataloged the methods to direct methods and module-assisted methods individually.
- Using integrated information. Another sensible strategy is to use many different data sources to increase the chances of obtaining function annotations for any given protein. For example, in Marcotte et al. (1999), proteins are grouped by experimental data, such as metabolic function, phylogenetic profiles, Rosetta stone results and correlated messenger RNA expression patterns to determine the functional relationships among proteins of the yeast. In fact, many methods are in this framework (Sanishvili et al. 2003; George et al. 2005; Pal and Eisenberg 2005; Zhao et al. 2008a, b), especially when data integration becomes the focus of the systems biology study.

In this review we highlight the relationships between protein local structures and protein functions since it is commonly believed that local regions on the structures are responsible for the performance of the particular functional tasks (Russell 1998; Ferre et al. 2005). Well-known examples include the Ser–His–Asp triad in enzymes and other known special structural frameworks that carry out certain functions of catalysis (Torrance et al. 2005). It is now widely recognized that some fold similarities suggest an 'analogous' rather than a 'homologous' relationship (Russell 1998). Proteins can adopt similar tertiary folds while performing different functions at different binding site locations. Given the existing status that the midnight zone functional linkages escape from the sequence and global structure similarity, only the local structures can be used to analyze detailed relationships with functions by determining the protein–protein interaction, protein–DNA interaction or other global performance from the physical perspective. Also, the local structures of protein provide more detail information on protein function not only from the single targeted action of that protein, but also from the integrative process due to the detailed components and the three-dimensional architecture. The local structures are also important in the design of drugs and bioengineering. In an interesting paper, Schnell and Chou (2008) convincingly provided nuclear magnetic resonance (NMR) data showing that the M2 proton channel of influenza a virus is typically controlled by the local conformational change with a pH-gated mechanism. The discovery provides sound evidence that the local structures are crucial for determining protein function, and it is vitally important in the search for effective anti-influenza drugs (Borman 2008). Bridging protein local structures and protein functions can timely provide useful information for structure-based drug design [e.g. see the methods in Chou et al. (2003) and Wang et al. (2007a) against severe acute respiratory syndrome (SARS), and that in Du et al. (2007) against chicken influenza A virus H5N1, as well as a review paper (Chou 2004)]. Thus, it a key task of researchers in this field is to investigate the relationships between protein functions and protein local structures.

This review is organized into four parts. First, we will describe the main molecular functions related to protein local structures. This is followed by a description of existing definitions and methods for detecting similarities in local structures. In the third part, the detailed methodologies to bridge local structures with functions are reviewed. Some discussion and future directions are summarized in the last part.

Molecular functions related to local structures

To bridge the relationship between local structures and functions, we first catalog the molecular functions of proteins strongly related to local structures. The local structures are often regarded as the protein–protein interfaces, catalytic sites, ligand-binding sites, metal-binding sites, post-translational modification sites or other miscellaneous active sites. Table 2 lists some of the important functional categories (Chakrabarti and Lanczycki 2007).

Protein–protein interaction

A protein generally interacts with other proteins in performing and regulating many processes in a cell. The pace of discovery of protein–protein interactions has recently accelerated due to rapid advances in new technologies (Salwinski and Eisenberg 2003; Chou and Cai 2006). The basis of protein–protein interactions often lie in local planar patches on the protein surface. The factors that influence the formation of protein–protein complexes can be cataloged into four different types—i.e. homodimeric protein, heterodimeric proteins, enzyme–inhibitor complexes and antibody–protein complexes (Jones and Thornton 1996). From the structural perspective, structural

Table 2 The categories of protein functions close related to local structures

Function	Descriptor
Protein binding	The protein–protein interfaces where the physical interactions take place
Ligand binding	Including nucleotide binding (e.g. DNA and RNA binding), lipid binding (e.g. cholesterol, glycerol, ganglioside, etc.), ligand; and carbohydrate binding (e.g. glucose, fructose, lactose, maltose, disaccharides, trisaccharides, etc.)
Metal binding	Functions of binding metals, such as zinc, magnesium and calcium
Catalytic site	Functional regions performing the catalytic functions
Miscellaneous sites	Active sites involving particular functions

characterization of macromolecular assemblies usually poses a more difficult challenge than structure determination of individual proteins (Russell et al. 2004). Effective approaches for the prediction of protein–protein interactions at physical interaction levels are also strongly in demand (Wodak and Mendez 2004). Zhou and Qin (2007) reviewed the methods currently being applied for interface prediction. The characteristics between interface and non-interface portions of a protein surface, such as sequence conservation, proportions of amino acids, secondary structure, solvent accessibility and side-chain conformational entropy, are often used to distinguish the specificity of local structures relating to protein binding function.

Protein–nucleotide binding

In the transcription and translation process, proteins always bind to DNA and RNA to fulfill various functions. Protein–nucleotide binding is a fundamental function of proteins. Luscombe et al. (2000) classified the DNA-binding proteins into eight different structural/functional groups. The helix–turn–helix (HTH) motif is one of the most common structures used by proteins to bind DNA, while protein–RNA binding involves a number of different structure specificities. A comparison between protein–RNA and protein–DNA complexes revealed that while base and backbone contacts (both hydrogen bonding and van der Waals) are observed with equal frequency in protein–RNA complexes, backbone contacts are more dominant in protein–DNA complexes (Jones et al. 2001). The positively charged residue, arginine, and the single aromatic residues, phenylalanine and tyrosine, all play key roles in the sites for the RNA-binding function.

Protein–ligand binding

Ligand binding is a key aspect of protein functions. Proteins recognize their natural ligands for transportation, signal transduction or catalysis (Campbell et al. 2003). The cleft volumes in proteins have strong relationships with their molecular interactions and functions. The ligands are always bound in the largest clefts (Laskowski et al. 1996).

Protein–metal binding

Metal ions have a role in a variety of important functions, including protein folding, assembly, stability, conformational change and catalysis (Barondeau and Getzoff 2004). In order to leverage the wealth of native metalloprotein structures into a deep understanding of metal ion site specificity and activity, high-resolution analyses of metal site structures and metalloprotein design are increasingly being performed. One of the most ubiquitous zinc-binding motifs is the C2H2 zinc finger motif, which was first identified in transcription factors (Ebert and Altman 2008).

Active sites

Another broad concept for protein local structures is the active site. Active sites of a protein are comprehensively related to functionally important local regions of the protein. The special features of functional local structure are to provide deep insights into the relationship between structure and function. For example, the catalytic triads provide a target of structure for finding the catalytic function of the proteins.

Identifying protein local structures

To date, many different types of local structures have been defined or identified based on the geometry of the local regions, protein surface patterns, chemical groups or the electronic features. Local structure features are believed to be the factors related to concrete functions. At the sequence level, the local regions may be scattered on the primary sequence, forming special motifs. Alternatively, at the folding level, they form locally spatial shapes. We can simply catalog the types of methods used to identify the local structures as follows: methods to detect profiles of

sequences with special local shapes, and methods to detect the substructures with special features based on folding.

Sequence-based local structures

The primary sequence of a protein consists of (combinations of) 20 different amino acids, which fold and pack together to constitute a special three-dimensional structure. Sequence motifs are conserved segments in protein primary sequences. Multiple sequence alignment is often used to identify the common patterns in several protein sequences, especially in the homology family. More advanced sequence comparison algorithms can detect the profiles of the functional residues in the primary sequence. Of these algorithms, one of the most common methods is the Hidden Markov Model (HMM). There are a number of important sequence pattern databases, which are publicly available from the Internet (Table 3).

Structure-based local structure

Local three-dimensional structural patterns, such as the surface cavities of protein (e.g. the clefts and pockets) also have conserved structural features. Table 4 lists a number of methods currently used to identify local structure patterns. The procedure of recognition can be generally divided into two parts. The first is to construct the local structures. The geometric structure patterns and biochemical properties can be used to segment the protein architecture into small substructures. The second is to search the annotated sites from the literature and databases.

The analysis of the protein surface is an active area of research in terms of the study of local structures. To date, two aspects of protein surface patches have attracted the most attention. The first is based on the defined features, such as surface curvature, surface cavities, electrostatic potential and hydrophobicity. CASTp (Binkowski et al. 2003b) uses the weighted Delaunay triangulation and the alpha complex for shape measurements. The local regions are defined by computational geometry, which identifies and measures surface accessible pockets as well as interior inaccessible cavities for proteins and other molecules. Computational geometry also measures analytically the area and volume of each pocket and cavity, both in solvent accessible surface (SA, Richards' surface) and molecular surface (MS, Connolly's surface). CASTp provides an

Table 3 Database of identified local structures based on sequences information

Database	URL	Descriptor
PROSITE	http://us.expasy.org/prosite/	A database of protein families and domains
PRINTS	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/	A compendium of protein fingerprints
Pfam	http://www.sanger.ac.uk/Software/Pfam/	A database of common protein domains and families by HMM
ProDom	http://prodom.prabi.fr/prodom/current/html/home.php	A database of protein domain families
SMART	http://smart.embl-heidelberg.de/	Simple Modular Architecture Research Tool
SUPERFAMILY	http://supfam.org/SUPERFAMILY/index.html	A database of structural and functional protein annotations

Table 4 Methods and databases to identify local structures based on structure information

Method	URL	Descriptor
CASTp	http://sts.bioengr.uic.edu/castp/	A database for identifying pockets and voids of proteins
pvSOAR	http://pvsoar.bioengr.uic.edu/	A web server of detecting similar pockets from CASTp
SURFNET	http://www.biochem.ucl.ac.uk/~roman/surfnet/	An algorithm for generating protein surfaces
SURFACE	http://cmb.bio.uniroma2.it/surface/	A database of protein surface patches
eF-Site	http://ef-site.hgc.jp/	A database for molecular surfaces of proteins' functional sites
LigSite	Unavailable	A fast algorithm to identify ligand-binding site
CSA	http://www.ebi.ac.uk/thornton-srv/databases/CSA/	A database documenting enzyme catalytic residues
PINTS	http://www.russell.embl-heidelberg.de/pints/	Finding local similarities between protein structures
SiteBase	http://www.modelling.leeds.ac.uk/sb/	A database of known ligand-binding sites
PDBSiteScan	http://www.mgs.bionet.nsc.ru/mgs/gnw/pdbsitescan/	Performing the best superposition of sites from PDBSite
SPASM	http://xray.bmc.uu.se/usf/spasm.html	Comparing user-defined motifs against a structure database
RIGOR	http://xray.bmc.uu.se/usf/	Searching a motif database to find matches, (opposite of SPASM, hence the name)
SuMo	http://sumo-pbil.ibcp.fr	A graph-based algorithm for finding similarities in substructures

online resource for locating, delineating and measuring concave surface regions on the three-dimensional structures of proteins. These include pockets located on protein surfaces and voids buried in the interior of proteins. pvSOAR (Binkowski et al. 2004) provides an online resource to identify similar protein surface regions. Kinoshita and Nakamura (2003) provided a molecular surface database of proteins' functional sites, named the eF-site. The method displays the electrostatic potentials and hydrophobic properties of proteins together on the Connolly surfaces of the active sites for analysis of the molecular recognition mechanisms. The Connolly surfaces are made by using the Molecular Surface Package program, and the electrostatic potentials are calculated by solving Poisson–Boltzmann equations with the self-consistent boundary method.

The second aspect of protein surface patches is based on a predefined segmentation size of the surface. The method uses a segmentation procedure to divide the surface into small segmentations that correspond to certain physical modules of the surface. SURFNET (Laskowski 1995) generates molecular surfaces and gaps between surfaces from three-dimensional coordinates supplied in a PDB-format file. The gap regions can correspond to the voids between two or more molecules or to the internal cavities and surface grooves within a single molecule. The program visualizes molecular surfaces, cavities and intermolecular interactions by segmenting the surfaces. Based on the SURFNET algorithm, SURFACE (Ferre et al. 2004) identifies clefts and explores the cleft boundaries called the surface patch. A non-redundant set of protein chains is then used to build a database of protein surface patches. LIGSITE (Hendlich et al. 1997) is a program for the automatic and time-efficient detection of pockets on the surface of proteins that act as binding sites for small molecule ligands. Pockets are identified with a series of simple operations on a cubic grid.

The special features of catalytic sites or other types of functional sites are also detected as local structures. Some functional annotations of residues can be found in databases and the literature, and the location of these residues can be represented as potential structural motifs. Although it is difficult to define just precisely what is the active site in protein structures, there are a number of methods for identifying active sites or functionally important residues. Wallace et al. (1997) described a geometric hashing algorithm, called TESS, to derive three-dimensional coordinate templates for motifs. TESS has been used to create a database of enzyme active site templates called PROCAT (Wallace et al. 1997). PROCAT provides facilities for interrogating a database of three-dimensional enzyme active site templates. It has been superseded by the Catalytic Site Atlas (CSA). The CSA (Porter et al. 2004; Torrance et al. 2005) is a database documenting enzyme active sites and catalytic residues in enzymes with a three-dimensional structure. It contains the original annotated entries derived from the primary literature by hand and the homologous entries found by the PSI-BLAST alignment. A HETATM and all annotated SITEs in the PDB also provide patterns of protein local structures strongly related to protein functions. Stark and Russell (2003a) reported patterns in non-homologous tertiary structures (PINTS) that can be used to uncover the recurring three-dimensional side-chain patterns based on the algorithm in Stark et al. (2003c). SiteBase (Gold and Jackson 2006a) is a database of known ligand-binding sites within the PDB. The search for an annotated position in the PDB constructs the location information of the ligand-binding sites. A collection of known sites from mining the annotations in the PDB has been designated as the PDBSite (Ivanisenko et al. 2005), which collects amino acid content structure features calculated by spatial protein structures, and physicochemical properties of sites and their spatial surroundings. The PDBSiteScan (Ivanisenko et al. 2004) provides an automatic search of three-dimensional protein fragments similar in structure to known functional sites.

A comparison of local structures in the PDB also provides valuable information for constructing the structural motifs. Kleywegt (1999) presented two programs, spatial arrangement of side-chains and main-chains (SPASM) and RIGOR, for recognizing spatial motifs in protein structure. SPASM can be used to find matches in the structural database for any user-defined motif. The program also has a unique capability to carry out "fuzzy pattern matching" with relax requirements on the types of some or all of the matching residues. RIGOR, on the other hand, can compare a database of pre-defined motifs against a perhaps newly determined structure. RIGOR scans a single protein structure for the occurrence of the pre-defined motifs from a database. Zemla (2003) presented a method for finding three-dimensional similarities in protein structure. This algorithm is able to generate different local superpositions between pairs of structures and to detect similar fragments. It allows the clustering of similar fragments and the use of such clusters to identify sequence patterns that represent local structure motifs. SUMO (Jambon et al. 2003) can detect the common site, which corresponds to the catalytic triad.

Bridges between local structures and protein functions

The general procedure of bridging the local structures with functions lies in constructing a candidate pool of local structures, identifying important features of function-related local structures and validating their functional

importance. The existing methods can be grouped into two categories, i.e. unsupervised and supervised methods, as shown in Fig. 2.

The unsupervised methods directly mine those local structures with special features and then detect their functional implications. The supervised methods use known function-related structures as the templates and match these similar patterns by comparison. There are strong relationships between the two kinds of methods. Most of the proposed methods are based on physical and/or biochemical patterns of the protein, and some particular patterns of local structures are strongly related to functions. In the unsupervised methods, the patterns are derived directly from a group of local structures without known functions. Their functional importance and characteristics are identified by analyzing the conserved factors in the common features of local structures. The identified function-related local structures can then be used to enlarge the pool of functional templates, which in turn can be used to measure the potential functional importance of the new substructures. Figure 2a shows these relations. These functionally important local regions can be referred to as functional motifs. The functional motif is the particular local structure pattern with factors that are the determinations of performing particular functions. Note that the functional motif is very important for studying the relationship between structure and function in theory, and it is of practical importance to the protein design of drug targets and other bioengineering fields.

We can investigate the functional patterns of the local structures in multiple ways. More specifically, we group existing methods to bridge protein local structure and function into three categories based on the hierarchical perspective, as shown in Fig. 2b.

1. Element-based methods. These identify the local structures from sequence, structure and/or other important amino acid residues information. The methods detect the common or conservation patterns in these elements of proteins and bridge the gaps between the local structures and functions at the micro level. During the bridging process, if prior knowledge is used to identify the functional importance or guide the detection, the method belongs to the supervised category, otherwise it belongs to the unsupervised division.
2. Feature-based methods. These investigate the putative features between the local structures and functions. This category can be further divided into two subcategories—i.e. scoring methods and learning methods. The identified functional features of local structures provide templates of functional motifs. In the scoring methods, the features of local structures are scored by a defined function, and then the scores are used to decide whether the targets are functionally important. Thresholds are often then chosen to provide guidance for detecting the importance of target local structures. In the learning methods, some features are chosen and learned from the known function-related local structures. The learned features in the trained machines can be used as the classifier to decide whether the testing targets are strongly related to the function. These methods belong to the supervised division.
3. Network-based methods. These are based on graph theory and network topology. The methods can be divided into two subcategories. The first is at the individual level and the second is at the mapping level. At the individual level, the protein can be represented as an interactive graph of the residues, with linkages

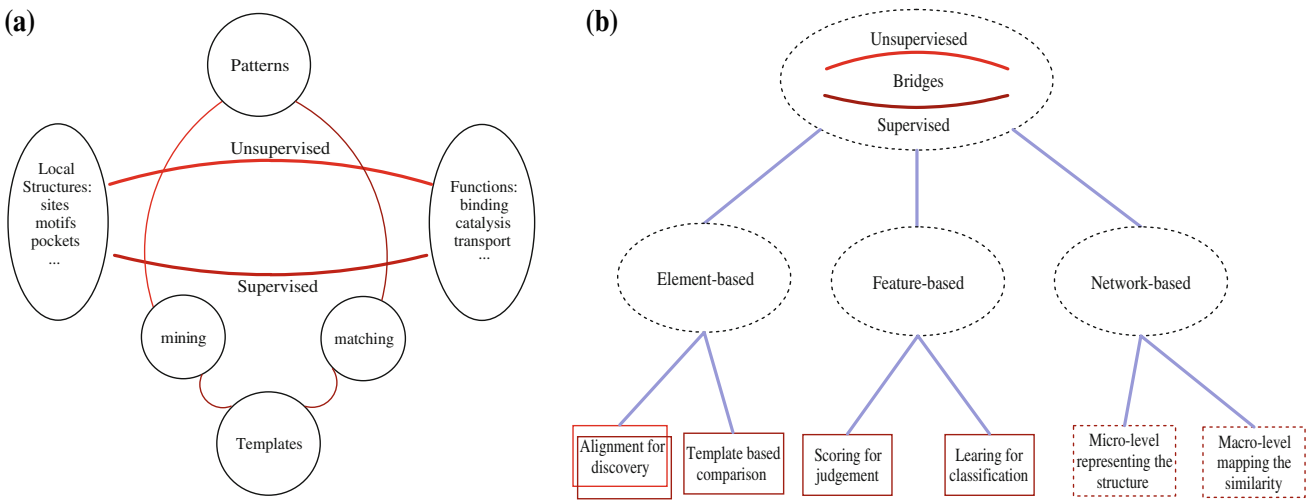

Fig. 2 Bridges between the local structures and the functions. **a** The schematic categories of the bridges, **b** the detailed and hierarchical classification of these bridges. In the lowest classes, the *bound color* implies the schematic category to which they belong

representing the close distance among them. Cliques of the graph, hub residues and residues with other special topology measures may correspond to functionally important regions and residues. At the mapping level, a network represents the similarity relations among the local structures. The functional motifs are mined from informative subgraphs. This approach lies in between the other two methods mentioned above and can be regarded as being semi-supervised because it uses some heuristic knowledge.

Element-based methods

Element-based methods are based on a basic intuition that the conserved part of a sequence and structure is an important functional motif (Aloy et al. 2001; Jones and Thornton 2004). The first step is a discovery process, which mines similar local structures from the sequences or structures of the target proteins. When similar local patterns of structures in some proteins are identified, the identified structure features of local regions will be the determinants of similar functions among the proteins. The second step is to match the process by comparing the target to the known functional templates. Based on the similarity between these, the function relationship is inferred. This method is also a basic tool for developing more advanced techniques to bridge the relationship between local structures and functions. The sequences, structures or other elements of the proteins are considered in the comparison. Table 5 lists the main methods that are currently being used. Depending on whether or not some prior knowledge is used in the assessment, the method is classified as being supervised or unsupervised.

Alignment method

Similar patterns of local structures can be identified in different proteins, even in proteins of the midnight zone with neither sequence homology nor structure homology. In this case, the alignment of the sequences and/or structure

Table 5 Element-based methods for identifying functional motifs

Local structure	Method	Software	Reference
Sequence motif			
Binding sites	Multiple sequence alignment	–	Ma et al. (2003)
Catalytic sites	Multiple sequence alignment	Conservation	Capra and Singh (2007)
Structural motif			
Functional active sites	Surface comparison	–	Rosen et al. (1998)
Recurring 3D motifs	Geometric hashing for structure alignment	–	Fischer et al. (1994)
Protein–protein interfaces	Comparison and querying	BID	Fischer et al. (2003)
Functional sites	All-vs-all comparison (from FSSP)	Phunctioner	Pazos and Sternberg (2004)
Constructed surface cavity	Pairwise alignment and querying	pvSOAR	Binkowski et al. (2003b)
Geometric and electrostatic surfaces	Pairwise alignment and querying	eF-site	Kinoshita and Nakamura (2003)
Surface chemical groups	Querying for similarity	SuMo	Jambon et al. (2003)
Binding pockets	Alignment all-vs-all and clustering	CavBase	Schmitt et al. (2002)
Binding sites and interface	Comparison for similarity	I2I-SiteEngine	Shulman-Peleg et al. (2005)
Documented motif			
Annotated sites	Alignment all-vs.-all and querying	PINTS	Stark and Russell (2003a)
Ligand-binding sites	Alignment all-vs.-all and querying	SiteBase	Gold and Jackson (2006a)
Known sites, especially interfaces	Querying for similarity	PDBSiteScan	Ivanisenko et al. (2004)
Sequence map to spatial motif			
Functional residues and sites	Multiple sequence alignment and phylogenetic	ET	Yao et al. (2003)
Functional residue clusters	Based on ET	–	Landgraf et al. (2001)
Patches of conserved residues	Based on ET	ConSurf	Armon et al. (2001)
Functional sites	Based on ET	–	Aloy et al. (2001)
Function template			
Functional 3D templates	Matching by geometric hashing	TESS	Wallace et al. (1997)
Metal-binding sites	Comparison with templates	PAR-3D	Goyal and Mande (2007)
Annotated functional sites	Comparison with templates	FIC	Chakrabarti and Lanczycki (2007)
Tertiary side-chain patterns	Subgraph-isomorphism matching	ASSAM	Artymiuk et al. (1994)

segments can imply similar functions of the local structures. These similar local structures of the proteins are important prognostic factors of their similar functions.

Multiple sequence alignment Ma et al. (2003) used ten protein interface families selected from two-chain interface entries in PDB, identified surface residues and filtered out contact residues. The alignment results of the residue properties revealed that polar residue hot spots occur frequently at the interfaces of macromolecular complexes, thereby distinguishing binding sites from the remainder of the surface. Using multiple structure alignment, these authors also showed the correspondence between energy hot spots and structurally conserved residues. Three residues (Trp, Phe and Met) were observed to be significantly conservative in binding sites. These identified local structures are linked with binding functions.

All residues in a protein are not equally important. Some are essential for certain structures or functions, whereas others can be readily replaced. Conservation analysis is one of the most widely used techniques for predicting these functionally important residues in protein sequences. Capra and Singh (2007) proposed a method focusing on the analysis of a multiple sequence alignment of the homologous sequences in order to find columns that are preferentially conserved. The results show that conservation is highly predictive in identifying catalytic sites and residues near bound ligands, while it is much less effective in identifying residues in protein–protein interfaces.

Structure alignment: geometric hashing Rosen et al. (1998) proposed a surface comparison algorithm in search of active sites and functional similarity. These authors first represents the surface by a face-center critical point technique and then derive active sites using geometric hashing to match the two surfaces. Finally, a clustering process is used to obtain the functional active sites. This method addresses the question of the usefulness of geometric comparisons and concludes that pure geometric surface matching is capable of obtaining biological meaningful solutions. Based on the geometric hashing algorithm, Leibowitz et al. (2001) presented a multiple structural alignment algorithm to detect a recurring substructural motif. Given an ensemble of protein structures, the algorithm automatically finds the largest common substructure (core) of C_α atoms that appears in all of the molecules in the ensemble. The detection of the core and the structural alignment are carried out simultaneously. Fischer et al. (1994) also presented an approach using geometric hashing to compare spatial, sequence-order independent atoms. It automatically detects a recurring three-dimensional motif in protein molecules without any predefinition of the motif.

Pairwise alignment of constructed local structures There are several methods that detect the functional relationship between local structures by structure alignment in an all-against-all manner. Pazos and Sternberg (2004) presented an automatic method to extract functional sites (residues associated to functions). The method relates proteins with the same GO functions through structural alignment in an all-against-all manner and extracts three-dimensional profiles of conserved residues.

Based on the identified local structures derived from geometry or physicochemical features, the functional relationship of these local regions can be detected and the comparison result is stored in a database. When querying a local structure, similar hits imply functional relationships. Binkowski et al. (2003a, 2005) described such an approach for inferring functional relationships of proteins based on the pvSOAR by detecting sequence and spatial patterns of the functional relationship of pockets on protein surfaces. The pvSOAR database provides a pairwise comparison of the pockets in the pocket database CASTp. Similar pockets in different match degrees are searched for in an advanced analysis of the function relationship among the local structural motifs. With respect to the pockets on the protein surface, Schmitt et al. (2002) developed a similar method based on a clique detection algorithm by comparing the query against the whole database. Kinoshita and Nakamura (2003) also provided an analogous method for comparing molecular surface geometries and electrostatic potential on the surfaces based on eF-site. Their method bridges the protein surface electronic features of the local region with the specific functions. Jambon et al. (2003) designed a new but similar approach for finding similarities using pairwise matching to detect common three-dimensional sites in proteins. The basis for their method is a representation of the protein structure by a set of stereochemical groups.

Protein surface regions with similar physicochemical properties and shapes may perform similar functions and bind similar partners. Shulman-Peleg et al. (2005) constructed two web servers and software packages for use in recognizing the similarity of binding sites and interface—SiteEngine and Interface-to-Interface (I2I)-SiteEngine. The input into the two methods is two protein structures or two protein–protein complexes; the output is the surface of the proteins for a region similar to the binding sites or the interfaces. The methods are efficient for large-scale database searches of the entire PDB. Obviously, the two locally identified structures are related to functions by searching similar local regions of their protein structures.

Pairwise alignment of annotated local structures Information on functional sites obtained from databases or the literature can be used to construct the function-related local

structure database, while the pairwise alignment method is used to detect the functional relationships. Stark and Russell (2003a) developed PINTS to uncover the recurring three-dimensional side-chain patterns based on the algorithm in Stark et al. (2003c). Their method queries the structural motif database constructed from the annotation mining from PDB to find similar three-dimensional motifs by a recursive, depth-first search algorithm, i.e. to find all possible groups of identical amino acids common to two protein structures independent of sequence order (Russell 1998). The search is conducted with distance constraints by ignoring those amino acids unlikely to be involved in the protein function. Stark et al. (2003b) identified some functional sites and compared these with PROCAT and RIGOR. Moreover, PINTS provides a measure of statistical significance based on a rigorous model for the behavior of RMSD (Stark et al. 2003c).

SiteBase (Gold and Jackson 2006a) is a database of known ligand-binding sites within the PDB. Gold and Jackson (2006a) provided a method that automatically identifies ligand-binding sites by searching for HETATM keywords in PDB files and constructing a database by excluding protein/peptide ligands and treating Het-groups as individual ligand-binding sites. Protein atoms within a 5-Å radius of any ligand atom were defined as its binding site in this work, and the ligand-binding was identified by comparison in an all-against-all way with geometric hashing. Similar functions of binding sites were detected regardless of the sequence and folding similarity (Gold and Jackson 2006b). PDBSiteScan (Ivanisenko et al. 2004) provides an automatic search of three-dimensional protein fragments that are similar in structure to known functional sites. A collection of known sites has been designated as the PDBSite (Ivanisenko et al. 2005), which is a database of amino acid content, structure features calculated by spatial protein structures and the physicochemical properties of sites and their spatial surroundings. Protein–protein interaction sites are also generated by an analysis of contact residues in heterocomplexes. The algorithm is developed based on an exhaustive examination of all possible combinations of protein positions. The BID (Fischer et al. 2003) database searches the primary scientific literature directly for detailed data on protein interfaces by text mining and stores the characterization of protein–protein binding interfaces at the amino acid level. The BID also organizes protein interaction information into tables, graphical contact maps and descriptive functional profiles.

Evolutionary tracing Protein functional sites have a number of similar and unique features. In order to explore the information fully, one can incorporate both sequence and structure data in a functional site prediction method. The Evolutionary Trace (ET) method is one such method that relies on both sequence and structure information. The most basic form of the algorithm requires a multiple sequence alignment of a protein family and an evolutionary tree, based on sequence identity, which can approximate the functional classification of the protein sequences (Lichtarge and Sowa 2002).

Yao et al. (2003) proposed an automatic ET method that ranks the evolutionary importance of amino acids in protein sequences. This was the first method to quantify the significance of the overlap observed between the best-ranked residues and functional sites. The information inherent in a phylogenetic tree is added to the analysis of conserved sequences, often revealing the more subtle aspects of protein function. Starting with a multiple sequence alignment, a representative structure and a phylogenetic tree, this method evaluates conservation at each position in the alignment for different sequence similarity cut-offs. In its original implementation, residues were classified as variable, conserved or a group-specific set that is specific to one branch of the phylogenetic tree. This analysis can be further expanded by the use of amino acid substitution matrices to evaluate conservation. In either case, a representative structure is used to visualize the distribution of scores at the end of the analysis.

Based on the ET method, Landgraf et al. (2001) presented a three-dimensional cluster analysis that offers a method for predicting functional residue clusters. This method requires a representative structure and a multiple sequence alignment as input data. Individual residues are represented in terms of regional alignments that reflect both their structural environment and their evolutionary variation, as defined by the alignment of homologous sequences. The overall and regional alignments are calculated from the global and regional similarity matrices, which contain scores for all pairwise sequence comparisons in the respective alignments. Three-dimensional clustering analysis is an easily applied method for the prediction of functionally relevant spatial clusters of residues in proteins.

Armon et al. (2001) proposed the ConSurf method, which takes into account the evolutionary relationships among the sequence homologues by closely approximating the evolutionary process and by considering the phylogenetic relationships among the sequences and the similarity between amino acids. ConSurf maps evolutionary conserved regions on the surface of proteins with a known structure; it also aligns sequence homologues of the protein and uses the alignment to construct phylogenetic trees. The trees are then used to infer the presumed amino acid exchanges that occur throughout the evolution. Each exchange is then weighted by the physicochemical distance between the exchanged amino acid residues. The results show that the patches of conserved residues correlate well

with the known functional regions of the domains and are more sensitive than the ET method.

To obtain an indication of the validity of functional inheritance, Aloy et al. (2001) proposed a method to evaluate the reliability by exploiting the conservative functional sites predicted by the ET method. Their method first used a fully automatic procedure to carry out the ET method, and then was benchmarked in terms of required sequence divergence and the resultant selectivity and specificity of the prediction. Finally, the results that were obtained using the prediction of location of functional sites to assist in filtering putative complexes were evaluated.

Template-based comparison

The functional importance of local structures can be detected by empirical methods or by computational methods. The identified functional motif can then be used as the structure template to detect the functional regions in other protein structures. The chosen method often consists of a comparison process, and the structure and physicochemical features can be considered in the comparison to the templates. In addition, a measurement of the similarity to the template is used to assess the functional importance of the testing of local structures.

Wallace et al. (1997) described a three-dimensional template matching method based on geometric hashing for automatically deriving three-dimensional templates from the protein structures deposited in PDB. In their paper, these researchers described a template derived for the Ser–His–Asp catalytic triad. Their results showed that the resultant template provides a highly selective tool for automatically differentiating between catalytic and non-catalytic Ser–His–Asp associations.

Goyal and Mande (2007) described the generation of three-dimensional structural motifs for metal-binding sites from known metalloproteins. Using three-residue templates and four-residue templates, the method scans all available protein structures in the PDB database for putative metal-binding sites. The search of the whole PDB database predicted many novel metal-binding sites, which are the identified functional motifs.

Chakrabarti and Lanczycki (2007) recently performed a detailed survey of compositional and evolutionary constraints at the molecular and biological functional levels for a large set of known functionally important sites extracted from a wide range of protein families. They compared the degree of conservation across different functionally important sites. The compositional and evolutionary information at functionally important sites was compiled into a library of functional templates. In their paper, these researchers developed a module that predicts functionally important columns of an alignment based on the detection of a significant 'template match score' to a library template. Benchmark studies showed good sensitivity/specificity for the prediction of functional sites and high accuracy in attributing correct molecular function type to the predicted sites.

The comparison between potential sites and the templates is very important in these kinds of methods. Artymiuk et al. (1994) developed a program called ASSAM, which represents a motif-by-distance matrix between pseudo-atoms and uses the subgraph-isomorphism algorithms to find matches. This is an elegant method for the detection of common tertiary side-chain patterns based on the use of the Ullman subgraph isomorphism algorithm. Singh and Saha (2003) formulated the problem of identifying a given structural motif (pattern) in a target protein and discussed the notion of complete and partial matches. They described the precise error criterion that has to minimized and also discussed different metrics for evaluating the quality of partial matches. They also presented a novel polynomial time algorithm for solving the problem of matching a given motif in a target protein.

Feature-based method

The functions of a protein are strongly related to the physicochemical features of that protein. The physical features (such as geometry, size, depth and shape) and the chemical features (such as energy, hydrophobicity, amino acid propensity and conservation) of the local structure are often measured by a score function or learned by a machine learning algorithm. The functional importance and specificity of a protein can be identified from the evaluation score or the trained standards of features. The main methods are listed in Table 6. The scoring method can often calculate an explicit value for the features, while the learning method can reveal the patterns inexplicitly.

Scoring methods

The properties of local structures are believed to be conserved in terms of determining their functions. The identified local regions of structure are analyzed based on the variations in their properties, which are investigated using the identified functionally important sets of local structures. The method to predict the functions of the local structures is often based on a scoring scheme that is used to analyze the properties of the targets. In particular, the scores of the features are used as the measurements to determine whether the local structure has functional importance, for example, for a particular function.

Scoring by physical features First, the physical features of the local structures, such as size, depth and shape, are

Table 6 Feature-based methods for identifying functional motifs

Local structure	Feature	Software	Reference
Scoring for every features: physical features, such as shape, size, depth and geometry, among others			
DNA-binding sites	Interfacial geometry	IAlign	Siggers et al. (2005)
Pockets for binding	Size and depth	PHECOM	Kawabata and Go (2007)
Binding pockets	Shape	–	Morris et al. (2005)
Binding pockets	Geometrical complementary	–	Kahraman et al. (2007)
Chemical features, such as energy, potential and conservation, among others			
Functional important residues	Electrostatic energy and conservation	–	Elcock (2001)
Protein–ligand binding sites	Physicochemical energy	Q-sitefinder	Laurie and Jackson (2005)
Protein–DNA binding sites	Five characteristics of patches	Web server	Jones et al. (2003)
Protein–RNA binding sites	As the former DNA-binding sites and van der Waals	Web server	Jones et al. (2001)
Protein–DNA binding sites	Hydrogen bonds and van der Waals interactions	Web server	Luscombe et al. (2001)
Protein interface	Energy score, propensity, conservation	PINUP	Liang et al. (2006)
Functional sites	Sequence, Rosetta free energy	Web server	Cheng et al. (2005)
Functional residues	Conservation score	–	Panchenko et al. (2004)
Functional sites	Functional groups	CFG	Innis et al. (2004)
Combined feature, such as the former features			
Ligand-binding sites	Geometry and conservation score	LIGSITE^{csc}	Huang and Schroeder (2006)
Protein–DNA binding sites	Shape and electrostatic potential	–	Tsuchiya et al. (2004)
Carbohydrate-binding sites	Six parameters	–	Taroni et al. (2000)
Protein–protein interfaces	Structure and physicochemical	ProMate	Neuvirth et al. (2004)
Docking pockets	Geometry and energy	–	Li et al. (2004)
Protein–protein interfaces	Five parameters	–	Hoskins et al. (2006)
Ligand binding pockets	Cleft volume and residue conservation	SURFNET-Consurf	Glaser et al. (2006)
Learning the features: SVM			
Protein–protein interfaces	Sequence profile, amino acid composition	–	Koike and Takagi (2004)
Protein–protein interfaces	Evolutionary conservation signal	–	Bordner and Abagyan (2005)
Protein–DNA binding sites	Composition, charge, positive potential patches	Web server	Bhardwaj et al. (2005)
Binding sites	Sequence and structural complementary	–	Chung et al. (2007)
Neural network			
Protein–protein interfaces	Composition	–	Ofran and Rost (2003)
Protein–protein interfaces	Conservation and residues structure properties	PPISP	Zhou and Shan (2001)
Catalytic residues	Conservation, ASA, structure, depth	–	Gutteridge et al. (2003)
Protein–protein interaction sites	Conservation and disposition	ISPRED	Fariselli et al. (2002)
Nucleic-acid-binding sites	Ensemble features of sequence and structure	–	Stawiski et al. (2003)
DNA-binding sites	Sequence profiles and solvent accessibility	DISPLAR	Tjong and Zhou (2007)
DNA-binding sites	Structure, ASA and electrostatic potential	DbHTH	Ferrer-Costa et al. (2005)
Metal-binding site residues	Sequence and structure data	MetSite	Sodhi et al. (2004)
Binding sites	Physical and chemical property lists	–	Keil et al. (2004)
DNA-binding sites	Evolutionary conservation	DP-BIND	Kuznetsov et al. (2006)
Metal-binding sites	Evolutionary profiles	–	Passerini et al. (2006)
Describing the features by statistical methods			
Functional sites	Calculated feature vectors	FEATURE	Liang et al. (2003a)
Protein–protein binding site	Six parameters	PPI-Pred	Bradford et al. (2006)
Protein–protein interface	Amino acid clusters	–	Yan et al. (2004)
Protein–DNA binding sites	Residues and sequence entropy	–	Yan et al. (2006)
Protein–protein interaction sites	Motifs and coexpression	InSite	Wang et al. (2007b)
DNA-binding sites	Geometrical measures	–	McLaughlin and Berman (2003)

Table 6 continued

Local structure	Feature	Software	Reference
Drug-binding sites	408 attributes, 8 broad categories	SCREEN	Nayal and Honig (2006)
Metal-binding sites	Geometric features	CHED	Babor et al. (2008)
Zinc-binding sites	A physicochemical feature set	Web server	Ebert and Altman (2008)

considered for scoring the function-related features. The shape features alone may provide basic information for the analysis of the functional features related to the protein function.

Siggers et al. (2005) introduced a new method to structurally align interfaces observed in protein–DNA complexes. Their method is based on a procedure that describes the interfacial geometry in terms of the spatial relationships between individual amino acid–nucleotide pairs. They subsequently provided a yet newer method to study the determinants of binding specificity. Kawabata and Go (2007) proposed a new definition for pockets using two explicit adjustable parameters, the radii of small and large probe spheres, which correspond to the two physical properties, 'size' and 'depth'. A pocket region was defined as a space into which a small probe can enter, but a large probe cannot. Based on the geometric standards of large probe spheres, this method identified the binding site positions.

From the geometrical viewpoint, the methods described above need further improvement to describe or compare the global shape and the local structures. Morris et al. (2005) presented a novel technique for capturing the global shape of a protein's binding pocket or ligand. This method uses the coefficients of a real spherical harmonics expansion to describe the shape of a protein's binding pocket. Shape similarity is computed as the L2 distance in coefficient space. Kahraman et al. (2007) used a recently developed shape matching method to compare the shapes of protein-binding pockets to the shapes of their ligands. Their results indicate that pockets binding the same ligand show greater variation in their shapes than those which can be accounted for by the conformational variability of the ligand. This result suggests that geometrical complementarity in general is not sufficient to derive molecular recognition.

Scoring by chemical features Chemical features of local structures are very important for determining their functional specificity. These feature scores of local structures can be used as standards to determine their functions.

The structural locations of functional sites are conserved between homologous proteins because functionally important residues tend to cluster together in space, forming three-dimensional residue clusters or surface patches. Panchenko et al. (2004) presented a method to assign each residue a score that depends on its own conservation in homologs and the conservation of residues in its spatial neighborhood. The high-scoring sites are more likely to be involved in specific binding or catalysis. Functionally important residues in a protein are known to be those computed to have energy among experimentally destabilized residues. Elcock (2001) proposed a method to predict functionally important residues based solely on the computed energetics of a protein structure. The energetic properties of binding surfaces in protein–protein interfaces and protein–ligand sites were shown to be different (Burgoyne and Jackson 2006). The pockets from Q-sitefinder (Laurie and Jackson 2005) were ranked by the scores of these properties—i.e. hydrophobicity, desolvation, electrostatics and conservation—which are used to determine binding sites.

Jones et al. (2003) developed a method to detect DNA-binding sites on a protein surface. The surface patches and the DNA-binding sites were initially analyzed for accessibility, electrostatic potential, residue propensity, hydrophobicity and residue conservation. In general, DNA-binding sites are among the top 10% of patches with the largest positive electrostatic scores. This knowledge was used to make predictions. Jones et al. (2001) presented a similar computational analysis of protein–RNA interactions. There are a number of differences between DNA-binding sites and RNA-binding sites. For the RNA-binding sites, van der Waals contacts play a more important role than hydrogen bond contacts. As to the protein–DNA binding local structures, Luscombe et al. (2001) investigated hydrogen bonds as well as van der Waals contacts and water-mediated bonds to assess whether there are universal rules that govern amino acid–base recognition. In a subsequent study, Luscombe and Thornton (2002) also identified the amino acid conservation and the effects of mutations on binding specificity.

In Liang et al. (2006), an empirical score function consisting of a linear combination of the energy score, interface propensity and residue conservation score is used to predict interface residues. The top-ranked patches are predicted to be the potential interface sites. The accuracy of prediction has been improved significantly, relative to any single or pairwise combination, by combining the three terms. Cheng et al. (2005) presented a method to predict protein function site using sequence alignment information

as well as Rosetta protein design and Rosetta free energy calculations. Logistic regression with the generalized linear model has been used to the determine weights of the sequence conservation, natural/designed sequence profile difference and natural/optimal residue free energy gap, all of which optimize the separation between functional and non-functional residues.

Innis et al. (2004) presented conserved functional group (CFG) analysis to predict function sites in proteins. The method relies on a simplified representation of the chemical groups found in amino acid side-chains to identify functional sites from a single protein structure and a number of its sequence homologs.

Scoring by physicochemical features Those features based only on physical geometry or chemical energy often can not represent functional features comprehensively. Most of the methods are used to integrate several important features together and then score these features for bridging the gaps between local structures and functions.

The LIGSITE algorithm is based only on the geometry. Huang and Schroeder (2006) presented an extension and implementation method, LIGSITE^{csc}, which is based on the notion of surface–solvent–surface events and the degree of conservation of the involved surface residues. The use of the Connolly surface has led to slight improvements, whereas the prediction re-ranking significantly improved the binding site predictions. Glaser et al. (2006) improved previous approaches by combining two known measures of 'functionality' in proteins, i.e. cleft volume and residue conservation, to develop a method for identifying the location of ligand-binding pockets in proteins.

Neuvirth et al. (2004) proposed a structure-based algorithm to identify the location of protein–protein interaction sites. The sites are defined based on Connolly's molecular dot surfaces. The method defines an interface score that combines the chemical and geometry features of the interaction sites. Interfacial residues are considered to be those with the 10% highest scores. Geometry and energy properties have also been used to analyze the pocket functions for docking (Li et al. 2004). Hoskins et al. (2006) considered the use of solvent accessibility, residue propensity and hydrophobicity in conjunction with secondary structure data as prediction parameters to predict protein–protein interaction sites. The influence of residue type and secondary structure on solvent accessibility is analyzed, and a measure of relative exposedness is defined. The high-scoring residues are clustered as a basis for predicting interaction sites.

Tsuchiya et al. (2004) provided a method for analyzing protein–DNA complexes, focusing on the shape of the molecular surface of the protein and DNA, along with the electrostatic potential on the surface, and calculated a new evaluation score. Based on the score, the method was used to classify DNA-binding from non-DNA-binding proteins. Taroni et al. (2000) provided an analysis of the characteristic properties of sugar-binding sites. For each site, six parameters were evaluated—i.e. solvation potential, residue propensity, hydrophobicity, planarity, protrusion and relative accessible surface area (ASA). Three of the parameters were found to distinguish the observed sugar-binding sites from the other surface patches. These parameters were then used to calculate the probability of a surface patch being a carbohydrate-binding site. The total score of the properties was used to determine whether the surface patch was a carbohydrate-binding site.

Learning methods

The features of the local structures play crucial roles in predicting protein function. To identify the relationship between protein local structure and protein function, the structural and/or physicochemical features can be learned implicitly using machine learning methods, such as the support vector machine (SVM) and neural network.

Support vector machine The support vector machine uses a linear model to implement nonlinear class boundaries through the input of a number of nonlinear mapping vectors into a high-dimensional feature space. It is based on mathematics theory and has many successful applications in statistical learning fields (Vapnik 1998). These methods have been confirmed to be able to learn the features of local structures with functional importance. The features can first be investigated in the learning process and used to detect whether these features relate some specific functions. Koike and Takagi (2004) proposed an SVM method to identify protein–protein interaction sites. The profiles of sequentially/spatially neighboring residues, plus additional information, constitute a feature vector, and the interaction site ratios are calculated by SVM regression. The predictive performance is evaluated and compared in different quantitative features. Cai et al. (2004) proposed an SVM algorithm to predict the catalytic triad of the serine hydrolase family. Bordner and Abagyan (2005) proposed a similar SVM to predict protein–protein interfaces. The local surface properties with a combination of an evolutionary conservation signal were used to train the machine on a large nonredundant data set of protein–protein interfaces. An SVM learning protocol was provided by Bhardwaj et al. (2005) for the prediction of DNA-binding proteins. The characteristics, including surface and overall composition, charge and positive potential patches on the protein surface, were derived, and the SVM was trained as a classifier to detect the DNA-binding proteins. The high accuracy value has been achieved in a large set of testing

proteins regardless of their sequence or structure homology. Chung et al. (2007) recently exploited the SVM approach to detect whether identified potential protein-binding sites interact with each other. The information related to sequence and structural complementary across protein interfaces were extracted from the PDB. This work also built a pipeline to predict the location of binding sites.

Neural network The neural network is a learning method which adapts the relationships of neurons; as such, it is a simplified model of the neural processing of the human brain (Zhang 2000). Based on the analysis of the both structures and sequences, Gutteridge et al. (2003) used a neural network to identify catalytic residues in enzymes. The locations of the active sites were predicted by the neural network output and spatial clustering of the highest scoring residues. In most testing cases, the likely functional residues were identified correctly, as were a number of potentially novel functional groups.

Ofran and Rost (2003) described a neural network to identify protein–protein interfaces from sequences. Since the compositions of contacting residues of the interaction sites were believed to be unique, the features of this known interaction sites were used to train the neural network. Zhou and Shan (2001) trained a neural network to predict protein–protein interactions. Their method combines conservation and structural properties of individual residues. Fariselli et al. (2002) reported a neural network-based system using information on evolutionary conservation and surface disposition. Chen and Zhou (2005) also provided a neural network method to predict interface residues in a protein–protein complex.

There are also neural network methods for predicting nucleic acid-binding (NA-binding) sites. Stawiski et al. (2003) presented an automatic neural network approach to predict NA-binding proteins, specifically DNA-binding proteins. This method uses an ensemble of features extracted from characterization of the structural and sequence properties of large, positively charged electrostatic patches. Structural and physical properties of DNA provide important constraints on the binding sites formed on the surfaces of the DNA-targeting proteins. The characteristics of DNA-binding sites may form the basis for predicting DNA-binding sites from the structures of proteins alone. Tjong and Zhou (2007) used a representative set of protein–DNA complexes from the PDB to analyze characteristics and to train a neural network predictor of DNA-binding sites. The input to the predictor consists of PSI-BLAST sequence profiles and solvent accessibility of each surface residue and 14 of its closest neighboring residues. Ferrer-Costa et al. (2005) provided a web-based method to detect if a protein structure contains a DNA-binding helix-turn-helix (DbHTH) motif. The method uses a neural network with no hidden layers, i.e. a linear predictor, to classify whether a protein is DNA-binding with the HTH motif. The linear predictor was trained on a non-homologous set of 79 structures of protein chains with a DbHTH motif and 490 without the motifs.

Sodhi et al. (2004) used a neural network to predict metal-binding sites residues in low-resolution structural models. The method involves sequence profile information combined with approximate structural data. Several neural networks were proposed to distinguish the metal sites from non-sites and then to detect these functionally important regions. In Keil et al. (2004), the patches of the molecular surface were segmented into overlapping patches. The properties of these patches were calculated based on the physical and chemical properties. A neural network strategy was then used to identify possible binding sites by classifying the surface patches as protein–protein, protein–DNA, protein–ligand or nonbinding sites.

Kuznetsov et al. (2006) applied an SVM method to predict DNA-binding sites using the features including amino acid sequence, profile of evolutionary conservation of sequence positions, and low-resolution structural information. The results indicate that an SVM predictor based on a properly scaled profile of evolutionary conservation in the form of a position specific scoring matrix (PSSM) significantly outperforms a PSSM-based neural network predictor. Such results imply that the combination of the two methods may improve the accuracy. Passerini et al. (2006) introduced a two-stage learning method for identifying histidines and cysteines that participate in binding of several transition metals and iron complexes. The first stage is an SVM, which is trained to locally classify the binding state of single histidines and cysteines. The second stage is a neural network trained to refine local predictions. The methods use only sequence information by utilizing position-specific evolutionary profiles.

Statistical methods Statistical learning also provides an effective way to link the features of local structures with their functional implication. Liang et al. (2003a) provided a supervised learning algorithm, FEATURE, for the automatic discovery of physical and chemical descriptions of protein microenvironments. The calculated feature vectors were used to predict functional motifs based on Bayesian inference. The method has also been proposed as an interactive web tool, WebFEATURE, for identifying and visualizing functional sites (Liang et al. 2003b).

Bradford et al. (2006) developed a method to predict both protein–protein binding site location and interface type (obligate or non-obligate) using a Bayesian network in combination with surface patch analysis. Two Bayesian network structures, naive and expert, were trained to

distinguish interaction surface patches. Wang et al. (2007b) proposed a computational method learned by the Expectation Maximization (EM) algorithm, InSite, to search for motifs whose presence in a pair of interacting proteins determined which motif pairs have high affinity that would lead to an interaction between proteins. Yan et al. (2004) also provided a two-stage method consisting of an SVM and a Bayesian classifier for predicting the surface residues of proteins that participate in protein–protein interaction. The method exploits the fact that interface residues tend to form clusters in the primary amino acid sequence. In addition, Chou and Cai (2004) provided a covariant discriminant algorithm to predict active sites of enzyme molecules. The high accuracy of prediction shows the effectiveness of the method.

Protein–DNA interactions are critical for deciphering the mechanisms of gene regulation. Yan et al. (2006) presented a supervised machine learning approach for the identification of amino acid residues involved in protein–DNA binding sites. A naive Bayesian classifier was trained for predicting whether a given amino acid residue is a DNA-binding residue based on its identity and the identities of its sequence neighbors. McLaughlin and Berman (2003) developed statistical models for discerning protein structures containing the DbHTH motifs. The method uses a decision tree model to identify the key structural features required for DNA binding. These features include a high average solvent-accessibility of residues within the recognition helix and a conserved hydrophobic interaction between the recognition helix and the second alpha helix preceding it. The Adaboost algorithm was used to search the PDB with the aim of identifying the structure containing the motifs with high probability.

Metal ions are crucial in facilitating the function of a protein. Identifying the features of metal binding sites provides crucial knowledge of the function performance of the local structures. Because the residues that coordinate a metal often undergo conformational changes upon binding, the detection of binding sites based on simple geometric criteria in proteins without bound metal is difficult. However, aspects of the physicochemical environment around a metal-binding site are often conserved, even when this structural rearrangement occurs. Ebert and Altman (2008) developed a Bayesian classifier using known zinc-binding sites as positive training examples and nonmetal-binding regions as negative training examples. Babor et al. (2008) reported an approach that identifies transition metal-binding sites in proteins by combining the decision tree and SVM. In the first step, the geometric search of structural rearrangements following metal binding was taken into account by a decision tree classifier. A second classifier based on SVMs was then used to identify the metal-binding sites.

Nayal and Honig (2006) proposed a comprehensive method to identify drug-binding sites in which 408 attributes were first computed for each cavity, and these were then used to distinguish drug-binding sites by the random forest classification scheme. The cavity properties cover eight broad categories, such as cavity size, cavity shape, hydrophobicity, electrostatics, hydrogen bonding, amino acid composition, secondary structure and rigidity.

Network-based method

An interesting method to identify function motifs is based on the graph theory and the network concept. The main methods are listed in Table 7. One subcategory of the method represents the protein structure as a complex network. A node represents a C_α of the backbone, and an edge linking two nodes represents the physical distance or the functional relationship between the nodes. Greene and Higman (2003) viewed protein structures as network systems. The systems are identified to exhibit small-world, single-scale and, to some degree, scale-free properties.

Using the network model, Amitai et al. (2004) identified active site residues. The method transforms a protein structure into a residue interaction graph, where graph nodes represent amino acid residues, and links represent their interactions. The active site, ligand-binding and evolutionary conserved residues are identified typically with a high closeness value, from which the functional residues are filtered out. del Sol et al. (2006) also represented a protein

Table 7 Network-based methods for identifying functional motifs

Local structure	Method	Software	Reference
Micro level: mining the special residues or subgraphs in the structure graphs			
Active site residues	High closeness value of residue interaction graphs	RIG	Amitai et al. (2004)
Functional residues	Residues of special topology in small-world network	–	del Sol et al. (2006)
Recurring side-chain patterns	Searching for similar subgraph	DRESPAT	Wangikar et al. (2003)
Structure motifs	Mining for cliques of the structure graph	CliqueHashing	Huan et al. (2006)
Macro level: similar groups of local structures			
Functional pockets	Similar pocket groups	PSN	Liu et al. (2007b)

structure as a small-world network and searched the topological determinants related to functionally important residues. The method investigates the performance of residues in protein families. The results indicate that enzyme active sites are located in surface clefts, and hetero-atom binding residues have deep cavities, while protein–protein interactions involve a more planar configuration.

Wangikar et al. (2003) reported a method for detecting recurring side-chain patterns using an unbiased and automatic graph theoretic approach. The method first lists all structural patterns as subgraphs. The patterns are compared in a pairwise manner based on content and geometry criteria. The recurring pattern is then detected using an automatic search algorithm from the all-against-all pairwise comparison proteins. Similarly, Huan et al. (2006) defined a labeled graph representation of a protein structure in which edges connecting pairs of residues are labeled by the Euclidian distance between the C_α atoms of the two residues. Based on this representation, a structural motif corresponding to a labeled clique occurs frequently among the graphical representation of the protein structures. The paper further presented an efficient mining algorithm aimed at discovering structure motifs in this setting.

In studies on protein structure and function, identifying calcium-binding sites in proteins is one of the first steps towards predicting and understanding the role of calcium in biological systems. Calcium-binding sites are often complex and irregular, and it is difficult to predict their location in protein structures. Deng et al. (2006) reported a rapid and accurate method for detecting calcium-binding sites. This algorithm uses a graph theory algorithm to identify oxygen clusters of the protein and a geometric algorithm to identify the center of these clusters. A cluster of four or more oxygen atoms has a high potential for calcium binding. A potential calcium-binding position is a clique and can be detected by a clique-detecting algorithm. The high accuracy of prediction shows that the majority of calcium-binding sites in proteins are formed by four or more oxygen atoms in a sphere center with a calcium atom.

The above network methods all focus on individual proteins and represent a protein structure a complex network. The specific topology features clearly imply a particular function module (Zhang and Grigorov 2006; Zhang et al. 2007). Recently, a novel category of network-based analysis of the protein local structures at the macro level has been proposed (Liu et al. 2008). The similarity of the local structures, specifically the pockets on the protein surface, is mapped to constitute a similarity network. The nodes represent the pockets, and the edges represent the certain similarity relationships among the pockets. The properties of the pocket similarity network are like other complex networks (Liu et al. 2008). The similar pockets are identified by the clusters and community structures, and the special features of the network are helpful in clustering the pockets into similar groups (Liu et al. 2007b), which may imply clusters of structure motifs and correspond to special functional implications (Liu et al. 2008). With the network concept, the pockets can also be used to characterize and predict protein functions by annotating the topology neighbors. In this way, the accuracy of the prediction is better than that with the global structural similarity approach (Liu et al. 2007a).

Discussion and future directions

Prediction of functions at the cellular level

Most of the methods used to annotate protein functions that are listed above are based on molecular function at the biological processing level. At the cellular component and location levels, the importance of protein local structure is also critical. In fact, information on the subcellular locations of proteins is important because it can provide useful insights into protein functions as well as how and in what kind of cellular environments they interact with each other and with other molecules. Such information is also fundamental and indispensable to systems biology because a knowledge of the localization of proteins within cellular compartments can facilitate our understanding of the intricate pathways that regulate biological processes at the cellular level. From this perspective, the functions of proteins at different levels are strongly inter-related to each other. At the cellular component level, local structures are still crucial in determining the roles of proteins and specific functions.

Many methods for predicting the subcellular location of proteins have been proposed recently because the location of such proteins in the cell can provide useful insights or clues about their functions (Chou and Shen 2007b). One of the more powerful methods applied in location prediction is based on an important descriptor of the protein sample, i.e. the pseudo-amino acid (PseAA) composition (Chou 2001). This descriptor can be used to represent a protein sequence with a discrete model yet without completely losing the sequence-order information. Since the concept of PseAA composition was introduced, various PseAA composition approaches have been developed, all with the aim of improving the prediction quality of protein attributes (Gao et al. 2005; Zhang et al. 2006; Zhou et al. 2007a, b; Diao et al. 2008; Fang et al. 2008; Li and Li 2008). The PseAA method has been widely used and extended. A very flexible PseAA composition generator (PseAAC) was established (Shen and Chou 2008) which enables users to generate 63 different kinds of PseAA composition. A web

server called Cell-PLoc (Chou and Shen 2008) has recently been developed that allows users to predict the subcellular locations of proteins in various different organisms. PseAA composition and PSSM have also been combined in various algorithms to improve the prediction quality for membrane protein type (i.e. MemType-2L: Chou and Shen 2007a), enzyme main-functional class and sub-functional class (i.e. EzyPred: Shen and Chou 2007a) and protein subnuclear localization (i.e. Nuc-PLoc: Shen and Chou 2007b). A comprehensive review (Chou and Shen 2007b) published recently provides a summary of these topics. In addition to sequence information, local structural information is useful, interesting and important in protein localization function prediction.

Validation of function prediction

A quality assessment of the results is necessary at all three levels of function prediction. The predicted functions of proteins can be taken as indicators of the directions to be taken by researchers when carrying out experiments to validate the functions of proteins. Many of the computational methods used to annotate protein functions as well as those used to predict functionally important local structures use cross-validation methods to assess the performance of a prediction; these include the independent dataset test, subsampling test and jackknife test (Chou and Zhang 1995). However, as elucidated by Chou and Shen (2008), of these cross-validation methods, the jackknife test is considered to be the most objective and has been increasingly used by investigators to examine the accuracy of various predictors (Zhou 1998; Zhou and Assa-Munt 2001; Zhou and Doctor 2003; Xiao et al. 2005; Zhou and Cai 2006; Chen et al. 2007; Shi et al. 2008). It is important to consider the relationship among the functional terms and the semantic similarity with the aim of avoiding biases in the assessment of functional similarity (Liu et al. 2007a).

Local versus global structure to function

The global structure similarity-based methods provide a straightforward approach to annotate protein functions. However, since the relationships between structures and functions are so complex, local structure-based methods can be used to predict protein function directly by identifying the local structures carrying out particular functions. Laskowski et al. (2005) proposed a novel method of predicting protein function using local three-dimensional templates. The authors build a template database and use four types of templates—enzyme active sites, ligand-binding residues, DNA-binding residues and reverse templates—to construct the relationship between templates and functions.

Ferre et al. (2005) described a method for the function-related annotation of protein structures based on the detection of local structural similarity with a library of annotated functional sites. An automatic procedure was used to annotate the function of the local surface regions, and then a sequence-independent algorithm was developed to compare exhaustively these functional patches with a larger collection of protein surface cavities. After tuning and validating the algorithm on a dataset of well-annotated structures, the results are able to provide functional clues to proteins that do not show any significant sequence or global structural similarity with proteins in the current databases.

Binkowski et al. (2005) provided similar methods to annotate protein functions from the protein surface similarity. Pockets are identified by CASTp from several proteins. These pockets are queried in the pvSOAR to locate similar pockets corresponding to annotated proteins. The conservation among the pockets can be detected by the sequence identities and other similarity metrics. Tseng and Liang (2006) developed a Bayesian Markov chain Monte Carlo method for rate estimation of the special substitution rates of the short sequence of local structure. Moreover, a method for protein function prediction is presented by surface matching using scoring matrices derived from estimated substitution rates for residues located on the binding surfaces. The method is effective in identifying functionally related proteins that have overall low sequence identity. The method provided by Pazos and Sternberg (2004) first identifies functional sites in proteins by bridging the local structures and functions, then the functions of a target proteins can be inferred from the similarity of the functional sites in the position-specific scoring matrices.

Information on the functional importance of local structure can facilitate the annotation of protein function more precisely. George et al. (2005) proposed an effective method to annotate protein function through the use of functional clues of conservation among the catalytic residues. This method improves the precision of annotation significantly.

The advantages of predicting protein functions from local structures are based on the fact that such methods can be implemented without any prior homology hypothesis. The methods can be used in proteins in midnight zone without sequence similarity, and local structures often provide concrete and specific functional annotations. To compare the precision and coverage of the global structural similarity and that of local structures, Liu et al. (2007a) proposed a novel method to predicted protein from the pockets on the protein's local surface region. The similarity of regional local surface pockets and the global similarity of proteins are all represented by networks. The prediction

is based on the network topology. A comparison of the results show that the local-structure-based prediction is better than the global-structure-based prediction (Liu et al. 2007a).

Future directions

In this paper, we have reviewed protein function prediction methods at different levels, i.e. sequence, structure, interaction and integration. We have mainly focused on the importance of local structures and the method used to predict functionally important local structures. In summary, we discuss possible future directions.

The interaction between proteins provides high-level information on protein function, especially in various biological processes. Although there are thousands of known interactions, a tiny fraction of these are available in precise molecular details. If we are able to examine structural details, systematic representation of the interaction would accurately reflect biological reality. For example, we can predict which part of the structures is most likely to be involved in interaction with other macromolecules, proteins, DNA or RNA by analyzing the properties of different local patches on the protein surface. The patch analysis, which considers properties of the surface such as flatness, hydrophobicity, charge and, in particular, residue conservation, is effective in identifying protein–protein interaction surfaces and has also been shown to successfully identify DNA-binding sites (Aloy and Russell 2006). Structural systems biology is a very effective approach that combines protein interactions and protein three-dimensional structures. The mechanisms of protein and protein interaction lie in the local structures between the two protein surfaces. From this perspective, structural systems biology provides us with a new direction in the fields of structural biology and systems biology. It combines the key features of the two directions to provide more insight into linking the single protein and systematic interaction between proteins. The relationships between local structures and functions are expected to play important roles in structural systems biology.

The computational methods used to bridge the relationship between local structures and functions can be further improved. The community of computational biology has a strong need for comprehensive feature selection in concise and effective ways. In addition, there is still much room for improvement in terms of the accuracy of the methods used to align the features between two local structures. The validation of the functions of structural motifs should also be conducted more carefully and by more reliable biological experiments. Recent advances in the field inspired by developments in sequences and structures demonstrate the great potential of such research in protein science in elucidating essential functional roles of the local structures. In our opinion, research aimed at bridging the gaps between local structures and function is still in its infant stage, and further advances in such areas will greatly enhance our ability to study the fundamental properties of proteins at a system-wide level. In other words, we expect to gain deep insight into essential mechanisms of biological systems from both structural and functional perspectives.

Different methods based on the local similarity, global similarity and interaction require and use different information, and they have different aspects, intentions and advantages. To our knowledge, the function annotation problem is still in its developing period and needs more comprehensive or hybrid approaches. None of the existing methods are likely to be successful in all cases to annotate a protein with its functions correctly and comprehensively. One reason for this is that protein functions not only rely on the sequence and/or folding characteristics, but also on the cell environment, the cycle of the biological processes and other chemical compounds. There are still many difficult-to-decipher proteins that researchers have been unable to annotate correctly by any existing method. Hence, a sensible strategy is to use different methods to incorporate data from multiple sources and to extensively utilize existing function annotations. Future directions include using combinations of different methods at different levels so as to efficiently explore the overall sequences, global structures and local structures and to obtain more information on interactions between the target proteins and others in the cellular context. Although computational methods generally cannot directly validate protein functions, the predefined tentative annotations provide valuable information as a basis for further efficient validation experiments.

Acknowledgments This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 10631070 and No. 60503004. LYW and XSZ are also supported by the Grant No. 5039052006CB from the Ministry of Science and Technology, China. The research was also supported by NSFC-JSPS collaborative project No. 10711140116. The authors are grateful to the anonymous referees as well as editors for comments and for helping to improve the earlier version. We recognize that this review is far from comprehensive, and we apologize for any papers related to the subject that were not mentioned.

References

Aloy P, Russell RB (2006) Structural systems biology: modelling protein interactions. Mol Cell Biol 7:188–197

Aloy P, Querol E, Aviles FX, Sternberg MJ (2001) Automatic 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 311:395–408

Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, Pietrokovski S (2004) Network analysis of protein structures identifies functional residues. J Mol Biol 344:1135–1146

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

Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P (1994) A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structure. J Mol Biol 243:327–344

Babor M, Gerzon S, Raveh B, Sobolev V, Edelman M (2008) Prediction of transition metal-binding sites from apo protein structures. Proteins 70:208–217

Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113

Barondeau DP, Getzoff ED (2004) Structural insights into protein–metal ion partnerships. Curr Opin Struct Biol 14:765–774

Barrett AJ (1997) Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement 4: corrections and additions. Eur J Biochem 250:1–6

Bhardwaj N, Langlois RE, Zhao G, Lu H (2005) Kernel-based machine learning protocol for predicting DNA-binding proteins. Nucleic Acids Res 33:6486–6493

Binkowski TA, Adamian L, Liang J (2003a) Inferring functional relationships of proteins from local sequence and spatial surface patterns. J Mol Biol 332:505–526

Binkowski TA, Naghibzadeh S, Liang J (2003b) CASTp: computed atlas of surface topography of proteins. Nucleic Acids Res 31:3352–3355

Binkowski TA, Freeman P, Liang J (2004) pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. Nucleic Acids Res 32:W555–W558

Binkowski TA, Joachimiak A, Liang J (2005) Protein surface analysis for function annotation in high-throughput structural genomics pipeline. Protein Sci 14:2972–2981

Bordner AJ, Abagyan R (2005) Statistical analysis and prediction of protein–protein interfaces. Proteins 60:353–366

Borman S (2008) Flu virus proton channel analyzed: structures of key surface protein suggest different drug mechanisms. Chem Eng News 86:53–54

Bradford JR, Needham CJ, Bulpitt AJ, Westhead DR (2006) Insights into protein–protein interfaces using a Bayesian network prediction method. J Mol Biol 362:365–386

Brenner SE (2001) A tour of structural genomics. Nat Rev Genet 2:801–809

Burgoyne NJ, Jackson RM (2006) Predicting protein interaction sites: binding hot-spots in protein–protein and protein–ligand interfaces. Bioinformatics 22:1335–1342

Cai YD, Zhou GP, Jen CH, Lin SL, Chou KC (2004) Identify catalytic triads of serine hydrolases by support vector machines. J Theor Biol 228:551–557

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R (2004) The gene ontology annotation (GOA) database: sharing knowledge in Uniprot with gene ontology. Nucleic Acids Res 32:D262–D266

Campbell SJ, Gold ND, Jackson RM, Westhead DR (2003) Ligand binding: functional site location, similarity and docking. Curr Opin Struct Biol 13:389–395

Capra JA, Singh M (2007) Predicting functionally important residues from sequence conservation. Bioinformatics 23:1875–1882

Chakrabarti S, Lanczycki CJ (2007) Analysis and prediction of functionally important sites in proteins. Protein Sci 16:4–13

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

Chen J, Liu H, Yang J, Chou KC (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33:423–428

Chen L, Wu LY, Wang Y, Zhang S, Zhang XS (2006) Revealing divergent evolution, identifying circular permutations and detecting active-sites by protein structure comparison. BMC Struct Biol 6:18

Cheng G, Qian B, Samudrala R, Baker D (2005) Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. Nucleic Acids Res 33:5861–5867

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition (Erratum: ibid., 2001, Vol. 44, 60). Proteins 43:246–255

Chou KC (2004) Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134

Chou KC, Cai YD (2004) A novel approach to predict active sites of enzyme molecules. Proteins 55:77–82

Chou KC, Cai YD (2006) Predicting protein–protein interactions from sequences in a hybridization space. J Proteome Res 5:316–322

Chou KC, Shen HB (2007a) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360:339–345

Chou KC, Shen HB (2007b) Recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16

Chou KC, Shen HB (2008) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3:153–162

Chou KC, Zhang CT (1995) Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349

Chou KC, Wei DQ, Zhong WZ (2003) Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS (Erratum: ibid., 2003, Vol.310, 675). Biochem Biophys Res Commun 308:148–151

Chung JL, Wang W, Bourne PE (2007) High-throughput identification of interacting protein–protein binding sites. BMC Bioinformatics 8:223

del Sol A, Fujihashi H, Amoros D, Nussinov R (2006) Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. Protein Sci 15:2120–2128

Deng H, Chen G, Yang W, Yang JJ (2006) Predicting calcium-binding sites in proteins—a graph theory and geometry approach. Proteins 64:34–42

Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41:98–107

Diao Y, Ma D, Wen Z, Yin J, Xiang J, Li M (2008) Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel–Ziv complexity. Amino Acids 34:111–117

Du QS, Wang SQ, Chou KC (2007) Analogue inhibitors by modifying oseltamivir based on the crystal neuraminidase structure for treating drug-resistant H5N1 virus. Biochem Biophys Res Commun 362:525–531

Ebert JC, Altman RB (2008) Robust recognition of zinc binding sites in proteins. Protein Sci 17:54–65

Eisenberg D, Marcotte EM, Xenarios I, Yeates TO (2000) Protein function in the post-genomic era. Nature 405:823–826

Elcock AH (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 312:885–896

Fang Y, Guo Y, Feng Y, Li M (2008) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34:103–109

Fariselli P, Pazos F, Valencia A, Casadio R (2002) Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 269:1356–1361

Ferre F, Ausiello G, Zanzoni A, Helmer-Citterich M (2004) SURFACE: a database of protein surface regions for functional annotation. Nucleic Acids Res 32:D240–D244

Ferre F, Ausiello G, Zanzoni A, Helmer-Citterich M (2005) Functional annotation by identification of local surface similarities: a novel tool for structural genomics. BMC Bioinformatics 6:194

Ferrer-Costa C, Shanahan HP, Jones S, Thornton JM (2005) HTHquery: a method for detecting DNA-binding proteins with a helix-turn-helix structural motif. Bioinformatics 21:3679–3680

Fischer D, Wolfson H, Lin SL, Nussinov R (1994) 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 3:769–778

Fischer TB, Arunachalam KV, Bailey D, Mangual V, Bakhru S, Russo R, Huang D, Paczkowski M, Lalchandani V, Ramachandra C, Ellison B, Galer S, Shapley J, Fuentes E, Tsai J (2003) The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. Bioinformatics 19:1453–1454

Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28:373–376

George RA, Spriggs RV, Bartlett GJ, Gutteridge A, MacArthur MW, Porter CT, Al-Lazikani B, Thornton JM, Swindells MB (2005) Effective function annotation through catalytic residue conservation. Proc Natl Acad Sci USA 102:12299–12304

Gerstein M, Levitt M (1998) Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. Protein Sci 7:445–456

Gibrat JF, Madej T, Bryant SH (1996) Surprising similarities in structure comparison. Curr Opin Struct Biol 6:377–385

Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM (2006) A method for localizing ligand binding pockets in protein structures. Proteins 62:479–488

Gold ND, Jackson RM (2006a) SiteBase: a database for structure-based protein–ligand binding site comparison. Nucleic Acids Res 34:D231–D234

Gold ND, Jackson RM (2006b) Fold independent structural comparisons of protein–ligand binding sites for exploring functional relationships. J Mol Biol 355:1112–1124

Goldsmith-Fischman S, Honig B (2003) Structural genomics: computational methods for structure analysis. Protein Sci 12:1813–1821

Goyal K, Mande SC (2007) Exploiting 3D structural templates for detection of metal-binding sites in protein structures. Proteins 70:1206–1218

Greene LH, Higman VA (2003) Uncovering network systems within protein structures. J Mol Biol 334:781–791

Gutteridge A, Bartlett GJ, Thornton JM (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. J Mol Biol 330:719–734

Huan J, Bandyopadhyay D, Prins J, Snoeyink J, Tropsha A, Wang W (2006) Distance-based identification of spatial motifs in proteins using constrained frequent subgraph mining. In: Proc LSS Computational Systems Bioinformatics Conference (CSB), pp 227–238

Huang B, Schroeder M (2006) LIGSITE^{csc}: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol 6:19

Hendlich M, Rippmann F, Barnickel G (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 15:359–63,389

Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. J Mol Biol 233:123–138

Holm L, Sander C (1996) Mapping the protein universe. Science 273:595–602

Hoskins J, Lovell S, Blundell TL (2006) An algorithm for predicting protein–protein interaction sites: abnormally exposed amino acid residues and secondary structure elements. Protein Sci 15:1017–1029

Hou J, Jun SR, Zhang C, Kim SH (2005) Global mapping of the protein structure space and application in structure-based inference of protein function. Proc Natl Acad Sci USA 102:3651–3656

Innis CA, Anand AP, Sowdhamini R (2004) Prediction of functional sites in proteins using conserved functional group analysis. J Mol Biol 337:1053–1068

Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA (2004) PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. Nucleic Acids Res 32:W549–W554

Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA (2005) PDBSite: a database of the 3D structure of protein functional sites. Nucleic Acids Res 33:D183–D187

Jambon M, Imberty A, Deleage G, Geourjon C (2003) A new bioinformatic approach to detect common 3D sites in protein structures. Proteins 52:137–145

Jones S, Thornton JM (1996) Principles of protein–protein interactions. Proc Natl Acad Sci USA 93:13–20

Jones S, Thornton JM (2004) Searching for functional sites in protein structures. Curr Opin Chem Biol 8:3–7

Jones S, Daley DTA, Luscombe NM, Berman HM, Thornton JM (2001) Protein–RNA interactions: a structural analysis. Nucleic Acids Res 29:943–954

Jones S, Shanahan HP, Berman HM, Thornton JM (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res 31:7189–7198

Joshi T, Xu D (2007) Quantitative assessment of relationship between sequence similarity and function similarity. BMC Genomics 8:222

Kahraman A., Morris RJ, Laskowski RA, Thornton JM (2007) Shape variation in protein binding pockets and their ligands. J Mol Biol 368:283–301

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30

Kawabata T, Go N (2007) Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. Proteins 68:516–529

Keil M, Exner TE, Brickmann J (2004) Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a neural network. J Comput Chem 25:779–789

Kinoshita K, Nakamura H (2003) Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. Protein Sci 12:1589–1595

Kleywegt GJ (1999) Recognition of spatial motifs in protein structures. J Mol Biol 285:1887–1897

Koike A, Takagi T (2004) Prediction of protein–protein interaction sites using support vector machines. Protein Eng Des Sel 17:165–173

Kolodny R, Koehl P, Levitt M (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J Mol Biol 346:1173–1188

Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Cryst D60:2256–2268

Kuznetsov IB, Gou Z, Li R, Hwang S (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins 64:19–27

Lackner P, Koppensteiner WA, Sippl MJ, Domingues FS (2000) ProSup: a refined tool for protein structure alignment. Protein Eng 13:745–752

Landgraf R, Xenarios I, Eisenberg D, (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 307:1487–1502

Laskowski RA (1995) SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions. J Mol Graph 13:323–330

Laskowski RA, Luscombe NM, Swindells MB, Thornton JM (1996) Protein clefts in molecular recognition and function. Protein Sci 5:2438–2452

Laskowski RA, Watson JD, Thornton JM (2003) From protein structure to biochemical function? J Struct Func Genomics 4:167–177

Laskowski RA, Watson JD, Thornton JM (2005) Protein function prediction using local 3D templates. J Mol Biol 351:614–626

Laurie AT, Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. Bioinformatics 21:1908–1916

Leibowitz N, Fligelman ZY, Nussinov R, Wolfson HJ (2001) Automatic multiple structure alignment and detection of a common substructural motif. Proteins 43:235–245

Li FM, Li QZ (2008) Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids 34:119–125

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

Liang MP, Banatao DR, Klein TE, Brutlag DL, Altman RB (2003a) WebFEATURE: an interactive web tool for identifying and visualizing functional sites on macromolecular structures. Nucleic Acids Res 31:3324–3327

Liang MP, Brutlag DL, Altman RB (2003b) Automatic construction of structural motifs for predicting functional sites on protein structures. Pac Symp Biocomput 8:204–215

Liang S, Zhang C, Liu S, Zhou Y (2006) Protein binding site prediction using an empirical scoring function. Nucleic Acids Res 34:3698–3707

Lichtarge O, Sowa ME (2002) Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 12:21–27

Liu ZP, Wu LY, Wang Y, Chen L, Zhang XS (2007a) Predicting gene ontology functions from protein's regional surface structures. BMC Bioinformatics 8:475

Liu ZP, Wu LY, Wang Y, Zhang XS, Chen L (2007b) An approach for clustering protein pockets into similar groups. In: Optimization and systems biology. Lecture Notes in Operations Research, vol 7. World Publishing, Beijing, pp 204–212

Liu ZP, Wu LY, Wang Y, Zhang XS, Chen L (2008) Analysis of protein surface patterns by pocket similarity network. Protein Pept Lett (in press)

Luscombe NM, Thornton JM (2002) Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J Mol Biol 320:991–1009

Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) An overview of the structures of protein–DNA complexes. Genome Biol 1:1–37

Luscombe NM, Laskowski RA, Thornton JM (2001) Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. Nucleic Acids Res 29:2860–2874

Ma B, Elkayam T, Wolfon H, Nussinov R (2003) Protein–protein interaction: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci USA 100:5772–5777

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. Nature 402:83–86

McLaughlin WA, Berman HM (2003) Statistical models for discerning protein structures containing the DNA-binding helix-turn helix motif. J Mol Biol 330:43–55

Morris RJ, Najmanovich RJ, Kahraman A, Thornton JM (2005) Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. Bioinformatics 21:2347–2355

Murzin A, Brenner S, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540

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

Neuvirth H, Raz R, Schreiber G (2004) ProMate: a structure based prediction program to identify the location of protein–protein binding sites. J Mol Biol 338:181–199

Ofran Y, Rost B (2003) Predicted protein–protein interaction sites from local sequence information. FEBS Lett 544:236–239

Orengo C, Michie A, Jones S, Jones D, Swindells M, Thornton J (1997) CATH—a hierarchic classification of protein domain structures. Structure 5:1093–1108

Orengo CA, Taylor WR (1996) SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol 266:617–635

Orengo CA, Todd AE, Thornton JM (1999) From protein structure to function. Curr Opin Struct Biol 9:374–382

Pal D, Eisenberg D (2005) Inference of protein function from protein structure. Structure 13:121–130

Panchenko AR, Kondrashov F, Bryant S (2004) Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci 13:884–892

Passerini A, Punta M, Ceroni A, Rost B, Frasconi P (2006) Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. Proteins 65:305–316

Pazos F, Sternberg MJE (2004) Automatic prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci USA 101:14754–14759

Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32:D129–133

Rosen M, Lin SL, Wolfson H, Nussinov R (1998) Molecular shape comparisons in searches for active sites and functional similarity. Protein Eng 11:263–277

Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12:85–94

Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M, Mewes HW (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 18:5539–5545

Russell RB (1998) Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 279:1211–1227

Russell RB, Alber F, Aloy P, Davis FP, Korkin D, Pichaud M, Topf M, Sali A (2004) A structural perspective on protein–protein interactions. Curr Opin Struct Biol 14:313–324

Salwinski L, Eisenberg D (2003) Computational methods of analysis of protein–protein interactions. Curr Opin Struct Biol 13:377–382

Sanishvili R, Yakunin AF, Laskowski RA, Skarina T, Evdokimova E, Doherty-Kirby A, Lajoie G A, Thornton JM, Arrowsmith CH, Savchenko A, Joachimiak A, Edwards AM (2003) Integrating

structure, bioinformatics, and enzymology to discover function—BioH, a new carboxylesterase from *Escherichia coli*. J Biol Chem 278:26039–26045
Schmitt S, Kuhn D, Klebe G (2002) A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol 323:387–406
Schnell JR, Chou JJ (2008) Structure and mechanism of the M2 proton channel of influenza A virus. Nature 451:591–595
Shah I, Hunter L (1997) Predicting enzyme function from sequence: a systematic appraisal. Proc Int Conf Intell Syst Mol Biol 5:276–283
Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Mol Syst Biol 3:88
Shen HB, Chou KC (2007a) EzyPred: a top–down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun 364:53–59
Shen HB, Chou KC (2007b) Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng Des Sel 20:561–567
Shen HB, Chou KC (2008) PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 373:386–388
Shi JY, Zhang SW, Pan Q and Zhou GP (2008) Using pseudo amino acid composition to predict protein subcellular location: approached with amino acid composition distribution. Amino Acids. doi:10.1007/s00726-007-0623-z
Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11:739–747
Shulman-Peleg A, Nussinov R, Wolfson HJ (2005) SiteEngines: recognition and comparison of binding sites and protein–protein interfaces. Nucleic Acids Res 33:W337–W341
Siggers TW, Silkov A, Honig B (2005) Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. J Mol Biol 345:1027–1045
Singh AP, Brutlag DL (1997) Hierarchical protein structure alignment using both secondary structure and atomic representations. Proc Intell Syst Mol Biol 4:284–293
Singh R, Saha M (2003) Identifying structural motifs in proteins. Pac Symp Biocomput 8:228–239
Sodhi JS, Bryson K, McGuffin LJ, Ward JJ, Wernisch L, Jones DT (2004) Predicting metal-binding site residues in low-resolution structural models. J Mol Biol 342:307–320
Stark A, Russell RB (2003a) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. Nucleic Acids Res 31:3341–3344
Stark A, Shkumatov A, Russell RB (2003b) Finding functional sites in structural genomics proteins. Structure 12:1405–1412
Stark A, Sunyaev S, Russell R (2003c) A model for statistical significance of local similarities in structure. J Mol Biol 326:1307–1316
Stawiski EW, Gregoret LM, Mandel-Gutfreund Y (2003) Annotating nucleic acid-binding function based on protein structure. J Mol Biol 326:1065–1079
Taroni C, Jones S, Thornton JM (2000) Analysis and prediction of carbohydrate binding sites. Protein Eng 13:89–98
The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. Nature Genet 25:25–29
Tjong H, Zhou HX (2007) DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. Nucleic Acids Res 35:1465–1477
Torrance JW, Bartlett GJ, Porter CT, Thornton JM (2005) Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. J Mol Biol 347:565–581
Tseng YY, Liang J (2006) Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: A Bayesian Monte Carlo approach. Mol Biol Evol 23:421–436
Tsuchiya Y, Kinoshita K, Nakamura H (2004) Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. Proteins 55:885–894
Vapnik V (1998) Statistical learning theory. Springer, New York
Vazquez A, Flammini A, Maritan A, Vespignani A (2003) Global protein function prediction from protein–protein interaction networks. Nat Biotechnol 21:697–700
Wallace AC, Borkakoti N, Thornton JM (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural database. Application to enzyme active sites. Protein Sci 6:2308–2323
Wang SQ, Du QS, Zhao K, Li AX, Wei DQ, Chou KC (2007a) Virtual screening for finding natural inhibitor against cathepsin-L for SARS therapy. Amino Acids 33:129–135
Wang H, Segal E, Ben-Hur A, Li Q, Vidal M, Koller D (2007b) InSite: a computational method for identifying protein–protein interaction binding sites on a proteome-wide scale. Genome Biol 8:R192
Wangikar PP, Tendulkar AV, Ramya S, Mali DN, Sarawagi S (2003) Functional sites in protein families uncovered via an objective and automatic graph theoretic approach. J Mol Biol 326:955–978
Watson JD, Laskowski RA, Thornton JM (2005) Predicting protein function from sequence and structural data. Curr Opin Struct Biol 15:275–284
Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. Q Rev Biophys 36:307–340
Wilson CA, Kreychman J, Gerstein M (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function throng traditional and probabilistic scores. J Mol Biol 297:233–249
Wodak SJ, Mendez R (2004) Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. Curr Opin Struct Biol 14:242–249
Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28:57–61
Yan C, Dobbs D, Honavar V (2004) A two-stage classifier for identification of protein–protein interface residues. Bioinformatics 20[Suppl]:i371–i378
Yan C, Terribilini M, Wu F, Jernigan RL, Dobbs D, Honavar V (2006) Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinformatics 7:262
Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, Lichtarge O (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. J Mol Biol 326:255–261
Ye Y, Godzik A (2004) FATCAT: a web server for flexible structure comparison and structure similarity searching. Nucleic Acids Res 32:W582–585
Zemla A (2003) LGA—a method for finding 3D similarities in protein structures, Nucleic Acids Res 31:3370–3374
Zhang XS (2000) Neural networks in optimization. Kluwer, Dordrecht
Zhang Z, Grigorov MG (2006) Similarity networks of protein binding sites. Proteins 62:470–478
Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on TM-score. Nucleic Acids Res 33:2302–2309
Zhang S, Jin G, Zhang XS, Chen L (2007) Discovering functions and revealing mechanisms at molecular level from biological networks. Proteomics 7:2856–2869
Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006) Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30:461–468

Zhao XM, Wang Y, Chen L, Aihara K (2008a) Gene function prediction using labeled and unlabeled data. BMC Bioinformatics 9:57

Zhao XM, Wang Y, Chen L, Aihara K (2008b) Protein domain annotation with integration of heterogeneous information sources. Proteins. doi:10.1002/prot.21943

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17:729–738

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins 44:57–59

Zhou GP, Cai YD (2006) Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. Proteins 63:681–684

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins 50:44–48

Zhou HX, Qin S (2007) Interaction-site prediction for protein complexes: a critical assessment. Bioinformatics 23:2203–2209

Zhou HX, Shan Y (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 44:336–343

Zhou XB, Chen C, Li ZC and Zou XY (2007a) Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine. Amino Acids. doi:10.1007/s00726-007-0608-y

Zhou XB, Chen C, Li ZC, Zou XY (2007b) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 248:546–551

Zhu J, Weng Z (2005) FAST: a novel protein structure alignment algorithm. Proteins 58:618–627