Chapter 11
3D Motifs

Jerome P. Nilmeier, Elaine C. Meng, Benjamin J. Polacco and Patricia C. Babbitt

Abstract Three-dimensional (3D) motifs are patterns of local structure associated with function, typically based on residues in binding or catalytic sites. Protein structures of unknown function can be annotated by comparing them to known 3D motifs. Many methods have been developed for identifying 3D motifs and for searching structures for their occurrence. Approaches vary in the type and amount of input evidence, how the motifs are described and matched, whether the results include a measure of statistical significance, and how the motifs relate to function. Compared to algorithm development, less progress has been made in providing publicly searchable databases of 3D motifs that are both functionally specific and cover a broad range of functions. A roadblock has been the difficulty of generating detailed structure-function classifications; instead, automated, large-scale studies have relied upon pre-existing classifications of either structure or function. Complementary to 3D motif methods are approaches focused on molecular surface descriptions, global structure (fold) comparisons, predicting interactions with other macromolecules, and identifying physiological substrates by docking databases of small molecules.

J.P. Nilmeier (✉)
Lawrence Livermore National Laboratory (LLNL) Division of Physical and Life Sciences Directorate, Biotechnology and Biosciences Division, 7000 East Avenue, Livermore, CA 94550-9234, USA
e-mail: nilmeier1@llnl.gov

E.C. Meng · B.J. Polacco · P.C. Babbitt
Department of Pharmaceutical Chemistry, University of California San Francisco (UCSF), 600 16th Street, San Francisco, CA 94158-2517, USA
e-mail: meng@cgl.ucsf.edu

B.J. Polacco
e-mail: benjamin.polacco@ucsf.edu

P.C. Babbitt
e-mail: babbitt@cgl.ucsf.edu

P.C. Babbitt
UCSF Department of Biopharmaceutical Sciences, 1700 4th Street, San Francisco, CA 94158-2330, USA

D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_11

Keywords Clique detection · Geometric hashing · Functional annotation · Function prediction · Active site · Binding site · Functional residues · Catalytic residues · Structural motifs · Pattern discovery

List of Abbreviations

3D	Three-dimensional
CSA	Catalytic Site Atlas
DRESPAT	Detection of REcurring Sidechain PATterns
EC	Enzyme Commission
FFF	Fuzzy Functional Form
GASPS	Genetic Algorithm Search for Patterns in Structures
GO	Gene Ontology
HMM	Hidden Markov Model
nr-PDB	Non-redundant PDB
NP	Nonpolynomial (scaling)
NOE	Nuclear Overhauser Effect
PAR-3D	Protein Active site Residues using 3-Dimensional structural motifs
PDB	Protein Data Bank
PINTS	Patterns in Non-homologous Tertiary Structures
RMSD	Root-mean-square Deviation
S-BLEST	Structure-Based Local Environment Search Tool
SCOP	Structural Classification of Proteins
SOIPPA	Sequence Order-Independent Profile-Profile Alignment
SPASM	SPatial Arrangements of Sidechains and Mainchains
TESS	TEmplate Search and Superposition

11.1 Background: Functional Annotation

The genomic approach to biology has resulted not only in copious amounts of new sequence and structure data, but also the prospect of obtaining a complete "parts list" for many organisms. However, a parts list is of little use without some understanding of what each part does. Even with entire genome sequences in hand, not all genes have been identified, and among identified genes, significant numbers have not been annotated with any function. The amount of sequence data far outweighs the available structures, so to a large extent, the assignment of functions, or ***functional annotation***, has been performed by large-scale sequence searching. In many cases, the function of an unknown sequence is inferred, or ***transferred***, through similarity to a sequence with a known function.

11.1.1 What Is Function?

Function can be described at many levels and from many perspectives (Radivojac et al. 2013). Objective classifications of function are needed for training and testing any method of functional annotation. The Gene Ontology (GO) system (Ashburner et al. 2000) is a hierarchical set of functional descriptors ranging from broad to specific in each of three categories: biological process, cellular component, and molecular function. For the specific molecular functions of enzymes, GO embeds the Enzyme Commission (EC) system (International Union of Biochemistry and Molecular Biology: Nomenclature Committee and Webb 1992) which is also hierarchical: catalysed reactions are described with four integers, where the first number refers to a broad class of reactions and the last number refers to a specific substrate. GO also includes molecular function terms for stable binding relationships (where binding is not functionally associated with membrane transport or catalytic activity). The KEGG annotation (Kanehisa and Goto 2000; Ogata et al. 1999), while used mostly for studying reaction pathways, can also be used to annotate enzyme function.

Other methods for classifying proteins, while less directly related to function, can be used to infer relationships related to function. These include Structural Classification of Proteins (SCOP) (Murzin et al. 1995; Conte et al. 2000; Andreeva et al. 2004, 2008) and Class, Architecture, Topology, and Homologous superfamily (CATH) (Orengo et al. 1997, 1999, 2003). Both methods are hierarchical classifications of protein substructures such as ***folds*** (Richardson 1981) or ***domains*** (Chothia and Lesk 1986; Rost 1997), that can be "mixed and matched" evolutionarily (Chothia et al. 2003). In SCOP, domains are classified into families, superfamilies, folds, and classes. Folds are, in general, only indirectly related to function (Babbitt and Gerlt 1997; Todd et al. 2001), but they can be very informative for many cases. The use of fold similarity for annotation transfer is discussed in Chap. 9.

The GO and EC annotations for functional annotation cover nearly all reactions found in biochemical systems. They do not, however, include details on enzymatic mechanism, or the role of the protein in the reaction (Babbitt 2003). Two enzymes that catalyze the same overall reaction would have the same EC number, even if their structures and catalytic intermediates are very different. Additionally, many enzymes are evolutionarily related because they share an intermediate step in the overall reaction, that is, a ***common partial reaction***. The EC and GO naming systems do not account for such similarities in any practical way, and yet such similarities are a defining feature for many protein superfamilies, with the enolase superfamily as the most notable example. Figure 11.1 illustrates the variety of reactions associated with the enolase superfamily.

11.1.2 Genomics and Functional Annotation

The progress in the genomics community in assigning functional annotations through sequence-based methods is impressive. Given that function is related

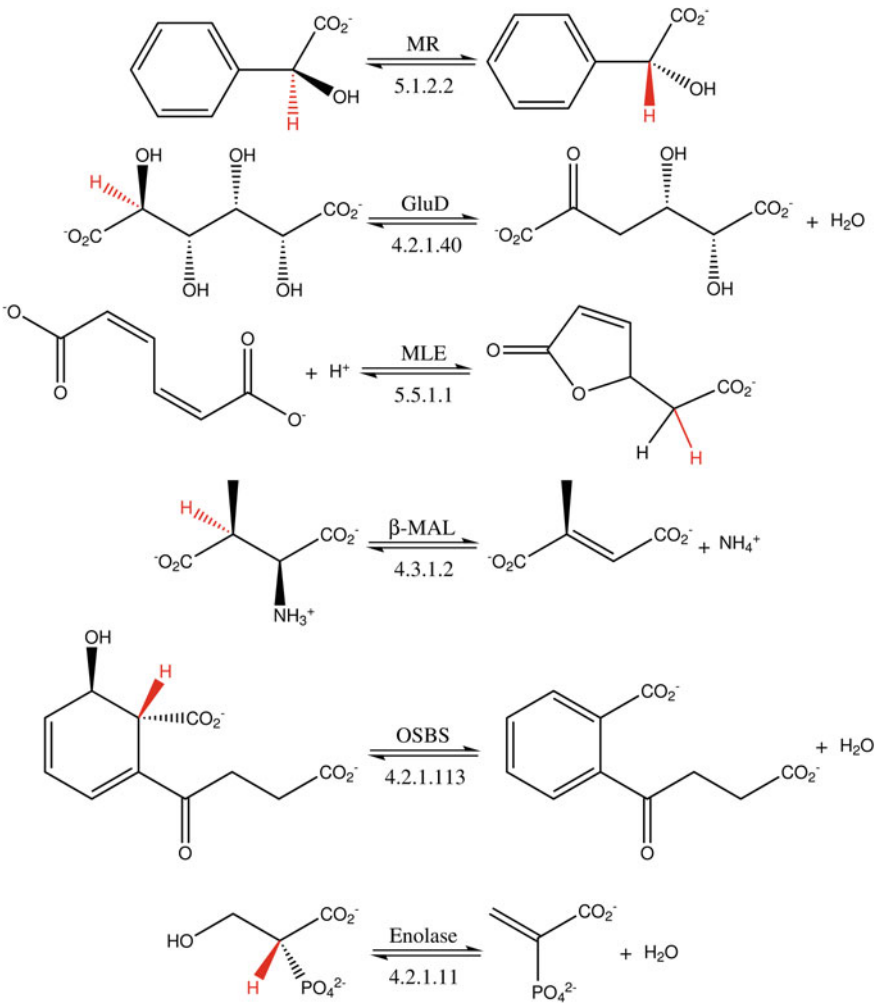


Fig. 11.1 Illustration of the common partial reaction in the enolase superfamily. The extraordinary diversity of reactions shown in these enzymes share one step in common, which is the initial abstraction of a proton (indicated in *red*). Abbreviations are *MR* mandelate racemase, *GlucD* glucarate dehydratase, *MLE* muconate lactonizing enzyme, *β-MAL* β methylaspartate ammonia lyase, *OSBS* O-succinylbenzoate synthase

indirectly to sequence through a protein structure, however, it makes sense to consider methods that incorporate protein structure more directly in the inference of function.

Sequence alignment methods such as BLAST (Altschul et al. 1990) and CLUSTALW (Larkin et al. 2007; Thompson et al. 1994) have enjoyed wide success in inferring function when sequence similarity is greater than 40–60% (Tian and Skolnick 2003; Devos and Valencia 2001; Rost 2002). More sophisticated

methods, including Hidden Markov Model (HMM) methods (Krogh et al. 1994; Sjölander et al. 1996), and ancestry-based methods such as the Evolutionary Trace (Lichtarge et al. 1996), INTREPID (Sankararaman and Sjölander 2008), Phylofacts (Glanville et al. 2007; Krishnamurthy et al. 2006), Bayesian Monte Carlo inference from phylogenetic trees (Tseng and Liang 2006) and EFICAz (Arakaki et al. 2009; Tian et al. 2004) combine sequence alignment procedures and machine learning techniques to specifically assign function to a sequence.

11.1.3 The Need for Structure-Based Methods

Protein structures, however, may reveal important similarities or possible evolutionary relationships that are not evident from their sequences alone. The natural analogue to a global sequence alignment is a global structure alignment. Methods like LGA (Zemla 2003), PINTS (Stark and Russell 2003) and CE (Shindyalov and Bourne 1998, 2001) can accomplish this alignment in various ways and sometimes reveal more significant relationships in the alignments.

Other approaches use combinations of sequence and structural information, such as SOIPPA (Xie and Bourne 2008, 2009; Ren et al. 2010), DISCERN (Sankararaman et al. 2010), and PevoSOAR (Tseng et al. 2009), and can provide improvements to sequence based methods alone. Additionally, methods like the FFF approach that are essentially structural in nature benefit from addition of sequence information (Cammer et al. 2003). The success of any of these global similarity-based techniques depends largely on the ability to distinguish conservation patterns that correspond to the actual functional or catalytic portions of a protein sequence or structure.

Related proteins may have diverged so far that global sequence or structure alignments are challenging. Conversely, proteins with highly similar folds can perform different functions (Babbitt and Gerlt 1997; Todd et al. 2001). This observation points to the need for a more fundamental definition of a structural unit, or ***3D motif*** which more specifically defines the functional aspects of a given protein structure.

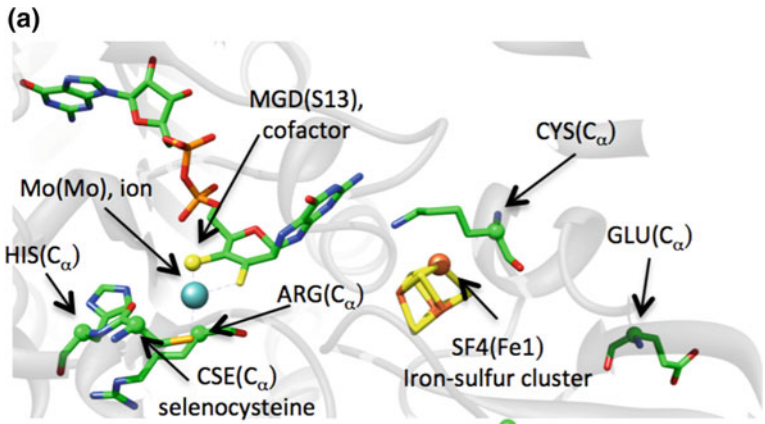
Structural genomics efforts have long recognized the fact that structural data is much more informative than sequence data alone. This data is used not only for annotation, but for homology modelling and in silico drug design. On principal driving idea behind this effort is to crystallize structures that are underrepresented in sequence space, so that more sequences can be more directly represented in structural forms (Berman et al. 2000; Baker and Sali 2001). The number of structures in the PDB from these initiatives has continued to grow at an increasing rate, and many target structures were previously completely unannotated, or annotated incorrectly using automated sequence-based methods.

Functional assignment to these proteins remains as a frontier challenge for structural genomics, and 3D motif-based methods are likely to play a prominent role for proteins where current methods fall short.

11.2 3D Motif Matching Techniques

11.2.1 What Is a 3D Motif?

3D motifs are spatial patterns of points based on a few residues (generally under a dozen) associated with some protein function or classification of interest. They are sometimes called ***active site templates***, since the residues may contribute to a



	ARG	CYS	GLU	HIS	LYS	MGD	SF4
ARG 333.A							
CYS 140.A	5.54						
GLU 30.A	24.15	27.56					
HIS 141.A	6.37	3.76	30.18				
LYS 44.A	17.05	19.15	11.22	22.21			
MGD 802.A	5.80	4.24	25.04	7.11	15.79		
MO 803.A	6.57	3.72	25.05	7.38	15.97	2.47	
SF4 800.A	13.65	17.00	11.71	19.50	5.15	13.93	14.51

Fig. 11.2 Example of a catalytic template constructed from a Catalytic Site Atlas (CSA) entry, which has a corresponding EC number along with a list of residues that comprise the site. Each residue has a centroid associated with it, which is labelled in parentheses and shown as spheres in (**a**) and (**b**). Cofactors, ions, and residues can often have either a single centroid or many centroids associated with them (see Fig. 11.3). In this example, Cα coordinates are used as the residue centroids, but centroids may be computed in other ways. For this templating approach, a graphical representation of the template is used, with nodes associated with the centroid identity, and edges defined by the interatomic distances. The template is stored as a distance matrix, shown in (**b**). The image was created with UCSF Chimera (Pettersen et al. 2004) (http://www.cgl.ucsf.edu/chimera)

binding or catalytic pocket, or ***structural templates***. The positions of one or a few atoms per residue are used, and the points are labelled with additional information, such as atom and residue type, used in matching. The residues are often strictly positioned in space but not necessarily in sequence. Figure 11.2 describes a typical binding site found in the Catalytic Site Atlas (CSA), and one way to represent it in a reduced form. In this example, the Cα atoms are used as pseudoatoms, but many approaches use atomic coordinates from the sidechains, or a centroid using clusters of atoms in the pseudoatom positions as well (Oldfield 2002), as is the case for the templates in Fig. 11.3.

3D motifs represent highly conserved patterns of local structure. Often the residues are conserved to sub-angstrom resolution, and the absence of one residue in the motif can completely eliminate its function. The remainder of the protein, however, can often vary substantially. Ideally, a 3D motif will describe exactly these function-critical structural components and serve as a sensitive and specific signature of the function.

Since such a motif can often be the only evolutionary constraint, many different structures can be present with the same motif, and there is no restriction on the

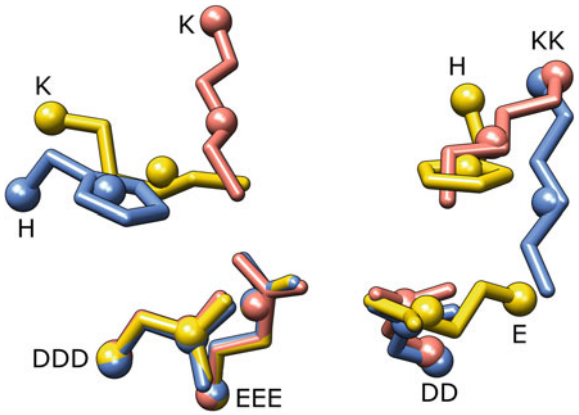


Fig. 11.3 Active site residues from members of the enolase superfamily, illustrating aspects of motif representation and specificity. The superimposed side chains of two basic and three acidic residues are shown from each of the following: mandelate racemase (*yellow*, PDB 2mnr), enolase (*salmon*, PDB 4enl), and methylaspartate ammonia lyase (*blue*, PDB 1kcz). Balls indicate alpha-carbon (Cα) and side chain centroid locations. Single-letter codes near the alpha-carbons indicate residue types: *H* for histidine, *K* for lysine, *D* for aspartic acid, and *E* for glutamic acid. While the acidic residues at the two lower left positions are highly conserved in type and conformation, variations in the sites include: *1* differing (albeit similar) residue types at the other three positions; *2* different side chain conformations, exemplified by the two lysines on the right; *3* different locations in primary sequence, where the basic residue on the upper left is C-terminal to the others in enolase but N-terminal in the sequences of the other two proteins. Using side chain centroids rather than the positions of functional atoms generally allows for more variety in backbone conformations, assuming the sidechain positions are well conserved across templates (Todd et al. 2002). The image was created with UCSF Chimera (Pettersen et al. 2004) (http://www.cgl.ucsf.edu/chimera)

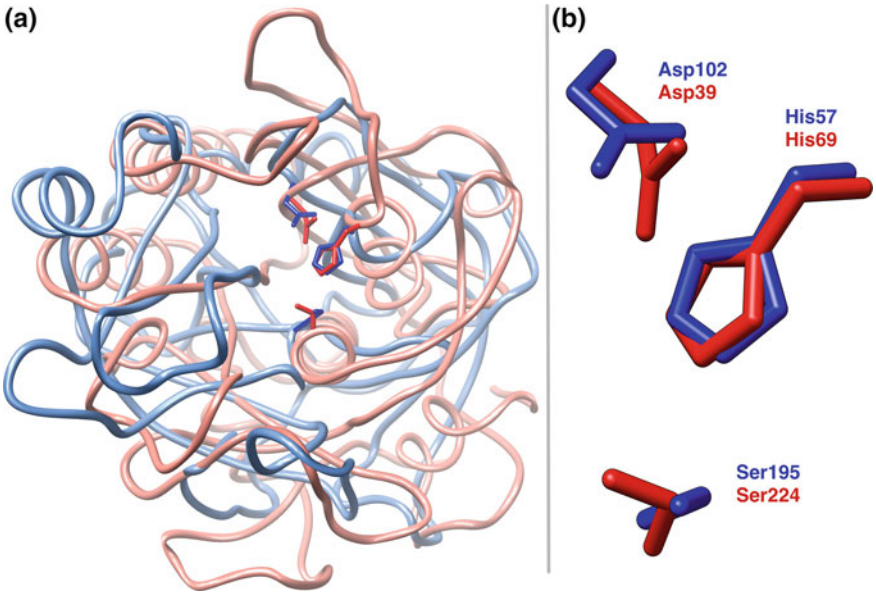
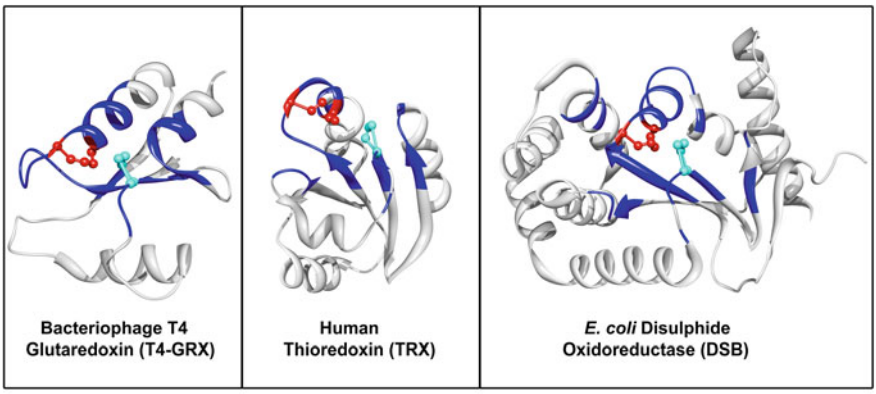


Fig. 11.4 Two serine proteases superimposed at their catalytic triads reveals the close similarity of residues in the active sites despite different overall folds. **a** Ribbon diagrams of trypsin (*blue*/*light blue*, PDB 1sgt) and proteinase *K*, a homolog of subtilisin, (*red*/*salmon*, PDB 2pkc) show that the two proteins have different folds with no corresponding secondary structure elements, yet their catalytic triads (displayed in stick representation) overlap. They are considered to have no common ancestor. **b** The sidechains of the catalytic triads are shown enlarged to display the similar orientations of the catalytic triad residues (1sgt: Asp102, His57, Ser195; and 2pkc: Asp39, His69, Ser224). The similarity of the catalytic triad in these non-homologous structures demonstrates the ability of 3D motifs to detect similar functions in a pair of proteins where homology-based methods will fail. The image was created with UCSF Chimera (Pettersen et al. 2004) (http://www.cgl.ucsf.edu/chimera)

location or relative order of residues in the sequence. Figure 11.4 shows a case of convergent evolution in the serine protease Asp-His-Ser catalytic triad. While the catalytic triad is highly conserved structurally, the remaining structural elements display noticeable variations. This particular catalytic triad was, historically, the first to be thought of as a 'motif' based on these observations. Variations in structure relative to a motif are even more pronounced in other more recent examples, including the disulfide oxidoreductase site shown in Fig. 11.5, which is taken from an example of a Fuzzy Functional Form (FFF) template (Fetrow and Skolnick 1998; Di Gennaro et al. 2001).




```
T4-GRX  ...................MFKVYGYDSNIHKCVYCDNAKRLLTVKKQPF....EFINIMPEKGVFDDEKIAEL
DSB     AQYEDGKQYTTLEKPVAGAPQVLEFFSFFCPHCYQFEEVLHISDNVKKKLPEGVKMTKYHVNFMGGDLGKDLTQ
TRX     KQIESKTAFQEALDAAGDKLVVVDFSATWCGPCKMIKPFFHSLSE...KYSN.VIFLEVDVD...........D

T4-GRX  LTKLGRDTQIGLTMPQVFAPDGSHIGGFDQLREYFK.....................................
DSB     AWAVAMALGVEDKVTVPLFEGVQKTQTIRSASDIRDVFINAGIKGEEYDAAWNSFVVKSLVAQQEKAAADVQLR
TRX     CQDVASECEVKCTPTFQFFKKGQKVGE..........FSGA.NKEKLEATINELV..................

T4-GRX  ........................................
DSB     GVPAMFVNGKYQLNPQGMDTSNMDVFVQQYADTVKYLSEKK
TRX     ........................................
```

Fig. 11.5 The FFF motif for the disulfide oxidoreductase active site is found in many proteins. Illustrated are T4 glutaredoxin, 1aaz, chain A (*left*), human thioredoxin, 4trx (*middle*) and proline disulfide isomerase, 1dsb, chain A. The three key residues which define this FFF are two cysteines (*red side* chains) and a proline (*cyan side* chain). The active site structure of these proteins is conserved, although the rest of the protein structures exhibit some differences. Using these three key residues, the active site signature for each protein was identified (fragments shown as *blue ribbons* in each protein). Global sequence alignment, produced using ClustalW, of these three proteins shows the location of the key residues (*red* and *cyan*, underlined) and the active site signature fragments (*blue*) within the whole sequence. The alignment illustrates the lack of overall sequence similarity between the three proteins, even though the active site structure itself is highly conserved

11.2.2 Historical Development of Motif Matching Methods

Early ideas about catalytic motifs were based on observation, and were not algorithmic in nature. The most widely studied motif is the Ser-His-Asp catalytic triad mentioned above, first recognized in serine proteases (Blow et al. 1969; Wright et al. 1969) and later in other hydrolases such as esterases and lipases. The catalytic triad occurs in different folds, and thus it encompasses cases of both divergent and convergent evolution (Fig. 11.4). Early discoveries of the catalytic triad found it present in entirely different folds of subtilisins, (Fischer et al. 1994). The Thornton group, studying triads in detail, formulated a more careful description of the site, based on the observation that only the relative positions of serine and aspartate

oxygens and the histidine ring were preserved across many examples (Wallace et al. 1996).

During this time, the concept of a 3D motif began to emerge in an algorithmic context, which is generally described as ***template matching*** or ***motif matching***.

Artymiuk et al. (1994) appear to be the first to apply such a procedure, which they called ASSAM, to enzymatic site detection. Their work used the ***subgraph isomorphism*** procedure, which is a graph theoretic method for finding a motif graph in a larger structure graph. The method, originally proposed by Ullmann (1976), is described in Sect. 11.2.1. Later work by Artymiuk et al. expanded the approach beyond catalytic sites to other structural applications, such as the identification of tertiary structures (Mitchell et al. 1990; Spriggs et al. 2003). In this work, many careful choices were made with regard to which atoms to use as part of the template, and particular attention was paid to reliable detection of residue triads, given the importance of catalytic triads as an archetypal motif.

During this period, Kleywegt also developed a site-matching procedure originally designed to identify patterns in distance matrices determined by Nuclear Overhauser Effect (Radivojac et al.) measurements (Kleywegt et al. 1989). Later Kleywegt introduced a program called DEJAVU that detects protein motifs (Kleywegt and Jones 1997). A technique based on DEJAVU was later generalized to identify enzymatic sites with a method called SPASM, along with a complementary approach, known as RIGOR (Kleywegt 1999), used to search a list of motifs for similarity to a given structure. Early work with this method focused on triad motifs as well. A notable example from the Kleywegt study (Kleywegt 1999) was the discovery of a family of glucanases.

A related set of approaches to the template matching problem uses a procedure known as ***geometric hashing*** (Wolfson and Rigoutsos 1997; Brakoulias and Jackson 2004). The main difference between the geometric hashing procedure and graph-based procedures is that geometric hashing uses a Cartesian grid (with a suitable coordinate system) to bin similar coordinates. It is used widely in image processing, and has been successfully adapted to structural approaches. It is dependent on the frame of reference, however, and additional overhead is required to accomplish optimal translations and rotations for comparison. The Thornton group proposed a template-matching procedure, named TESS (Wallace et al. 1997), built on such an approach. A later iteration, known as JESS (Barker and Thornton 2003), incorporated recursive ideas and threshold constraints to improve searching procedures. More recently, the Kavraki group developed a series of procedures built on a match augmentation method, MASH, that iteratively grows a template match from pairwise matches obtained through geometric hashing (Chen et al. 2007a). Later developments from this group include the addition of residue hash matching, the LabelHash algorithm (Moll et al. 2010; Moll and Kavraki 2008), along with impressive optimizations at the hardware and software level to improve performance. Other geometric hashing approaches include SitesBase (Gold and Jackson 2006a, b), and GIRAF (Kinjo and Nakamura 2007).

Success of template-matching methods, within the Thornton group and elsewhere, led to the important recognition that a high quality curated database of enzymatic sites was needed. This recognition led directly to the development of the Catalytic Site Atlas (CSA) (Porter et al. 2004), which is a manually curated table of enzymes and binding site residues, as well as tabulated Enzyme Commission (EC) numbers (Bairoch 1994). The CSA is somewhat limited in coverage, however, and the scale of such a database will always be strictly limited by the capacity of expert manual curators. As a result, many approaches have been developed which attempt to automatically locate structural features that may be used as templates. These approaches include physics-based approaches (Halgren 2007, 2009) and statistical modelling of measures (Liang et al. 2003; Brylinski and Skolnick 2008; Skolnick and Brylinski 2009). Methods that consider protein dynamics (Yang and Bahar 2005; Glazer et al. 2008) represent a promising direction as computational capabilities improve (see also Chap. 12).

Other valuable resources related to this effort, including the MACiE database (Holliday et al. 2007), and the ProFunc server (Laskowski et al. 2005), as well as metaservers like ProKnow (Whisstock and Lesk 2003), resulted from the success and utility of structure-based approaches to understanding function. Table 11.1 lists some of the database resources that have resulted from efforts in this field.

The Babbitt and Gerlt groups have gone beyond matching of catalytic residues and matched enzymes by their chemical mechanism. They established the concept of a mechanistically diverse superfamily, where the similarity among members is governed by the conservation of partial reactions within the protein family, rather than by sequence or structure conservation alone (Galperin et al. 1998; Gerlt and Babbitt 2001; Gerlt et al. 2012). This approach is in contrast to a sequence-based approach, which relies on global sequence similarity with the expectation that conservation patterns can point to residues of functional interest. It also presents an alternative to the Enzyme Commission (EC) classification scheme (Webb 1992), which builds a hierarchy based on the substrate reaction chemistry. This alternative approach to classification, with its emphasis on binding site architecture and conservation of partial reactivity, led to the development of the Structure-Function Linkage Database SFLD (Pegg et al. 2005, 2006). These ideas led to the larger Enzyme Function Initiative (Gerlt et al. 2011), which has the goal of large-scale enzyme characterization and classification based on experimental and computational work (Gerlt et al. 2012; Kalyanaraman et al. 2008; Song et al. 2007). Template-matching procedures using superfamily template libraries were applied (Meng et al. 2004), and led to a procedure known as GASPS and the database GASPSdb. GASPS is designed to develop new template libraries based on any classification of structures into those with and without a function (or other property) of interest (Polacco and Babbitt 2006).

Table 11.1 Servers and other web resources for 3D motif searching and comparison

Server name and citation	Server URL
Description of resource	Motif database description
Catalytic site atlas (CSA) (Furnham et al. 2014)	http://www.ebi.ac.uk/thornton-srv/databases/CSA/
Basic interface to motif database (CSA)	Cα and Cβ functional atom motifs for 147 well-characterized enzyme families. Database freely available for download
ProFunc (Laskowski et al. 2005)	http://www.ebi.ac.uk/thornton-srv/databases/profunc/
Multi-search including motif search with JESS: whole structure query vs. motif database, fragment query versus whole chains	CSA motifs, 13,057 ligand-binding and 1200 DNA-binding modes from PDB. Motifs contain both sidechain and backbone atoms
Catalytic site identification (Kirshner et al. 2013)	http://catsid.llnl.gov/
Finds matches to motifs with user defined target and/or protein databank. Uses subgraph isomorphism and machine learning	2244 motifs, including modified CSA and enolase superfamily templates. User can also search for unannotated structures by EC number
Uppsala Software Factory (Kleywegt 1999)	http://xray.bmc.uu.se/usf/
Software is available for download. SPASM compares a query motif to a database of targets. RIGOR compares a query structure to a database of motifs	RIGOR database contains 73,164 motifs from PDB. 57,719 motifs have residue type labels. The remaining are unlabelled (engineerable)
ProKnow (Pal and Eisenberg 2005)	http://proknow.mbi.ucla.edu/
Multi-search, including RIGOR motif searches. GO annotations included in output	10,230 motifs with GO annotations from their source structures, 7819 if electronic annotations are excluded
GASPSdb (Polacco and Babbitt 2006)	http://gaspsdb.rbvi.ucsf.edu/
Browse database of 3D motifs representing SCOP families and superfamilies	Motifs have Cα and side chain coordinates. RIGOR-formatted database files are available for download
funClust (Ausiello et al. 2008)	http://pdbfun.uniroma2.it/funclust/
Uses Query3D to identify motifs shared by groups of 3–20 structures	User supplied structures for consensus motif
pdbFun (Ausiello et al. 2005b)	http://pdbfun.uniroma2.it/
Compares specified probe and target residue sets using Query3D	>12 M individual residues. Subsets are defined with Boolean descriptors combinations
ProBIS (Konc and Janežič 2012)	http://probis.cmm.ki.si/
Detects similar binding sites using a clique detection algorithm (ProBIS)	Database contains pre-calculated matches for non-redundant (95%) pdb
The LabelHash server (Moll et al. 2011)	http://labelhash.kavrakilab.org/
Compares motifs with PDB or user structures using LabelHash algorithm	17 predefined motifs derived from CSA. User defined motifs are allowed
WebFEATURE (Liang et al. 2003; Buturovic et al. 2014)	http://feature.stanford.edu/webfeature/

(continued)

Table 11.1 (continued)

Server name and citation	Server URL
Uses radially symmetric patterns as motifs	Motifs are derived from PROSITE v20.81, and are available for individual download
PAR-3D (Goyal et al. 2007)	http://sunserver.cdfd.org.in:8080/protease/PAR_3D/access.html
Compares query to motifs expressed as distance and angle ranges	Cα and Cβ motifs for 6 protease classes and 10 glycolytic enzymes. Metal chelating sites have sidechain centroids as well
PDBSiteScan (Ivanisenko et al. 2004)	http://wwwmgs.bionet.nsc.ru/mgs/gnw/pdbsitescan/
Compares query to all or a subset of motifs in the PDBSite database	36,273 backbone-atom motifs from SITE annotations. Also includes interfaces with DNA, RNA, or other proteins
PINTS (Stark and Russell 2003)	http://www.russelllab.org/cgi-bin/tools/pints.pl
Compares query structure to motif database, query motif to PDB, or two proteins to each other	Ligand-binding and SITE-annotated motifs consisting of side chain points from polar residues
SuMo (Jambon et al. 2005)	http://sumo-pbil.ibcp.fr/
Compares query structure, chain, or ligand-binding site to database	Database contains 34,210 ligand-binding sites, and also whole structures. Motifs are built from functional groups
S-BLEST (Schmitt et al.) (Mooney et al. 2005)	http://www.sblest.org/
Queries residue-centred patterns against nr-PDB. Returns best-matching chains and annotations	Searches for similarity to uploaded structure only
SiteEngine (Shulman-Peleg et al. 2005)	http://bioinfo3d.cs.tau.ac.il/SiteEngine/
Compares the binding site of a ligand-bound structure to the entire surface region of another structure	Linux executable for non-commercial use only
Nestor3D (Nebel et al. 2007)	http://staffnet.kingston.ac.uk/~ku33185/Nestor3D.html
Generates a consensus motif with input structures and structure alignments	User supplies input structures for comparison. Software is available for download

11.3 Algorithmic Approaches to Motif Matching

The historical development of motif matching methods and current methods suggest the following categorization of these methods.

11.3.1 Methods Using 3D Motifs

Many elements can make up the definition of a motif, but the majority of approaches consider a motif as a constellation of labelled points derived directly from an important subset of atomic coordinates of a structure or set of structures. A side chain centroid, for example, is simply a pseudoatom at the average position of the atoms in the side chain. Up to a few points are used per residue in the motif, and the points are labelled with additional information such as atom type, residue type, or physicochemical characteristics.

Searching can be computationally intensive, especially considering that thousands of structures may be compared to thousands of motifs; 3D motif searching has relied on the development of efficient algorithms, often involving one or more of the following:

- **Geometric hashing**. Hashing is a broad term for reducing complex data to a simpler form that can be compared more rapidly. In its most basic form, a geometric hash can be a lookup table of Cartesian coordinate points (Fischer et al. 1994) and pseudoatom identities as well as many other properties, including distances, angles, and other residue features (Shulman-Peleg et al. 2004). In general, hash comparisons are very fast, especially compared to the time required to align the coordinates (Pennec and Ayache 1998). Hashing or preprocessing the data takes time, but only needs to be done once per structure and can greatly speed up comparisons.
- **Graph Theoretic Methods**. A graph consists of vertices (Kaminski et al. 2001) and edges (lines that connect pairs of vertices). A molecular structure or 3D motif can be treated as a labelled graph. Figure 11.2 shows how a catalytic site might be represented as a group of labelled vertices with interatomic distances used as edges. ***Subgraph isomorphism*** algorithms look for the occurrence of a subgraph (the 3D motif) in a larger graph (the structure). While the subgraph isomorphism is formally treated as a method for identical matches, many modifications to this basic approach are used for imperfect matches, including a variety of distance tolerances, as well as allowances for substitutions (Nilmeier et al. 2013). ***Clique detection*** (Schmitt et al. 2002) is essentially a similar algorithm, but the graph in this case describes the geometries of both structures together. A vertex in the graph represents a pair of atoms or pseudoatoms, one from structure A and one from structure B (where "structure" could be a 3D motif). Only atoms with matching types are allowed to pair. Two vertices are connected by an edge if the distance between the two atoms in A matches the distance between the two atoms in B within a specified tolerance. *A clique is a graph in which every vertex is connected to every other vertex*. Thus, clique detection identifies a set of atoms from A with internal distances completely consistent with those among a paired set of atoms from B.

11.3.2 Efficiency Considerations for 3D Motifs

Motif matching algorithms can be very fast for perfect matches. A challenge in the design of these algorithms, however, is that the extension to imperfect matches can lead to exponential scaling—sometimes referred to as nonpolynomial (Larkin et al. 2007) scaling—with respect to template and structure size, with concomitant losses in speed and efficiency.

To address this challenge variations of branch and bound approaches are used. These approaches leverage combinations of ***breadth-first*** and ***depth-first*** searches, and usually build a series of partial templates for comparison. In template matching algorithms, a breadth-first search typically refers to a method whereby a partially constructed template with few vertices and a 'breadth' of candidate edges are compared for fitness. The best candidates are then selected for the next iteration. Alternatively, a depth-first search builds a 'deeper' partial template with many vertices and fewer edges before iterating to the next comparison step. While described graphically, these ideas can be used in the geometric hashing comparisons as well.

During the buildup procedure, the list of candidates in the search is usually pruned using a heuristic similarity cutoff that can be highly specific to the algorithms and templates that are used. This buildup procedure is discussed in some of the isomorphism searches (Nilmeier et al. 2013), and in variants of the geometric hashing technique (Chen et al. 2007b).

Care must be applied in determining these cutoffs, especially in the time-intensive search portions. If the similarity cutoffs are relaxed, false positives may be obtained. More importantly, however, the scaling can rapidly become unmanageable, since each list is carried into the next iteration. On the other hand, if the cutoffs are too strict, then good matches are discarded. In addition to the pruning criterion, other measures are applied to restrict the search space. For example, in graph comparison algorithms the default description of the resulting graph would contain all distances, resulting in a large, fully connected (clique) structure graph. Nearly all of these edges are unnecessary when comparing the graphs, so careful construction of the graphs beforehand will vastly improve performance.

Application of similarity thresholds can be a nontrivial effort, and very specific to the templates under consideration. Consider the residues in the lower right hand corner of Fig. 11.3. The active site residues are represented as Cα and side chain centroids (Oldfield 2002). In this case, centroid position is highly conserved, but the Cα position is not, and the residue identity is also different (Asp $\longrightarrow$ Glu). The choice of which constraints to apply and which to relax in this case would require detailed knowledge about the significant elements involved (in this case, the proton abstraction residue).

11.3.3 Methods with Nonstandard Motif Information

It is not always straightforward to differentiate between methods that use 'standard' 3D motifs from methods that incorporate additional information. For example, many techniques have multiple stages. In these techniques, a fast template matching algorithm is used to generate an initial candidate list, followed by a more complex scoring procedure to refine results (Laskowski et al. 2005; Kirshner et al. 2013; Nilmeier et al. 2013). While the second stage scoring procedure may incorporate more complex representations of the catalytic site, the core search algorithm uses the classic definition of a motif.

Other methods, however, incorporate a fundamentally different definition of a motif in the primary search machinery. For example, hybrid point-surface and single-point-centred descriptions of local structure do not fall under our working definition of a 3D motif approaches, but they do share many similarities. Methods primarily based on surface descriptions are covered in Chap. 10.

- **Single-Point-Centred Descriptions**. The program FEATURE (Bagley and Altman 1995) describes local structure as a set of properties in concentric shells emanating from a single point. The properties include descriptors of atoms, functional groups, residues, secondary structure, and simple biophysical characteristics. Because values are summed over spherical shells, however, directional information is lost. Both the WebFEATURE server (Liang et al. 2003; Buturovic et al. 2014) and the Structure-Based Local Environment Search Tool S-BLEST web server (Mooney et al. 2005; Peters et al. 2006) use FEATURE templates, and each provide their own results, along with enhanced annotations (Table 11.1).
- **Hybrid (Point-Surface) Descriptions**. Cavbase (Schmitt et al. 2002; Kuhn et al. 2006) and SiteEngine (Shulman-Peleg et al. 2004) describe binding sites as collections of pseudoatoms and their associated surface patches. The pseudoatoms represent surface-exposed functional groups of various types, such as a hydrogen bond donor or acceptor. Comparisons involve finding geometrically and physicochemically consistent sets of pseudoatoms, superimposing structures based on those matches, and then scoring based on surface patch overlap and physicochemical similarity. Surface points typically far outnumber the pseudoatoms, so scoring is relatively computationally demanding. The SiteEngine web server (Shulman-Peleg et al. 2005) performs pairwise comparisons but not database searches(Table 11.1). Other surface-based methods include eF-site (Kinoshita and Nakamura 2003), SuMo (Jambon et al. 2003), SiteEngine (Shulman-Peleg et al. 2004), and Query3D (Ausiello et al. 2005a)

11.3.4 Interpretation of Results

The previous sections have discussed the technical challenge of finding a given motif in a structure. However, there are still questions that must be answered when applying these methods. What can be said about the function of the structure if a positive match is found? What constitutes a positive match, and how reliable is it?

Several issues must be considered when deciding what a positive match means. The ideal case is when the motif perfectly defines the residues for a particular annotated function. In these cases, the interpretation of the match is straightforward: the structure has the annotated function that the matching motif has. Developing a motif library with these desirable properties is a challenge in itself, and is discussed in Sect. 11.3. This simple mapping of function from a motif to the structure is not always straightforward, as motifs may be only indirectly associated with a specific function. For example, if a motif is derived from a SCOP superfamily, a match may only imply some function which is commonly found in the SCOP structure.

Any given motif-to-structure comparison is an NP-hard challenge, and even an efficient procedure may still yield several different candidate matches. Additionally, motif libraries can number on the order of thousands, while the PDB has tens of thousands of structures. A comparison of the full set of possibilities can quickly lead to an intractable problem unless sensible cutoffs to candidate matches are applied during the evaluation steps.

It is even more important to be able to report a manageable list of matches that can be easily interpreted and understood by users. This list will likely contain trivial matches of nearly exact motifs found in proteins with very similar global structure. The more interesting matches in the list should include somewhat distant but still plausible relationships; possibly with residue substitutions, or noticeable differences in global structures.

Basic measures of structural similarity are usually the starting point for scoring. The root mean square deviation, or RMSD, is one very common measure. It has many limitations, however. Most notably, it is not a useful measure when comparing matches to motifs of different sizes. Many other nuances begin to become apparent, including substitution allowances as well as subtle geometric relationships that may not be properly represented by the reduced geometric form of the motif.

To account for these issues and provide a better ranking of hits, some groups apply a multistage method. The fast, coarse search method will generate a large candidate list that is then subjected to a more rigorous scoring procedure. Sometimes the scoring procedure is intended to have a direct statistical interpretation, much like a p-value or other probabilistic score (Barker and Thornton 2003; Nilmeier et al. 2013; Kirshner et al. 2013). The determination of the cutoff score, which indicates whether the candidate is a positive match, can often be heuristic. There are, however, classic machine learning techniques that can be applied to determine appropriate cutoffs.

The ability of a procedure to identify true positives, measured by the true positive rate (TPR) or ***sensitivity***, while also minimizing the false positive rate (FPR) is usually the measure of performance of many of these techniques. One technique that is used frequently is the Receiver Operating Characteristic (Bairoch), which is simply a plot of these values as the cutoff is adjusted, in which the Area Under the Curve (AUC) indicates a quality measure of the prediction procedure. This is only one of many techniques to identify good cutoff values, but is widely used in the motif matching literature and elsewhere.

Another approach to interpretation is to take the predictions of multiple methods into consideration. This can often prove to be more useful than relying on any one particular method. Some servers provide predictions from multiple sources, leaving the final determination to the user. Notable examples include the ProKnow server (Pal and Eisenberg 2005) and ProFunc (Laskowski et al. 2005) servers, and are listed in Table 11.1. These servers are also discussed in detail in Chap. 13.

Finally, common sense must be applied. Many confounding factors will still present themselves, even in the most carefully constructed procedures. For example, a motif may be correctly located in a structure, but there is no actual binding cavity to accommodate the substrate. It is prudent, if not essential, to inspect matches visually and to evaluate them using biologically relevant criteria when inferring the function from a match. Many of the most useful servers and software have some visualization process as an integral part of the procedure for studying matches, simply because expert evaluation of the matches is still the best way to determine if algorithms are working as expected.

11.4 Methods for Deriving Motifs

Most of the effort in motif matching approaches is invested in locating a motif in a protein structure. This challenge, however, assumes that the motif is available as a ground truth. Sometimes the methods allow the user to supply a motif, while other methods use a library of motifs. How, then, are these motifs generated in the first place?

Ideally, for ***motif discovery***, the set of positive examples should be as diverse as possible while retaining the common feature, and the negative examples should be as similar as possible to the positive examples while lacking that feature. In practice, the positive and negative sets may not be ideal, and part or all of a derived 3D motif could still reflect common ancestry or coincidence rather than shared function.

Others treat motif discovery or generation of motif libraries essentially as the primary goal of their method.

11.4.1 Literature Search and Manual Curation

Perhaps the most reliable approach to motif discovery is to mine the published literature for experimental evidence. For 3D motifs, the focus is on residues that provide a specific binding or catalytic capability.

The Catalytic Site Atlas (CSA) (Table 11.1) contains several hundred families of enzymes, each comprised of a structure with catalytic residue annotations from the literature (Barker and Thornton 2003; Porter et al. 2004; Torrance et al. 2005). The atlas library also includes structures related through sequence homology. Representative structural templates (3D motifs) are based on side-chain functional atoms, alpha carbons (Cα) and beta carbons (Cβ). In all, more than 2200 unique motifs were generated, whose function is verified through literature values, which often include experimental verification of the function.

The generation of this dataset was a fundamental advance in the field. Other servers rely on this dataset, including multiservers like ProFunc, (Laskowski et al. 2005), and groups who have curated or modified this Atlas and incorporated it in their own servers (Moll et al. 2011; Kirshner et al. 2013; Nilmeier et al. 2013).

11.4.2 Annotated Sites in PDB Structures

Another approach is to use the annotations given to the crystallographic structures in the PDB. In practice, this means looking at the SITE records of a given protein databank file, or at residues around molecules labelled as LIGAND. Sometimes even the residues around nonspecific heteroatoms (HET) or analysis of the residues of macromolecular interfaces can give some clue as to what portions of a protein may be involved in catalysis. This is not always informative, as these annotations are not guaranteed to point to the catalytic site of the protein of interest. It is often a very good starting point, however, and can provide new hypothesis for motifs.

Several databases of 3D motifs have been generated using only information from each source structure individually. For example, binding site motifs can be collected by taking residues within a cutoff distance of ligands, nucleic acids, or even other protein chains. The PINTS (Patterns in Non-homologous Tertiary Structures) server (Stark and Russell 2003) derives its database from binding sites defined as residues within three angstroms of a ligand as well as motifs annotated in the PDB as a SITE record (Russell 1998), along with careful statistical models (Stark et al. 2003, 2004) that estimate the statistical significance of matches. The PDBSite database (Ivanisenko et al. 2005) (Table 11.1) includes SITE records, along with interfacial reaction sites with other proteins, RNA, and DNA. Residues with at least three atoms within five angstroms of the other chain are included in an interaction site. The search machinery is called PDBSiteScan (Ivanisenko et al. 2004) (Table 11.1). The pdbFun web server (Ausiello et al. 2005b) uses sites defined as residues within 3.5 angstroms of a ligand (Ausiello et al. 2005a).

11.4.3 Mining for Emergent Properties

When groups of structures are studied, local structural features shared among proteins may be taken as a 3D motif. The process of identifying these common features may be described as the ***mining*** step. It is helpful to separately identify the grouping methods as either ***undirected*** or ***directed***. In general, undirected (unsupervised) mining methods do not specifically use labels or annotations in the grouping step, while directed (supervised) mining methods tend to use labelled structures. Each approach will be discussed in the following sections.

In some cases investigators provide a mining toolset for the user. The technology is focused on mining the pattern or motif from a group, rather than in how the groups are defined. The *applications* of these methods are, in general, directed mining approaches. At the heart of these techniques is a search for a ***clique*** that is common to the grouping that can be interpreted as a functionally important motif. Methods such as the common structural cliques method (Milik et al. 2003), the maximum common clique (ProBIS) algorithm (Konc and Janežič 2010), as well as the Detection of REcurring Sidechain PATterns (DRESPAT), (Kar et al. 2012) are all designed to locate maximal cliques among sets of structures.

In other cases, the approaches for determining a motif are more dependent on the nature of the groupings: these are discussed in the next sections.

11.4.3.1 Undirected Mining

Undirected mining refers to finding common patterns in unannotated, or ***unlabelled*** structures. The undirected mining approaches have elements of what is usually considered ***unsupervised learning***. For example, many of these approaches make ***all-to-all*** similarity comparisons (Russell 1998), which has some analogy to the notion of a ***distance matrix*** as seen in traditional clustering methods. Structures with sufficiently similar measures are grouped as a cluster. Other methods count motifs that appear with relatively high frequency (Oldfield 2002), and consider the structures having those motifs as a grouping.

Mining techniques apply to both unlabelled and labelled groupings, as well as cases where the distinction between unlabelled and labelled is not always straightforward. For example, a study that used groups of structures with similarity to sites with hypothesized function (Ausiello et al. 2007) was able to detect and propose new motifs. The reference structures were based on sequence similarity, proximity to a co-crystallized ligand, or contact with a cavity, but did not have a specific functional annotation.

11.4.3.2 Directed Mining

In directed mining, the focus is on the use of ***labelled*** examples to suggest geometric features (residues) that are common, both within the labelled dataset and other structures that may be deemed similar to the labelled dataset. Directed mining may also be considered to be more of a targeted search for motifs and themes within a given group.

In general, only ***positive examples*** are used for motif discovery. Positive examples are those structures whose labels indicate a positive membership in the functional set. The motif discovery process is then to find what essential features define that set. The use of ***negative examples*** is not as frequent in the motif discovery process. It does, however, appear in the validation of the models. One notable exception to this approach is the GASPS method (Polacco and Babbitt 2006), which uses both positive and negative examples in the motif discovery process, and is discussed in the next section.

It is often more practical to develop motifs from crystallographic structures where the ligand is present. Studies of this sort tend to be more specific to the ligand types of interest. For example, one of the early approaches was developed for adenine mononucleotide sites, based on the fact that there were over 100 structures available for comparison at the time (Kobayashi and Go 1997). A high similarity was found between structures of different folds, which is a hallmark of a good motif. Later, after many more structures had become available, a similar approach was used to generate consensus binding-site motifs (Nebel et al. 2007), and the study was expanded to study mono-, di-, and tri-phosphate complexes as well, resulting in 13 high quality motifs. The same group developed motifs specifically for porphyrin-binding sites (Nebel 2006). Another study used phosphate groups as the ligand in protein-nucleotide complexes, and applied a clique detection algorithm to discover motifs (Brakoulias and Jackson 2004).

Other methods use more standard template-matching programs, but on smaller motifs, with emergent motifs built from the smaller ones. The funClust server (Ausiello et al. 2008) (Table 11.1) identifies 3D motifs shared by up to 20 input structures. The structures are then filtered by sequence identity and other geometric filters, and the comparison is made with Query3D (Ausiello et al. 2005a). Another method, the PAR-3D (Protein Active site Residues using 3-Dimensional structural motifs) server (Goyal et al. 2007) (Table 11.1) compares a structure to motifs for proteases, glycolysis enzymes, and metalloenzyme sites with only three or four residues (Goyal and Mande 2008) that are common to the broadly defined functions. The motifs returned are given as allowed ranges of interatomic distances to the library of motifs. Another approach, termed Geometric Sieving, starts with an existing motif or list of putatively important residues (Chen et al. 2007b), and develops candidate motifs by comparing them to a representative sample of structures. It is assumed that the low-RMSD tail in a distribution represents true positives.

11.4.3.3 Directed Mining with Positive and Negative Examples

In most of the approaches listed above, only sets with known positives are used to discover the emergent features of a binding-site. Sometimes, however, it is important to know not just consensus features of a catalytic site, but the *essential* features. For this more subtle delineation, negative examples are needed to more precisely define what is an outlier.

For example, a simple mutation from Asp to Glu in a set of binding site residues may still preserve function, while a mutation from Asp to Asn may remove function completely if the residue needs to be protonated at some point in the catalysis. If, however, the residue only needs to be polar, then the Asp to Asn mutation might still be allowable.

These types of differences may not be easily seen by consensus methods, but some very carefully chosen negative examples can reveal these more subtle differences. The use of negative training examples is well understood in machine learning approaches with linear models. Here, the goal is to discover geometric features, rather than to apply a fitting procedure to determine parameters for a linear model. This presents a fundamentally different optimization problem.

One very successful approach to this problem is GASPS (Genetic Algorithm Search for Patterns in Structures), which finds patterns of residues that best separate the two groups (Polacco and Babbitt 2006). No prior residues list is required, and how the positive/negative groups are defined is independent of the method. The underlying search tool is SPASM (Kleywegt 1999), with residues represented by alpha-carbons and side chain centroids and only identical residue types allowed to match. To limit the search space, GASPS considers only the 100 most conserved residues in a structure chain, based on an automatically constructed sequence alignment. An initial candidate motif is constructed by picking one residue randomly and then choosing four more, also randomly except in the vicinity of the first. Each of 50 initial candidates is scored on how well it separates the positive and negative structures in terms of best match RMSD values. In each round of the genetic algorithm, the 16 highest-scoring motifs are used as the parents of 36 new motifs, and the top-scoring motif after 50 rounds is declared the winner. Motifs are allowed to contain from three to ten residues. Sensitive and specific motifs were obtained for diverse superfamilies (Babbitt and Gerlt 2000) and serine proteases. Most of the residues in the motifs were functionally important, but in some cases, residues with no known functional role were found to be equally predictive (Polacco and Babbitt 2006).

The GASPSdb database (Table 11.1) allows browsing and downloading 3D motifs previously generated by GASPS for SCOP families and superfamilies.

11.5 Molecular Docking for Functional Annotation

Ultimately, the ligand specificity and catalytic capabilities of a protein depend on the arrangement of atoms in its binding or active site(s). The use of 3D motifs can be seen as informatics approaches that are informed by the chemistry of the protein. These methods are limited in that they can only associate function to known motifs. There are many cases, however, where a high resolution target structure is available (either experimentally or through homology modelling), but there is no identifiable motif in the structure. For these cases, a more fundamental physical approach can fill in gaps in knowledge that the informatics approaches do not provide.

A computational method known as ***ligand docking*** can provide a different perspective on the problem of functional annotation (Jacobson et al. 2014). This technique (also mentioned in Chap. 10) uses molecular mechanics forcefields to directly estimate ligand protein energetics and complementarities. The field of docking is vast, and we list only a few examples for reference (Meng et al. 1992; Wang et al. 2003). As the name suggests, the molecule is 'docked' into the target protein, and the quality of the resulting pose is evaluated for fitness. Figure 11.6 illustrates a typical workflow that uses docking as a method for functional assignments. In general, the target is held rigid, but more recent approaches also allow for sidechain flexibility (Sherman et al. 2006). Since it is based on molecular interaction energies, this technique can conceivably predict molecular binding modes that are novel, but still physically reasonable.

Traditionally, database docking, or **in silico screening** has been applied to the ***lead discovery*** phase of drug design pipelines. As such, the technique is highly automated, and designed to dock large libraries of small molecules to selected targets (on the order of a million of compounds or more in some cases). While most ligand docking studies are focused on finding *inhibitors* to the target, the functional annotation effort seeks to find the *native metabolite* that is catalysed in the target. Many of the technical challenges in ligand docking are common to both goals, however, such as the need to distinguish true positives from false positives, or ***decoys*** (Huang et al. 2006). These studies highlight the need not only for high quality poses, but also for scoring procedures that will properly rank ligand affinities. Metabolite docking can be distinct from inhibitor docking, most notably due to the fact that most metabolites are highly charged (Song et al. 2007).

Despite these challenges, this approach has received considerable attention (Favia et al. 2008; Kalyanaraman et al. 2005; Macchiarulo et al. 2004; Paul et al. 2004; Tyagi and Pleiss 2006; Jacobson et al. 2014) In particular, studies of alpha-beta barrel enzymes (Song et al. 2007) and amidohydrolases (Hermann et al. 2007) have firmly established the capabilities of docking approaches as a supplement to approaches using sequence- and motif-based comparative approaches.

As these approaches have progressed, an emergent challenge for functional annotation is to not only generate comparative affinities for a particular target, but also to be able to compare affinities *across* targets. While inhibitor design is usually focused on a single target, the goal of functional annotation is to characterize entire

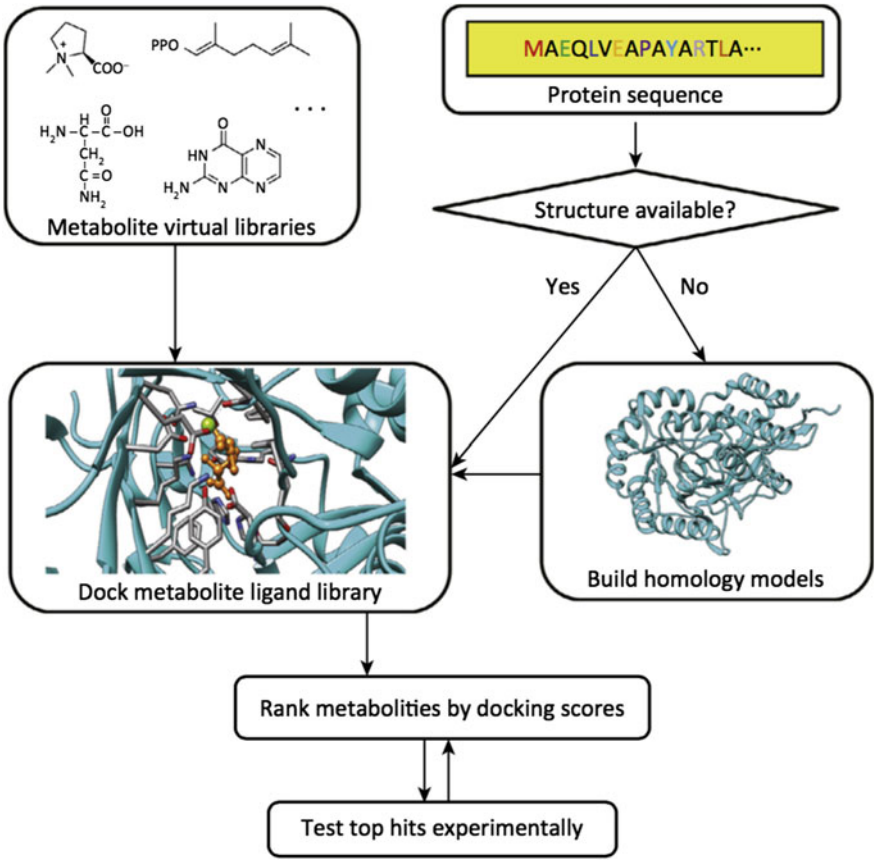


Fig. 11.6 Structure-based virtual metabolite docking protocol for enzyme activity prediction. When no structure has been experimentally determined for a protein sequence, a model can be built using a variety of comparative modelling methods, if sequence identity is approximately 30% or more. Whether using a structure of a model, it is critical that active site metal ions and cofactors are present, and that catalytic residues are positioned appropriately for catalysis. Virtual metabolites libraries can be constructed and 'docked' against the putative active sites of structures or models using computational tools more commonly used in structure-based drug design (e.g., Glide or DOCK). The docking scoring functions can be used to rank the ligands according to their estimated relative binding affinities. Top-scoring metabolites are typically inspected for plausibility and then selected for in vitro testing. (This Figure was reprinted from Jacobson et al. (2014) with permission from Elsevier License #3624901501981)

synthetic pathways or proteomes in an automated fashion. Studying target groups for entire synthetic pathways provides a much larger perspective, as the molecules are related by a chain of incremental modifications, and the targets are often expressed from the same '***genome neighborhood***'. Applying these additional guidelines for self consistency, while also using homology modelling to construct missing targets, can allow for elucidation of complete pathways that were

previously unknown (Zhao et al. 2013), with potential applications to synthetic biology and other efforts that have not traditionally relied on structure-based techniques (Jacobson et al. 2014).

While molecular recognition techniques are significantly more computationally demanding than 3D motif matching, docking has the potential to extrapolate to functions not associated with previously characterized structures, and represents a frontier direction in the field for the most challenging of catalytic sites.

11.6 Discussion and Conclusions

The question of how best to describe the function of a protein with a meaningful language remains. While fold-based methods and ligand-based methods have been shown to be very useful, the use of a 3D motif as a signature for protein function has offered new perspectives on catalytic sites, and could ultimately form the foundation of a functional annotation language. Challenges remain on how to identify these motifs, and even with knowledge of the substrate and many examples, it can be nontrivial to identify the ideal 3D motif that uniquely and completely defines function for a given enzyme.

What, then, is the most natural classification of protein function, if we choose 3D motifs as a basis for classification? In enzymes, individual residues or functional groups play different roles in the course of a reaction: substrate recognition, catalysis of a particular step in the reaction, stabilization of an intermediate, or some combination of these. As proteins evolve to perform new functions, they can make use of existing pieces of catalytic machinery that carry out a ***common partial reaction*** (Babbitt and Gerlt 2000; Bartlett et al. 2003). This explains in part why members of a homologous but diverse group of enzymes often make use of the same configuration of a small number of amino acids, despite catalysing different overall reactions. It may well be that these subunits (which are 3D motifs) will form the basic building blocks of all enzymes, and a functional classification scheme should include these basic units in its language.

Acknowledgements We acknowledge support from NIH GM60595 and NSF DBI-0234768. Molecular graphics were produced with the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41-GM103311). We thank Jacquelyn Fetrow and Stacy Knutson (Wake Forest University) for providing Fig. 11.5 as an example of a result from their FFF/DASP/PASS motif analysis software. We gratefully acknowledge Dan Kirshner for enlightening discussions and a critical reading of the manuscript.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32(suppl 1):D226–D229

Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36 (suppl 1):D419–D425

Arakaki A, Huang Y, Skolnick J (2009) EFICAz2: enzyme function inference by a combined approach enhanced by machine learning. BMC Bioinform 10(1):107

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 structures. J Mol Biol 243(2):327–344

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25(1):25–29

Ausiello G, Gherardini PF, Marcatili P, Tramontano A, Via A, Helmer-Citterich M (2008) FunClust: a web server for the identification of structural motifs in a set of non-homologous protein structures. BMC Bioinform 9(Suppl 2):S2

Ausiello G, Peluso D, Via A, Helmer-Citterich M (2007) Local comparison of protein structures highlights cases of convergent evolution in analogous functional sites. BMC Bioinform 8 (Suppl 1):S24

Ausiello G, Via A, Helmer-Citterich M (2005a) Query3D: a new method for high-throughput analysis of functional residues in protein structures. BMC Bioinform 6(Suppl 4):S5

Ausiello G, Zanzoni A, Peluso D, Via A, Helmer-Citterich M (2005b) pdbFun: mass selection and fast comparison of annotated PDB residues. Nucleic Acids Res 33 (Web Server issue):W133–137

Babbitt PC (2003) Definitions of enzyme function for the structural genomics era. Curr Opin Chem Biol 7(2):230–237

Babbitt PC, Gerlt JA (1997) Understanding enzyme superfamilies. Chemistry As the fundamental determinant in the evolution of new catalytic activities. J Biol Chem 272(49):30591–30594

Babbitt PC, Gerlt JA (2000) New functions from old scaffolds: how nature reengineers enzymes for new functions. Adv Protein Chem 55:1–28

Bagley SC, Altman RB (1995) Characterizing the microenvironment surrounding protein sites. Protein Sci 4(4):622–635

Bairoch A (1994) The ENZYME data bank. Nucleic Acids Res 22(17):3626–3627

Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294 (5540):93–96

Barker JA, Thornton JM (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics 19(13):1644–1649

Bartlett GJ, Borkakoti N, Thornton JM (2003) Catalysing new reactions during evolution: economy of residues and mechanism. J Mol Biol 331(4):829–860

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242

Blow DM, Birktoft JJ, Hartley BS (1969) Role of a buried acid group in the mechanism of action of chymotrypsin. Nature 221(5178):337–340

Brakoulias A, Jackson RM (2004) Towards a structural classification of phosphate binding sites in protein–nucleotide complexes: an automated all-against-all structural comparison using geometric matching. Proteins Struct Funct Bioinf 56(2):250–260

Brylinski M, Skolnick J (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci 105(1):129

Buturovic L, Wong M, Tang GW, Altman RB, Petkovic D (2014) High precision prediction of functional sites in protein structures. Publ Libr Sci One 9(3):e91240

Cammer SA, Hoffman BT, Speir JA, Canady MA, Nelson MR, Knutson S, Gallina M, Baxter SM, Fetrow JS (2003) Structure-based active site profiles for genome analysis and functional family subclassification. J Mol Biol 334(3):387–401

Chen BY, Bryant DH, Cruess AE, Bylund JH, Fofanov VY, Kristensen DM, Kimmel M, Lichtarge O, Kavraki LE (2007a) Composite motifs integrating multiple protein structures increase sensitivity for function prediction. Comput Syst Bioinform Conf 6:343–355

Chen BY, Fofanov VY, Bryant DH, Dodson BD, Kristensen DM, Lisewski AM, Kimmel M, Lichtarge O, Kavraki LE (2007b) The MASH pipeline for protein function prediction and an algorithm for the geometric refinement of 3D motifs. J Comput Biol 14(6):791–816

Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. Science 300(5626):1701–1703

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5(4):823–826

Conte LL, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C (2000) SCOP: a structural classification of proteins database. Nucleic Acids Res 28(1):257–259

Devos D, Valencia A (2001) Intrinsic errors in genome annotation. Trends Genet 17(8):429–431

Di Gennaro JA, Siew N, Hoffman BT, Zhang L, Skolnick J, Neilson LI, Fetrow JS (2001) Enhanced functional annotation of protein sequences via the use of structural descriptors. J Struct Biol 134(2–3):232–245

Favia AD, Nobeli I, Glaser F, Thornton JM (2008) Molecular docking for substrate identification: the short-chain dehydrogenases/reductases. J Mol Biol 375(3):855–874

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

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(5):769–778

Furnham N, Holliday GL, de Beer TA, Jacobsen JO, Pearson WR, Thornton JM (2014) The catalytic site atlas 2.0: cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Res 42 (D1):D485–D489

Galperin MY, Walker DR, Koonin EV (1998) Analogous enzymes: independent inventions in enzyme evolution. Genome Res 8(8):779–790

Gerlt JA, Allen KN, Almo SC, Armstrong RN, Babbitt PC, Cronan JE, Dunaway-Mariano D, Imker HJ, Jacobson MP, Minor W (2011) The enzyme function initiative. Biochem

Gerlt JA, Babbitt PC (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu Rev Biochem 70(1):209–246

Gerlt JA, Babbitt PC, Jacobson MP, Almo SC (2012) Divergent evolution in enolase superfamily: strategies for assigning functions. J Biol Chem 287(1):29–34

Glanville JG, Kirshner D, Krishnamurthy N, Sjölander K (2007) Berkeley phylogenomics group web servers: resources for structural phylogenomic analysis. Nucleic Acids Res 35(suppl 2): W27–W32

Glazer DS, Radmer RJ, Altman RB Combining molecular dynamics and machine learning to improve protein function recognition. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 2008. NIH Public Access, p 332

Gold ND, Jackson RM (2006a) Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. J Mol Biol 355(5):1112–1124

Gold ND, Jackson RM (2006b) SitesBase: a database for structure-based protein–ligand binding site comparisons. Nucleic Acids Res 34(suppl 1):D231–D234

Goyal K, Mande SC (2008) Exploiting 3D structural templates for detection of metal-binding sites in protein structures. Proteins 70(4):1206–1218

Goyal K, Mohanty D, Mande SC (2007) PAR-3D: a server to predict protein active site residues. Nucleic Acids Res 35 (Web Server issue):W503–505

Halgren T (2007) New method for fast and accurate binding-site Identification and analysis. Chem Biol Drug Des 69(2):146–148

Halgren TA (2009) Identifying and characterizing binding sites and assessing druggability. J Chem Inf Model 49(2):377–389

Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, Raushel FM (2007) Structure-based activity prediction for an enzyme of unknown function. Nature 448 (7155):775–779

Holliday GL, Almonacid DE, Bartlett GJ, O'Boyle NM, Torrance JW, Murray-Rust P, Mitchell JBO, Thornton JM (2007) MACiE (mechanism, annotation and classification in enzymes): novel tools for searching catalytic mechanisms. Nucleic Acids Res 35(suppl 1): D515–D520

Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49(23):6789–6801

International Union of Biochemistry and Molecular Biology: Nomenclature Committee, Webb EC (1992) Enzyme nomenclature 1992: recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes. Academic Press, San Diego

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(Web Server issue):W549–554

Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA (2005) PDBSite: a database of the 3D structure of protein functional sites. Nucleic Acids Res 33(Database issue):D183–187

Jacobson MP, Kalyanaraman C, Zhao S, Tian B (2014) Leveraging structure for enzyme function prediction: methods, opportunities, and challenges. Trends Biochem Sci 39(8):363–371

Jambon M, Andrieu O, Combet C, Deléage G, Delfaud F, Geourjon C (2005) The SuMo server: 3D search for protein functional sites. Bioinformatics 21(20):3929–3930

Jambon M, Imberty A, Deléage G, Geourjon C (2003) A new bioinformatic approach to detect common 3D sites in protein structures. Proteins Struct Funct Bioinf 52(2):137–145

Kalyanaraman C, Bernacki K, Jacobson MP (2005) Virtual screening against highly charged active sites: identifying substrates of alpha-beta barrel enzymes. Biochemistry 44(6):2059–2071

Kalyanaraman C, Imker HJ, Fedorov AA, Fedorov EV, Glasner ME, Babbitt PC, Almo SC, Gerlt JA, Jacobson MP (2008) Discovery of a dipeptide epimerase enzymatic function guided by homology modeling and virtual screening. Structure 16(11):1668–1677

Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 105(28):6474–6487

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30

Kar S, Vijayakeerthi D, Tendulkar AV, Ravindran B Functional site prediction by exploiting correlations between labels of interacting residues. In: Proceedings of the ACM conference on bioinformatics, computational biology and biomedicine, 2012. ACM, pp 76–83

Kinjo AR, Nakamura H (2007) Similarity search for local protein structures at atomic resolution by exploiting a database management system. Biophysics 3:75–84

Kinoshita K, Nakamura H (2003) Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. Protein Sci 12(8):1589–1595

Kirshner DA, Nilmeier JP, Lightstone FC (2013) Catalytic site identification—a web server to identify catalytic site structural matches throughout PDB. Nucleic Acids Res 41 (W1):W256–W265

Kleywegt GJ (1999) Recognition of spatial motifs in protein structures. J Mol Biol 285 (4):1887–1897

Kleywegt GJ, Jones TA (1997) Detecting folding motifs and similarities in protein structures. Methods Enzymol 277:525–545
Kleywegt GJ, Lamerichs RMJN, Boelens R, Kaptein R (1989) Toward automatic assignment of protein 1H NMR spectra. J Magn Reson 85(1):186–197
Kobayashi N, Go N (1997) A method to search for similar protein local structures at ligand binding sites and its application to adenine recognition. Eur Biophys J 26(2):135–144
Konc J, Janežič D (2010) ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. Bioinformatics 26(9):1160–1168
Konc J, Janežič D (2012) ProBiS-2012: web server and web services for detection of structurally similar binding sites in proteins. Nucleic Acids Res 40(W1):W214–W221
Krishnamurthy N, Brown DP, Kirshner D, Sjölander K (2006) PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. Genome Biol 7(9):R83
Krogh A, Brown M, Mian IS, Sjolander K, Haussler D (1994) Hidden Markov models in computational biology: applications to protein modeling. J Mol Biol 235(5):1501–1531
Kuhn D, Weskamp N, Schmitt S, Hullermeier E, Klebe G (2006) From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. J Mol Biol 359(4):1023–1044
Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, McWilliam H, Valentin F, Wallace I, Wilm A, Lopez R (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23 (21):2947–2948
Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33(Web Server issue):W89–W93
Liang MP, Banatao DR, Klein TE, Brutlag DL, Altman RB (2003) WebFEATURE: an interactive web tool for identifying and visualizing functional sites on macromolecular structures. Nucleic Acids Res 31(13):3324–3327
Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257(2):342–358
Macchiarulo A, Nobeli I, Thornton JM (2004) Ligand selectivity and competition between enzymes in silico. Nat Biotechnol 22(8):1039–1045
Meng EC, Polacco BJ, Babbitt PC (2004) Superfamily active site templates. Proteins Struct Funct Bioinf 55(4):962–976
Meng EC, Shoichet BK, Kuntz ID (1992) Automated docking with grid-based energy evaluation. J Comput Chem 13(4):505–524
Milik M, Szalma S, Olszewski KA (2003) Common structural cliques: a tool for protein structure and function analysis. Protein Eng 16(8):543–552
Mitchell EM, Artymiuk PJ, Rice DW, Willett P (1990) Use of techniques derived from graph theory to compare secondary structure motifs in proteins. J Mol Biol 212(1):151–166
Moll M, Bryant DH, Kavraki LE (2010) The LabelHash algorithm for substructure matching. BMC Bioinform 11(1):555
Moll M, Bryant DH, Kavraki LE (2011) The LabelHash server and tools for substructure-based functional annotation. Bioinformatics 27(15):2161–2162
Moll M, Kavraki LE (2008) LabelHash: a flexible and extensible method for matching structural motifs. Available from Nature Precedings. http://dx.doi.org/10.1038/npre.2008.2199.1
Mooney SD, Liang MH, DeConde R, Altman RB (2005) Structural characterization of proteins using residue environments. Proteins 61(4):741–747
Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4):536–540
Nebel JC (2006) Generation of 3D templates of active sites of proteins with rigid prosthetic groups. Bioinformatics 22(10):1183–1189
Nebel JC, Herzyk P, Gilbert DR (2007) Automatic generation of 3D motifs for classification of protein binding sites. BMC Bioinformatics 8(1):321
Nilmeier JP, Kirshner DA, Wong SE, Lightstone FC (2013) Rapid catalytic template searching as an enzyme function prediction procedure. Publ Libr Sci One 8(5):e62535

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27(1):29–34

Oldfield TJ (2002) Data mining the protein data bank: residue interactions. Proteins 49(4):510–528

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH–a hierarchic classification of protein domain structures. Structure 5(8):1093–1108

Orengo CA, Pearl FMG, Bray JE, Todd AE, Martin A, Conte LL, Thornton JM (1999) The CATH Database provides insights into protein structure/function relationships. Nucleic Acids Res 27 (1):275–279

Orengo CA, Pearl FMG, Thornton JM (2003) The CATH domain structure database. Struct Bioinform 249–271

Pal D, Eisenberg D (2005) Inference of protein function from protein structure. Structure 13 (1):121–130

Paul N, Kellenberger E, Bret G, Muller P, Rognan D (2004) Recovering the true targets of specific ligands by virtual screening of the protein data bank. Proteins 54(4):671–680

Pegg SC, Brown S, Ojha S, Huang CC, Ferrin TE, Babbitt PC (2005) Representing structure-function relationships in mechanistically diverse enzyme superfamilies. Pac Symp Biocomput 358–369

Pegg SCH, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC (2006) Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. Biochemistry 45(8):2545–2555

Pennec X, Ayache N (1998) A geometric algorithm to find small but highly similar 3D substructures in proteins. Bioinformatics 14(6):516–522

Peters B, Moad C, Youn E, Buffington K, Heiland R, Mooney S (2006) Identification of similar regions of protein structures using integrated sequence and structure analysis tools. BMC Struct Biol 6:4

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera–a visualization system for exploratory research and analysis. Journal of Computaional Chemistry 25(13):1605–1612

Polacco BJ, Babbitt PC (2006) Automated discovery of 3D motifs for protein function annotation. Bioinformatics 22(6):723–730

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(suppl 1):D129–D133

Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A (2013) A large-scale evaluation of computational protein function prediction. Nat Methods 10(3):221–227

Ren J, Xie L, Li WW, Bourne PE (2010) SMAP-WS: a parallel web service for structural proteome-wide ligand-binding site comparison. Nucleic Acids Res 38(suppl 2):W441–W444

Richardson JS (1981) The anatomy and taxonomy of protein structure. Adv Protein Chem 34:167–339

Rost B (1997) Protein structures sustain evolutionary drift. Fold Des 2(3):S19–S24

Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318(2):595–608

Russell RB (1998) Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 279(5):1211–1227

Sankararaman S, Sha F, Kirsch JF, Jordan MI, Sjölander K (2010) Active site prediction using evolutionary and structural information. Bioinformatics 26(5):617–624

Sankararaman S, Sjölander K (2008) INTREPID—INformation-theoretic TREe traversal for Protein functional site IDentification. Bioinformatics 24(21):2445–2452

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(2):387–406

Sherman W, Day T, Jacobson MP, Friesner RA, Farid R (2006) Novel procedure for modeling ligand/receptor induced fit effects. J Med Chem 49(2):534–553

Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11(9):739–747
Shindyalov IN, Bourne PE (2001) A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. Nucleic Acids Res 29(1):228–229
Shulman-Peleg A, Nussinov R, Wolfson HJ (2004) Recognition of functional sites in protein structures. J Mol Biol 339(3):607–633
Shulman-Peleg A, Nussinov R, Wolfson HJ (2005) SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. Nucleic Acids Res 33(Web Server issue):W337–W341
Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci CABIOS 12(4):327–345
Skolnick J, Brylinski M (2009) FINDSITE: a combined evolution/structure-based approach to protein function prediction. Briefings Bioinform 10(4):378–391
Song L, Kalyanaraman C, Fedorov AA, Fedorov EV, Glasner ME, Brown S, Imker HJ, Babbitt PC, Almo SC, Jacobson MP (2007) Prediction and assignment of function for a divergent N-succinyl amino acid racemase. Nat Chem Biol 3(8):486–491
Spriggs RV, Artymiuk PJ, Willett P (2003) Searching for patterns of amino acids in 3D protein structures. J Chem Inf Comput Sci 43(2):412–421
Stark A, Russell RB (2003) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. Nucleic Acids Res 31(13):3341–3344
Stark A, Shkumatov A, Russell RB (2004) Finding functional sites in structural genomics proteins. Structure 12(8):1405–1412
Stark A, Sunyaev S, Russell RB (2003) A model for statistical significance of local similarities in structure. J Mol Biol 326(5):1307–1316
Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673–4680
Tian W, Arakaki AK, Skolnick J (2004) EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. Nucleic Acids Res 32(21):6226–6239
Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333(4):863–882
Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307(4):1113–1143
Todd AE, Orengo CA, Thornton JM (2002) Plasticity of enzyme active sites. Trends Biochem Sci 27(8):419–426
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 (3):565–581
Tseng YY, Dundas J, Liang J (2009) Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. J Mol Biol 387(2):451–464
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(2):421–436
Tyagi S, Pleiss J (2006) Biochemical profiling in silico–predicting substrate specificities of large enzyme families. J Biotechnol 124(1):108–116
Ullmann JR (1976) An algorithm for subgraph isomorphism. J ACM (JACM) 23(1):31–42
Wallace AC, Borkakoti N, Thornton JM (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 6:2308–2323
Wallace AC, Laskowski RA, Thornton JM (1996) Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Sci 5(6):1001–1013

Wang R, Lu Y, Wang S (2003) Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 46(12):2287–2303

Webb EC (1992) Enzyme nomenclature 1992. In: Recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes, vol Ed. 6. Academic Press

Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. Q Rev Biophys 36(03):307–340

Wolfson HJ, Rigoutsos I (1997) Geometric hashing: An overview. Comput Sci Eng IEEE 4(4):10–21

Wright CS, Alden RA, Kraut J (1969) Structure of subtilisin BPN' at 2.5 angstrom resolution. Nature 221(5177):235–242

Xie L, Bourne PE (2008) Detecting evolutionary relationships across existing fold space, using sequence order-independent profile–profile alignments. Proc Natl Acad Sci 105(14):5441

Xie L, Bourne PE (2009) A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. Bioinformatics 25(12):i305–i312

Yang LW, Bahar I (2005) Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. Structure 13(6):893–904

Zemla A (2003) LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 31(13):3370–3374

Zhao S, Kumar R, Sakai A, Vetting MW, Wood BM, Brown S, Bonanno JB, Hillerich BS, Seidel RD, Babbitt PC (2013) Discovery of new enzymes and metabolic pathways by using structure and genome context. Nature 502(7473):698–702