1 Introduction

While functional motifs are commonly detected and studied in protein sequences, few three-dimensional (3D) motifs, that is sets of residues spatially close in three dimensions but not necessarily adjacent in the sequence, have been identified so far, mostly through manual approaches. However, structural motifs may reveal novel and important functional sites and allow the detection of evolutionary relationships that are often unrecognizable when the linear amino acid sequence is inspected. The occurrence of similar 3D motifs in unrelated protein structures can in principle allow the transfer of functional annotation and highlight interesting examples of independently evolved functional sites (convergent evolution). Furthermore, the increasing number of experimentally solved protein structures arising from structural genomics projects, oftentimes poorly annotated, requires structure-based methods for fast and reliable functional inference. Such methods generally rely on matching regions of the query proteins with structural motifs associated with known biochemical functions. The systematic and, whenever possible, automatic identification and annotation of new 3D motifs is an important challenge in bioinformatics. In particular, the compilation of a large and robustly annotated collection of structural motifs, conceptually similar to the several existing resources for sequence motifs (e.g. PROSITE (Hulo, et al. 2006), ELM (Gould, et al.), MmM (Ball, et al. 2006), would fill a substantial void in the field.

This chapter will cover several aspects of structural motifs and discuss both the approaches aimed at detecting and the procedures used to associate them to a function. The main issues and limitations of the methodologies based on 3D motif for function prediction will also be discussed.

1.1 Sequence Functional Motifs: a well-Established Research Area

To quote Arthur Lesk, one of the fundamental laws of molecular biology is that “The gene sequence determines amino-acid sequence, the amino-acid sequence determines the protein structure and the protein structure determines the protein function. Selection acts on function to shape the gene sequence.” (Lesk 2010). We can add here that the molecular function of a protein is often explicated through a limited number of amino acids. This means that function conservation across evolution is reflected in the conservation of sets of amino acids that are directly or indirectly involved in the protein function.
In 1988, Amos Bairoch started developing PROSITE, a database of biologically significant sites and patterns to be used for the inference of protein function(s) in the context of protein sequence analysis (Hulo, et al. 2006). PROSITE is based on the idea that conserved patterns reflect evolutionary constraints on limited and functionally relevant regions of protein sequences. In the PROSITE user manual, Bairoch cites Arthur Lesk as follows: "In some cases, the structure and function of an unknown protein which is too distantly related to any protein of known structure to detect its affinity by overall sequence alignment may be identified by its possession of a particular cluster of residues types classified as a motif. The motifs, or templates, or fingerprints, arise because of particular requirements of binding sites that impose very tight constraint on the evolution of portions of a protein sequence." (Lesk 1988).

Most PROSITE patterns are retrieved from the literature whereas new patterns are developed from regions of multiple sequence alignments thought or proved to be important for the biological function of the aligned proteins.

After that, and especially in the last decade, a fertile research area has developed around the idea of sequence motifs and several resources have been established for their discovery, annotation and use in function prediction. Among these, Pfam (Bateman, et al. 1999), SMART (Schultz, et al. 1998), PRINTS (Attwood, et al. 1994), BLOCKS (Henikoff, et al. 1991), DOMO (Gracy, et al. 1998), Prodom (Corpet, et al. 1998), ELM (Gould, et al.) and MnM (Balla, et al. 2006), plus several resources devoted to specific classes of functional sites (e.g. Phospho.ELM (Diella, Gould, et al. 2008), Scansite (Obenauer, et al. 2003), etc.).

Many of these resources provide collections of functional motifs, patterns or domains, usually accompanied by computational tools to search the occurrence of the motifs in sequences of unknown function and by appropriate statistical supports to assess the significance of motif occurrences. A prominent example is represented by the ScanProsite tool (de Castro, et al. 2006), which relies on the Nicodème methodology to handle motif statistics (Nicodeme 2001). Some resources, such as MnM and ELM, also provide biological filters to help users discriminate between true and false motifs.


1.2 Structural Motifs: Problems and Challenges

The research area of structural motifs is grounded on the same evolutionary assumptions that appertain to the field of sequence motifs, i.e. on the idea that functionally significant amino acids will be conserved during evolution in different proteins sharing a related function. There are several issues, however, making the 3D motif identification, characterization and use in function prediction more problematic and challenging. First of all, a clear definition of a 3D motif is needed. By analogy with the concept of sequence motif, a structural motif is commonly defined as a similar spatial arrangement of a limited set of amino acids encoding a specific biochemical function in the context of different protein structures. However, the discovery of so-defined motifs is a difficult task. In fact, we must take into account that functional amino acids, close in the protein structure as a rule, are neither necessarily close nor always co-linear in the protein sequence. This observation makes it difficult to use sequence alignments for 3D motif identification. On the other hand, as it will be discussed below, a conserved and almost identical 3D motif may occur in proteins with a rather different overall structure and the structural alignment of proteins with different structure is not straightforward.

Another important issue concerns the statistical significance of a 3D motif occurrence. As mentioned above, robust statistics have been developed to assess the significance of finding a motif in a protein sequence: in fact, in order to reliably associate the occurrence of a motif with the presence of a specific function, it is important to be able to estimate the probability that the motif occurs in a given sequence by mere chance. Nevertheless, statistical procedures that are effective for sequence motifs, cannot be easily applied to 3D motifs, due to the higher dimensionality of the system.

The chapter is organized as follows: first, we will focus on the definition and discovery of 3D functional motifs, by illustrating the paradigmatic example of the enzyme active sites and discussing the computational
approaches for the identification of novel 3D motifs. Next, we will describe databases collecting 3D functional motifs and introduce available resources for their annotation and analysis. Finally, we will discuss the issue of structure-guided function inference and the related problem of 3D motif statistical and biological significance.

2 Definition and Discovery of 3D Functional Motifs

2.1 A Clear Definition for 3D Functional Motifs

Three-dimensional functional motifs are commonly defined as recurring spatial arrangements of specific sets of amino acids participating in the protein function. Amino acids defining a 3D motif do not need to be close and/or co-linear in the protein sequence: only their vicinity in space, relative orientation and biochemical properties characterize them.

Sometimes, however, we have to recur to a looser definition: in some cases, a 3D motif can be described by a set of local structural properties (charge, secondary structure, solvent accessibility, occurrence in a cavity, etc) that are related to or even specific for a particular function. In other words, there are cases where we cannot claim that a “3D consensus” exists in terms of position and identity of the participating amino acids, as we often do for sequence consensi, but we can detect one or more local structural features that suggest or even unambiguously point out the presence of a functional site. This definition is implicitly adopted by computational methods that predict functional sites on the surface of proteins using distinctive structural and physicochemical characteristics (see Section 5).

Since protein structure determines protein function, it is reasonable that globally diverse proteins sharing a function display conserved local structural features. The structural conservation of a 3D functional site can be the result of the evolutionary pressure on function preservation in a process of divergent evolution or the effect of a convergent evolution process that has produced the same structural solution in unrelated proteins to achieve the same functional result.

One beautiful example of convergent evolution is represented by the serine protease active site. Serine proteases are a family of well-characterized proteases whose most prominent functional group in the active site, formed by three amino acids that constitute the catalytic triad, is a serine. Members of this family are usually evolutionarily related even if there are examples of non-related members such as the mammalian trypsin and the bacterial subtilisin. The latter enzyme is not evolutionarily related to the former; however, the atoms in subtilisin that participate in the enzymatic function are in almost identical positions relative to one another in the 3D structure as they are in trypsin and its relatives (Branden, et al. 1999). Figure 1 highlights the conservation of the catalytic triad residues in the trypsin and subtilisin structures. Another nice example of convergent evolution is represented by the phosphate-binding loop (P-loop) motif (Saraste, et al. 1990) (Via, Ferre, Brannetti, Valencia, et al. 2000) (Figure 2).

In these two cases, the atom coordinates of the structurally conserved residues can be used to describe two motifs encoding for the serine protease function and for phosphate-binding, respectively; these motifs can then be searched – by means of appropriate algorithms - in proteins structures the function of which is not yet characterized, such as those determined in structural genomics experiments. Structure-based function prediction will be discussed in Section 5.
**Protein Structural Motifs: Identification, Annotation and Use in Function Prediction**

**Figure 1:** The conservation of the catalytic triad residues in the trypsin and subtilisin structures. 

a) The trypsin structure (PDB:1FXY (Hopfner, et al. 1998)) is represented in blue marine ribbon. The catalytic triad is represented in yellow (left: SER195, middle: HIS57, right: ASP102); b) The subtilisin structure (PDB:1SBN (Heinz, et al. 1991)) is represented in orange ribbon. The catalytic triad is represented in red (left: SER221, middle: HIS64, right: ASP32); c) The serine protease 3D motif represented by the trypsin (yellow) and subtilisin (red) residues belonging to the catalytic triad.

**Figure 2:** Phosphate-binding 3D motif (P-loop). The GNP molecule, a non-hydrolyzable analog of GTP, is represented in yellow. The residues belonging to the 3D P-loop motif taken from different phosphate-binding proteins (c-H-ras p21, Ran, cdc 42, Rac 1, Rab 3A, Transducin alpha, Adp ribosylation factor-1, Elongation factor G, Rap 1A) are represented in red/orange color-scale. Notice that the P-loop residues are not aligned in sequence.
2.2 Discovery of Novel 3D Motifs

2.2.1 Manual Extraction of 3D Motifs

The choice of the most appropriate definition/description of 3D motifs is clearly essential to develop tools for their discovery. The question that arises is: how can novel 3D motifs be discovered? In other words, what can be done to identify a set of 3D variables (atoms or amino acids coordinates, structural or phisico-chemical properties, etc) that are necessary and sufficient to capture a specific biochemical function of a protein? The discovery of a sequence motif basically relies either on the detection of a conserved consensus in the multiple sequence alignment (MSA) of a set of distantly related proteins or on the idea that a linear motif could well be the only common sequence feature of a set of non-homologous proteins sharing a function (e.g. an interaction partner) and might thus be detectable simply by virtue of over-representation. Using non-homologous proteins ensures that over-representation is not the result of homology. Notice that MSAs of, or over-representation in, closely related protein sequences helps detect more general structural and functional relationships such as those encoded in family domains, rather than short functional consensi. It is important to pinpoint that, in order to assess if a short motif is over-represented in a set of sequences, it is necessary to be able to determine the probability of observing it in a similar set of sequences by chances, which implies that a background set of random but similar sequences must be constructed.

These relatively simple approaches for discovering functional sequence motifs cannot be trivially transferred to protein structures; in fact, on one hand, the structural alignment of distantly related proteins is a complex task and its does not necessarily have a unique solution and, on the other hand, the concept of over-representation applied to 3D motifs entails computational and statistical problems of difficult resolution (this issue is discussed in Section 6).

Therefore, in most cases, 3D motifs are manually extracted from the literature and/or determined by an expert (e.g. a crystallographer) through the manual analysis of the structure of proteins the function of which is experimentally known. One example is represented by 3D motifs stored in the Catalytic Site Atlas (Porter, et al. 2004), a database consisting of catalytic residue annotation for enzymes in the Protein Data Bank (PDB) (Berman, et al. 2000). Part of the CSA catalytic 3D motifs are manually annotated from the primary literature and part are inferred by sequence alignment to one of the manually collected sites. Figure 3 shows an example of the CSA representation of a 3D motif.

Manual extraction of 3D motifs is a time-consuming and difficult process, which does not keep up with the pace of the increasingly growing number of available protein structures, meaning that only a few 3D motifs are currently known compared to thousands of proteins that use them in protein-protein and protein-ligand interactions as well as in catalytic sites.

2.2.2 Automated Discovery of 3D Motifs

Mapping sequence motifs onto protein structures. In the last decade, several algorithms for functional sequence motif discovery have been published (e.g. (Bailey, et al. 2006, Frith, et al. 2008) (Davey, et al. 2006, Leung, et al. 2007, Redhead, et al. 2007); for a review see: (Bailey 2007) and, as a consequence, sequence motifs are continuously discovered through computational tools (even if the validation of these novel motifs and their association with biological functions still requires expensive and laborious experimental approaches).

Unfortunately, the research area of 3D motif computational discovery has not yet experienced a comparable expansion. However, in the recent years, some automated approaches for the identification and/or analysis of structural motifs have been developed.

Some of these approaches are based on the transfer of information from sequence motifs to protein structures. They rely on the assumption that, for a conserved sequence motif, there should be a common 3D structure (a “template”), which can be considered characteristic of the function. In other words, since sequence motifs contain functionally critical amino acids, they are expected to share structural homology. For example,
Kasuya and Thornton (Kasuya, et al. 1999) carried out a 3D structure analysis of the 20 PROSITE patterns with the largest numbers of hits, searching for common structural features among true positives and for structural distinctions between true and false positives. They found that, for many of the considered patterns, the RMSD values for the true hits were low enough to deduce that they have an overall common 3D structure, whereas the structures of false positives were clearly different from those of true positives. In their work, Kasuya and Thornton, did not map flexible PROSITE pattern with long spacers (stretches of wildcard consecutive positions in the pattern regular expression) onto protein structure. This is, in principle, reasonable, being amino acids in variable regions less likely to be functional. In this regard, Lin and colleagues (Lin, et al. 2000) carried out a 3D conformational analysis of long spacers in PROSITE patterns in order to check if a given pattern has a common backbone conformation and if this can be exploited to improve the accuracy of pattern recognition. These authors found patterns with long spacers that adopt a well-defined backbone conformation, in addition to conserved residues, even if this was not always the case. They suggest using this type of structural information to improve the specificity of the original PROSITE pattern search. The work of Via et al. (Via, et al. 2004) also proposes a procedure aimed at exploiting structural information to optimize PROSITE sequence motif accuracy.

Figure 3 - Serine protease catalytic 3D motif representation in Catalytic Site Atlas. The trypsin catalytic site 3D motif found in PDB:1A0J is represented as a list of the three catalytic residues (His, Asp, Ser), their position number in the PDB structure and in the Uniprot sequence. The structure unit (chain) of the catalytic residues is also reported.

All these findings not only support the idea that 3D information might be used to improve the performance of sequence motif searches but also that it is possible to use sequence patterns to identify structural templates from the analysis of which extra features, distinctive of the function, can be extracted.

In contrast, Mondal et al (Mondal, et al. 2003) propose an idea that might have an important impact in the field of 3D motifs: the structural shape of a 3D motif may hinge upon the protein context, such as its interaction with a ligand or with another macromolecule. These authors found indeed different true hits of PROSITE patterns displaying significantly different local conformations depending on the context. Therefore, they recommend taking into account the possibility of context-based backbone conformational changes when creating 3D templates for PROSITE patterns.
This leads to consider a more general issue: Many algorithms for either 3D motif discovery or structure-based function prediction try to identify sets of residues displaying reciprocal spatial conservation in the crystallographic structure of different proteins sharing a biochemical function. However, these algorithms do not consider all the potential conformational states of a functional surface region. This implies that a 3D template, being only representative of a “frozen conformation”, would generally miss the information inherent in other conformations of the same functional site.

Other important tools are those aimed at mapping and visualizing sequence motifs (e.g. PROSITE patterns) in 3D. As such, they are very useful in structural motif analyses and can be actually used to discover novel 3D motifs, but they do not play an instrumental role in 3D motif discovery. These approaches will be described in Section 4.

The identification of 3D motifs based on mapping sequence motifs of known function on protein structures has advantages and limits. The main advantages are that, in this case, the functional annotation already exists (by definition) and that the identified local structure can be used both to analyze the biochemical details of the associated function and to scan a dataset of uncharacterized structures in search for similar local configurations. Clearly, though, the main limit of this approach is that it basically consists in transferring annotations and therefore will not discover genuinely novel 3D motifs.

3D motif detection based on local structure similarities. Most of the available programs for 3D motif discovery take advantage of functional site evolutionary conservation principles – also exploited in the field of sequence motif discovery. The underlying assumption is that proteins with common functional properties, i.e. common interacting partners or binding ligands, will share common 3D functional motifs. Therefore, several groups have developed comparison algorithms in order to detect local similarities in completely different protein structures. Basically, these algorithms can differ in two aspects: the protein structure representation and the computational procedure for similarity detection. The structure representations may range from very approximate (e.g. Ca atoms) to very detailed ones (e.g. functional atom positions labelled with partial charge or atom types (Hoffmann, et al.)). Notice that a more elaborate view of the protein structure is not necessarily an advantage. In some cases, a more approximate description might better take into account the different conformational states (e.g. bound/unbound) of the protein. Computational procedures can be based on different search strategies, such as geometric hashing or subgraph isomorphism. A complete review of 3D comparison methods is out of the scope of this chapter, and interested readers may consult (Gherardini, et al. 2008) for a detailed review of the topic.

A special case is represented by 3dLOGO (Via, et al. 2007), a Web server for the identification, analysis and use of conserved protein substructures. A sequence logo is a graphical representation of an MSA which provides an instant way of visualizing variable and conserved positions of the MSA. In the sequence logo, the logo is constructed by calculating the information content of each position of the aligned sequences, and then displaying the characters representing the amino acids stacked on top of each other. The height of each letter is made proportional to its frequency. The height of each stack is then adjusted according to the information content at that position, as detailed in (Schneider, et al. 1990). 3dLOGO is a resource specifically focusing on the identification and improvement of 3D functional consensi, given a set of structures and a number of specified residues in at least one of them (i.e. a substructure). First, 3dLOGO detects the substructures common to the set of input structures and performs a 3D alignment using the substructures’ residue coordinates. This procedure results in a 3D consensus, i.e. a 3D motif, formed by residues structurally conserved in all the input structures. Two amino acids belonging to different structures are assumed to be conserved if they are spatially well superimposed and display physico-chemical similarities, according to a substitution matrix. A 3D motif is presented in the form of a tabular view, which also displays other similar and well-superimposed amino acids. This implies that, if additional conserved residues exist in the region surrounding the original substructure, they can be used to build a larger 3D motif. 3dLOGO also provides a tool to visualize the 3D motif in the context of the structures, and a Java application (3dProLogo) that shows a three-dimensional view of the 3D motif sequence logo (Schneider, et al. 1990). An example of 3dLOGO output page is shown in Figure 4. Finally, 3dLOGO makes it possible to use the identified structurally conserved residues to build a sequence consensus,
to be used with tools such as ScanProsite for sequence database searches. Notice that sequence motif consensi would greatly improve their function prediction specificity if they can be optimized by exploiting structural conservation information (Via, et al. 2004).

Figure 4 – The 3dLOGO (http://3dlogo.uniroma2.it/) output page. The input structures are those of trypsin and subtilisin of figure 1; the superimposition has been carried out on their catalytic residues. Notice that, besides the catalytic His, Asp and Ser, eight additional residue pairs are found to be conserved in the catalytic triad surroundings. The inset on the left reports the AstexViewer (Hartshorn 2002) view of the structural superimposition, whereas the inset on the right shows a three-dimensional view of the 3D motif sequence logo; residue numbers are ordered by PDB:1FXY structure.

3dLOGO has some limitations: the analysis of a large number of protein structures makes the motif identification very time consuming and the position of a motif in at least one input structure must be provided. This implies that the method is not applicable if no motif information for at least one protein structure is available. Moreover, the source code cannot be downloaded and some steps of the procedure require manual intervention preventing the automation of annotation procedures based on this tool.

In the context of computer vision, various 3D shape-searching techniques have been developed and applied to the search of local similarities between protein structures. Computer vision is a discipline that extracts information and produces 3D models from images. Visual-based methods are extensively used to find similarities between protein structures automatically. See, for example, (Ballester, et al. 2007) or (Daras, et al. 2006), and, for a detailed review, see (Iyer, et al. 2005). Notice that 3D shape comparison methods implicitly
redefine 3D motifs in terms of their geometrical shape, generally using geometry-based descriptor vectors.

All methods based on local structural comparison make it in principle possible to identify truly novel 3D motifs and, in particular, to highlight cases of convergent evolution (Ausiello, et al. 2007, Gherardini, et al. 2007), which cannot be easily detected by other existing strategies. However, their success strongly depends on the chosen protein structure representation and on the specific features of the implemented algorithm. Moreover, when a candidate motif is found, its role in a biological function has to be further verified and manual inspection and/or experimental validation is required.

The above-mentioned algorithms in principle allow the detection of any kind of structural similarity. If used for all-against-all searches in whole protein structures, they might also find, for example, elements of secondary structure or other structural motifs. In order to avoid this, and limit the search to 3D motifs encoding a biochemical function, many authors restricted the domain of application of their algorithms to specific regions of protein structures, including protein surfaces and/or cavities, ligand binding sites, protein-protein interaction interfaces, domain-peptide binding regions.

**Protein surfaces** Computer vision-based methods are explicitly developed to compare protein surfaces. However, more traditional methods exist that limit their searches to protein surfaces. Ferré and collaborators (Ferre, et al. 2005), for example, employ a sequence-independent algorithm to exhaustively compare functional surface regions with a collection of protein surface cavities. Binkowski and Joachimiak (Binkowski, et al. 2008) describe a methodology for evaluating the similarity between a pair of surfaces. The methodology is used to search the Global Protein Surface Survey (GPSS), a library of annotated surfaces derived from structures in the PDB. Our group developed a methodology (Sanchez, personal communication) that uses the FunClust algorithm (Ausiello, et al. 2008) to identify structural motifs on the surface of a set of non-homologous proteins all interacting with a “hub” protein. A survey of methods to describe and compare protein surfaces is reported in (Via, Ferre, Brannetti & Helmer-Citterich 2000).

**Ligand binding sites** Many authors try to identify 3D motifs as sets of protein amino acids that occur in more than one protein fold and interact with a common ligand molecule. It is worth noting that ligand-binding 3D motifs provide structural basis for rational drug design on target proteins of medical relevance. Most of the methods for ligand-binding site detection first identify regions in diverse proteins that are likely to interact with a specific ligand and next search for common structural features shared by these regions. The Protein Data Bank (Berman, et al. 2000) stores the coordinate files of many protein/ligand co-crystals. This makes it possible to easily identify amino acids that are within a certain distance from a specific ligand. Generally, ligand-binding regions are identified on the basis of distance requirements from the atoms of the ligand. A 3D ligand-binding motif can be then extracted from (geometrical, physico-chemical, etc.) traits common to regions belonging to different proteins that interact with the same ligand.

Nebel et al (Nebel, et al. 2007), for example, identify ligand-binding sites by superimposing the corresponding ligands. The similarity between ligand binding regions is then evaluated by calculating the number of atoms of the same type which share equivalent spatial positions. Clusters of similar sites are used to produce consensus 3D patterns.

Hoffmann and collaborators (Hoffmann, et al.) represent ligand binding pockets by clouds of atoms and identify similarities between two pockets by aligning their atoms in the 3D space.

For reviews on methods for the detection of protein-ligand binding sites see, for example, (Leis, et al.) and (Laurie, et al. 2006).

Many methods focus on specific classes of ligands, such as nucleotide phosphate (Brakoulas, et al. 2004); ATP, CoA, NAD, NADP, and FAD (Denessiouk, et al. 2001), adenine and AMP (Denessiouk, et al. 2000), magnesium (Dudev, et al. 2007), etc.

**Protein-protein and protein-DNA interfaces** The principles underlying methods for the discovery of 3D motifs involved in protein-protein and protein-DNA interactions are similar to those for the discovery of ligand-binding 3D motifs: protein-protein and protein-DNA interfaces can easily be identified from the PDB and compared in search of structural similarities.
The application of local structural similarity search algorithms to specific regions of proteins, where functional residues are more likely to be found, represents one of the most effective strategies for 3D motif discovery. In fact, this combined approach restricts the search space, thus overcoming speed problems and permits to easily associate a potential function to a newly detected motif. The main disadvantage of this type of methods is that they are generally calibrated on specific classes of functional sites (ligand binding sites, protein-protein interaction and domain-peptide binding residues, etc.) and therefore have a limited range of application. Methods described in this section are listed in Table 1.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Type of prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Via, et al. 2007)</td>
<td>3dLOGO: identification, analysis and use of conserved protein substructures</td>
</tr>
<tr>
<td>(Ferre, et al. 2005)</td>
<td>identification of protein surface similarities</td>
</tr>
<tr>
<td>(Binkowski, et al. 2008)</td>
<td>identification of protein surface similarities</td>
</tr>
<tr>
<td>(Nebel, et al. 2007)</td>
<td>identification of ligand-binding sites</td>
</tr>
<tr>
<td>(Hoffmann, et al.)</td>
<td>identification of similarities between two ligand binding pockets</td>
</tr>
<tr>
<td>(Brakoulias, et al. 2004)</td>
<td>identification of nucleotide phosphate binding sites</td>
</tr>
<tr>
<td>(Denessiouk, et al. 2001)</td>
<td>identification of ATP, CoA, NAD, NADP, and FAD binding sites</td>
</tr>
<tr>
<td>(Denessiouk, et al. 2000)</td>
<td>identification of adenine and AMP binding sites</td>
</tr>
<tr>
<td>(Dudev, et al. 2007)</td>
<td>identification of magnesium binding sites</td>
</tr>
</tbody>
</table>

Table 1: Methods for 3D motif discovery

3 Databases and Other Resources

As mentioned in the Introduction of this chapter, one important void in the structural motif research field consists in the lack of a comprehensive repository of 3D functional motifs, conceptually similar to PROSITE (Hulo, et al. 2006), ELM (Gould, et al.), MnM (Balla, et al. 2006) for sequence motifs. Resources devoted to the collection and analysis of specific classes of 3D functional motifs are very limited in number, and spread in various laboratories. This represents a particularly serious problem in the post-genomic era, where a large number of macromolecule structures belong to proteins of unknown function and 3D motifs could be successfully used in structure-based function prediction (see Section 5).

The disadvantage of specialized resources is that they do not share a unique representation for structural motifs, making it difficult to compare them or to use similarity search algorithms to search them. In contrast, ELM and MnM motifs are encoded in regular expressions as well as most PROSITE patterns. Moreover, these three resources provide tools to scan a sequence against its patterns or, in the case of PROSITE, to search a given pattern in the whole UniProtKB database (Uniprot, et al. 2010).

Similar “all inclusive” resources do not exist for 3D motifs and we hope that, in the future, the scientific community will invest resources in this direction.

Some specialized repositories for 3D motifs and structural functional sites are described in the following.

The Catalytic Site Atlas (Porter, et al. 2004), mentioned in Section 2.2.1, is a database collecting enzyme active sites and catalytic residues in enzymes of 3D structure, i.e. enzymatic 3D motifs. Each CSA motif consists of the catalytic residues found in a catalytic site, using PDB residue numbering. The database can be searched by PDB identifier, Swiss-Prot code or EC number.

If we consider functionally annotated important residues in protein structures as rudimentary 3D functional motifs, we can regard FireDB (Lopez, et al. 2007) as a comprehensive and detailed repository of 3D motifs.

The SURFACE database is a repository of annotated and compared protein surface regions (Ferre, et al. 2004) that the authors call “patches”. A patch, which can be assimilated to a 3D motif, is defined as the set of
residues surrounding a surface cavity of a protein structure. Patches are annotated for the presence of a PROSITE or ELM pattern hit and/or for binding a ligand (binding is defined on the basis of the distance from the atoms of the ligand). A procedure (Ausiello, Via, et al. 2005) for structure comparison is used to perform an all-versus-all patch comparison.

Phospho3D (Zanzoni, et al. 2007) collects 3D structures of phosphorylation sites and their surrounding “3D zone”, i.e. the 3D region defined by the set of residues at a distance not exceeding 12 Å from the phosphorylation site. A zone is essentially a 3D phosphorylation motif. In the latest version of Phospho3D (Zanzoni, et al. 2011) the authors provide a tool to search the 3D zones in a user-provided structure. Even though Phospho3D is a specialized resource, it is conceptually designed as PROSITE, i.e. as a collection of functional motifs associated with a tool for searching the motifs in proteins (sequence in the case of PROSITE and structure in the case of Phospho3D).

MODBASE (Pieper, et al. 2009) is a database of annotated comparative protein structure models. However, it also provides integrated resources such as LigBase (Stuart, et al. 2002), which comprises all ligand-binding sites of known structure aligned with all related protein sequences and structures, or AnnoLyze, a collection of predicted ligand binding sites. Sequence and structure alignments provide clues to assess evolutionary site conservation. LigBase conserved sites can be considered as 3D ligand binding motifs.

Databases and other resources for 3D motif analysis are being made increasingly available even if not yet as systematically as for sequence functional motifs.

Databases and resources described in this section are listed in Table 2.

<table>
<thead>
<tr>
<th>Resource name</th>
<th>Reference</th>
<th>Contents</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catalytic Site</td>
<td>(Porter, et al. 2004)</td>
<td>enzyme active sites and catalytic residues in enzymes of 3D structure, i.e.</td>
<td><a href="http://www.ebi.ac.uk/thornton-srv/databases/CSA/">http://www.ebi.ac.uk/thornton-srv/databases/CSA/</a></td>
</tr>
<tr>
<td>Atlas (CSA)</td>
<td></td>
<td>enzymatic 3D motifs</td>
<td></td>
</tr>
<tr>
<td>FireDB</td>
<td>(Lopez, et al. 2007)</td>
<td>functionally annotated important residues in protein structures</td>
<td><a href="http://firedb.bioinfo.cnio.es/">http://firedb.bioinfo.cnio.es/</a></td>
</tr>
<tr>
<td>SURFACE</td>
<td>(Ferre, et al. 2004)</td>
<td>annotated and compared protein surface regions</td>
<td><a href="http://cbm.bio.uniroma2.it/surface">http://cbm.bio.uniroma2.it/surface</a></td>
</tr>
<tr>
<td>Phospho3D</td>
<td>(Zanzoni, et al. 2007)</td>
<td>3D structures of phosphorylation sites and their surrounding “3D zone”</td>
<td><a href="http://www.phospho3d.org/">http://www.phospho3d.org/</a></td>
</tr>
<tr>
<td></td>
<td>(Zanzoni, et al. 2011)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LigBase</td>
<td>(Stuart, et al. 2002)</td>
<td>ligand-binding sites of known structure aligned with all related protein</td>
<td><a href="http://modbase.compbio.ucsf.edu/ligbase/">http://modbase.compbio.ucsf.edu/ligbase/</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>sequences and structures</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Resources and databases for 3D motifs

4  Annotation and Analysis of Structural Motif

Once a set of conserved residues in non-homologous protein structures has been identified, it is important to have instruments to annotate and analyze it. In many cases, the function related to a newly discovered motif is already known. For example, if a motif is identified by comparing or superimposing ATP binding pockets, it is likely that it will be related to ATP binding. In other cases, a 3D motif may result from an all-versus-all
comparison of two or more complete protein structures and it is not always possible to associate a function to it. For example, the FunClust (Ausiello, et al. 2008) web server for the identification of structural motifs in a set of non-homologous protein structures, basically finds local structural similarities without associating them with a biochemical function.

Fortunately, the bioinformatics community has developed several tools for the analysis and visualization of 3D motifs, thus facilitating their functional annotation.

An important class of tools for the analysis of 3D motifs is represented by software products that have been developed to visualize and explore sequence motifs in known protein structures.

A precursor of structural representations of conserved sequence data is the software JOY (Mizuguchi, et al. 1998), which annotates protein sequence alignments with three-dimensional structural features. JOY allows a pattern of structural characteristics (e.g. accessibility to the solvent, secondary structure, etc.) to be assigned to conserved regions of a MSA, but does not provide 3D visualization of the pattern.

In 2001 PDBsum (Laskowski 2001) introduced a new package for the visualization of PROSITE patterns on structures. Matching residues are coloured according to their conservation (and hence importance) and can be viewed in RasMol (Sayle, et al. 1995). The package, gives the opportunity to explore the structural and functional significance of the PROSITE pattern residues.

3MOTIF and 3MATRIX (Bennett, Lu, et al. 2003) (Bennett, Nevill-Manning, et al. 2003) are two tools belonging to a web-based visualization system for displaying sequence motif information in its 3D context. 3MOTIF accepts a sequence motif in the form of a regular expression and 3MATRIX accepts a position-specific scoring matrix, and search them in the PDB for matches. Structural hits are then visualized in three dimensions, together with the sequence conservation of residues and the solvent accessible surface area. SeeMotif (Chang, et al. 2009) is conceptually similar to 3MOTIF, i.e. it provides an interactive interface for visualizing sequence motifs in PDB protein structures.

MSDmotif (Golovin, et al. 2008) is a component of the Macromolecular Structure Database (MSD) (Boutselakis, et al. 2003) aimed at querying the PDB for binding sites, small 3D and sequence motifs and at exploring them. The MSDmotif web interface provides functionality for visualization, data integration, search criteria creation and sequence and structural multiple alignment options. This resource can also be used to discover novel 3D motifs.

Together with specialized resources, any molecular visualization system, allowing the view of protein structures, can be used to analyze known 3D motifs, their location in the protein, the side-chain orientation of their residues, their spatial neighbourhoods, and the charge properties of their environment. PyMol (http://www.pymol.org/), Jmol (http://jmol.sourceforge.net/) and RasMol (http://rasmol.org/) are well-known examples.

Moreover, any system for the annotation of protein structures will turn to be effective in 3D motif annotation. Two interesting examples are COLORADO3D (Sasin, et al. 2004) and FeatureMap3D (Wernersson, et al. 2006).

COLORADO3D (Sasin, et al. 2004) is not a viewer of 3D protein structures; however, given a PDB input file, it identifies, among other properties, residue evolutionary conservation and returns a PDB-formatted file, where the B-factor values are replaced by conservation values. The output file can be easily displayed with structure viewers such as RasMol in order to visualize the conservation of residues that belong, for example, to a 3D motif. FeatureMap3D (Wernersson, et al. 2006) is a web-based application that maps protein features onto protein structures. Starting from pre-annotated sequences, FeatureMap3D maps the sequence features, for example annotated glycosylation or phosphorylation sites, onto the 3D structure of homologous proteins. This tool can be easily used for 3D motif annotation. Finally, SPICE (Prlic, et al. 2005) is a tool that can be used to visualize protein sequence and structure annotations as well.

Tools described in this section are reported in Table 3.
<table>
<thead>
<tr>
<th>Tool</th>
<th>Reference</th>
<th>Function</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>JOY</td>
<td>(Mizuguchi, et al. 1998)</td>
<td>annotation of protein sequence alignments with three-dimensional structural features</td>
<td>not available</td>
</tr>
<tr>
<td>PDBsum</td>
<td>(Laskowski 2001)</td>
<td>(among others) visualization of PROSITE patterns on structures</td>
<td><a href="http://www.ebi.ac.uk/pdbsum/">http://www.ebi.ac.uk/pdbsum/</a></td>
</tr>
<tr>
<td>MSDmotif</td>
<td>(Golovin, et al. 2008)</td>
<td>querying the PDB for binding sites, small 3D and sequence motifs and at exploring them</td>
<td><a href="http://www.ebi.ac.uk/pdbe-site/pdbemotif/">http://www.ebi.ac.uk/pdbe-site/pdbemotif/</a></td>
</tr>
<tr>
<td>PyMol</td>
<td>-</td>
<td>3D structure visualization system</td>
<td><a href="http://www.pymol.org/">http://www.pymol.org/</a></td>
</tr>
<tr>
<td>Jmol</td>
<td>-</td>
<td>open-source Java viewer for chemical structures in 3D</td>
<td><a href="http://jmol.sourceforge.net">http://jmol.sourceforge.net</a></td>
</tr>
<tr>
<td>RasMol</td>
<td>(Sayle, et al. 1995)</td>
<td>3D structure visualization system</td>
<td><a href="http://rasmol.org/">http://rasmol.org/</a></td>
</tr>
</tbody>
</table>

Table 3 – Tools for annotation, visualization and analysis of structural motifs

5 3D Motifs in Structure-Guided Function Inference

Some might consider that 3D motif discovery and function prediction are one and the same thing. This would be true if we assumed that the identification of a 3D motif corresponds to the identification of the protein function. However, even though 3D motifs and protein functions are strictly correlated, they are not the same: a 3D motif is a set of local structural features capturing a specific molecular function. As such, it can be found in non-homologous protein structures sharing a function. On the other hand, whereas it would in principle be possible that a biochemical function exists that cannot be ascribed to an identifiable set of local structural features (i.e. a specific 3D motif), the contrary is indeed impossible: a 3D motif must be associated with a function by definition, otherwise it would be just a recurring sub-structure of a protein (e.g. a helix) without a specific functional meaning. Notice that, in the field of sequence motifs, motif discovery and function inference are generally based on different approaches. Similarly, here, we assume that the discovery of a 3D motif consists in identifying one or more structural features that are conserved in distinct proteins and encode for a specific biochemical function, whereas the structure-guided function inference involves computational procedures developed for determining the presence of characterized 3D motifs in protein structures the function of which is unknown. The function can then be inferred by transferring the annotation of the 3D motif to the uncharacterized protein.

Structure-based methods for predicting function have recently become important particularly in the context...
of structural genomics initiatives, which are determining an increasing number of protein structures for which little or no functional information is available (Friedberg 2006). This is an intrinsic consequence of the structural genomics initiatives most of which, in order to have the largest sampling of the structure space, tend to select their structure determination targets among proteins with very low sequence identity to proteins of known structure.

Methods that use pre-defined 3D templates to predict the function(s) of still uncharacterized proteins, are essentially based on comparative approaches, i.e. on procedures aimed at identifying local similarities between the original template and a protein structure. In particular, a dataset of known 3D motifs can be searched in few or many protein structures, which is usually a time consuming task; search algorithms, moreover, must be designed taking into account the structure representation adopted to describe the query motif(s) and the target protein structure(s). The most important issue, however, concerns the assessment of the statistical and biological significance of the resulting 3D hits. Since, as it will be discussed in Section 6, it is difficult if not impossible to assess the statistical significance of 3D matches, it is crucial to integrate similarity search algorithms with functional annotation. For example, in the PdbFun web server (Ausiello, Zanzoni, et al. 2005) users can select a set of residues in a protein structure by combining a number of features, including solvent exposure, ligand binding ability, location in a protein cavity, secondary structure, residue type, sequence functional pattern, protein domain and catalytic activity. Selections can be used as probe and target in multiple structure comparison searches. The annotated set of residues can then be used as probe in multiple structure comparison searches for structurally related and similarly annotated targets.

The new version of Phospho3D (Zanzoni, et al. 2011) implements P3Dscan, a tool based on a similarity search algorithm (Gherardini, et al. 2010), that allows users to compare a PDB structure against the dataset of 3D zones collected in the Phospho3D database. In order to obtain as much as possible reliable results, the search parameters are rather restrictive: only hits that display an RMSD lower than 0.7 Å and very similar physicochemical properties to the query 3D zone residues are retained. However, in principle, this does not ensure that the detected hit is a true positive one, i.e. a real phosphorylation site. In order to help users to evaluate the biological and structural context of each match, P3Dscan also supplies the resulting hits with a number of annotations at the residue level (solvent accessibility, secondary structure, occurrence in a cavity, etc), which can be compared with the query 3D zone residue features: it is in fact reasonable to expect that a true positive hit will occur in a structural context similar to that of the original 3D zone. Furthermore, the structural superimposition of query and target 3D motifs, in the context of their entire structures, can be visualized using Jmol. These functionalities can help users assess the biological plausibility of a match, but do not tell much about its statistical significance.

PdbFun relies on the Query3d program for local structural comparison (Ausiello, Via, et al. 2005) and Phospho3D on Superpose3D (Gherardini, et al. 2010). Due to the difficulty of statistically assessing 3D motif matches, neither Query3D nor Superpose3D provide the general performance of the models in terms of statistical parameters such as sensitivity and specificity. As for Query3D, a number of successful examples have been reported, whereas for Superpose3D the authors showed that, by following their guidelines for interpreting the results, scanning a database of structures using the p-loop residues of H-RAS as probe can reach an AUC value of 0.87. Thus, in the case of PdbFun and Phospho3D, the discrimination between true and false positives results from a filtering procedure based on the functional annotation of motif matches rather than from a statistical assessment.

A class of methods for structure-based function prediction includes procedures based on the observation that functional regions have specific physicochemical characteristics, such as being located in a cavity, being in evolutionary conserved areas, having pKa values that differ from the standard values in solution, etc. Examples are ProMate (Neuvirth, et al. 2004), THEMATICS (Wei, et al. 2007), ConSurf (Armon, et al. 2001), HotPatch (Pettit, et al. 2007). The type of predicted sites by these methods and their performance are briefly summarized in Table 4. These methods do not make use of 3D motifs and, therefore, will not be described in this chapter. For a review see (Gherardini, et al. 2008).
### Table 4 – Methods for structure-guided function inference that are not based on local structural comparison

<table>
<thead>
<tr>
<th>Method name</th>
<th>PMID</th>
<th>Type of predicted site</th>
<th>Rate of success (%)</th>
<th>Test dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProMate</td>
<td>15050833</td>
<td>protein-protein interface</td>
<td>~70</td>
<td>67 structures of transient protein-protein heterodimers</td>
</tr>
<tr>
<td>THEMATICS</td>
<td>17419878</td>
<td>enzyme active sites</td>
<td>86(^1) 93(^2)</td>
<td>169 enzymes from CatRes/CSA database (Bartlett, et al. 2002, Porter, et al. 2004)</td>
</tr>
<tr>
<td>ConSurf</td>
<td>11243830</td>
<td>evolutionary conserved residues on the surface of proteins</td>
<td>good on the examples reported</td>
<td>Src SH2 domain She PTB domain</td>
</tr>
<tr>
<td>HotPatch</td>
<td>17451744</td>
<td>surface patches of unusual physicochemical properties (virtually any kind of functional site)</td>
<td>60-80(^3)</td>
<td>618 proteins of diverse mixed functions</td>
</tr>
</tbody>
</table>

6 **Statistical Significance of 3D Motif Occurrences**

Different authors have successfully tackled the problem of determining the statistical significance of linear motif occurrences in protein sequences. The statistical framework, in this case, relies on three basic observations: 1) it is possible to generate random models that can be used as a basis for significance assessment; 2) sub-sequences that are overrepresented in a given set of protein sequences with respect to a background set, are likely to encode a functional property; 3) datasets of true positive instances and false positive hits exist that can be used to assess motif performance. Statistical analysis of three-dimensional motif occurrences in protein structures is still in its infancy when compared with its analogue in sequence pattern matching. However, there are good reasons why this happens. We will discuss each of these three observations above in relation to 3D motif occurrences in protein structures. 1) Random models can be either simulated or determined analytically. In both cases, the idea is that the model should be able to generate sequences with no biological meaning but with the same constraints and amino acid composition of the biological ones. For protein sequences, this can be done in several ways, for example by reversing the original sequences or by reshuffling them. However, the definition of ‘random’ in the case of protein structures is clearly problematic due to the higher dimensionality of the problem: not only protein reshuffling in the 3D space would not be a trivial operation, but the generation of a random structure with physicochemical constraints similar to those of a native structure might turn out to be very difficult if not impossible. 2) The most over-represented substructures in a given set of structures are expected to be helices, sheets, or supersecondary structures. The occurrence of these structural elements may be indicative, for example, of a similar fold, but do not necessarily reflect their role in a specific biochemical functions. This implies that, being the occurrence of secondary structure elements in general more statistically significant than the occurrence of sets of functional residues, we can expect that the former will mask the latter. On the other hand, while it makes sense to explicitly mask globular domains or low complexity regions in protein sequence pattern matching (see for example (Neduva, et al. 2005), masking secondary structure elements in 3D similarity searches would lead to senseless results. 3) In order to evaluate the performance of a motif in function prediction, datasets of true and false positives are needed. As stated above, several collections of true positive instances of sequence motifs are available in the literature and sets of false positives can be built,  

---

\(^1\) using only CatRes/CSA annotation as reference set.  
\(^2\) using an expanded reference set which includes literature annotation.  
\(^3\) depending on the specific class of functional sites considered (enzyme active sites, DNA-RNA-interacting sites, small-molecule interacting sites, etc.). For generic functional sites the rate of success is >63%.
Protein Structural Motifs: Identification, Annotation and Use in Function Prediction

even if with some limits and variable reliability. This is not true in the case of 3D motifs, for which large benchmarking datasets essentially do not exist.

In spite of the mentioned difficulties, different authors tried to solve the problem of the statistical significance of 3D motif occurrences using either empirical or semiempirical methods. For example, a universal similarity measure for comparing protein structures has been proposed by Betancourt and Skolnick (Betancourt, et al. 2001); these authors argue that one of the most commonly used measures to detect structural similarities between proteins, the root mean square distance (RMSD) of their corresponding atoms, is dependent on the protein structure size. Therefore they introduce the RRMSD, a measure independent of the protein size, as a dimensionless variant of the RMSD for comparing protein structures. RRMSD is defined as the ratio between the RMSD and the average RMSD between two random polypeptides, with sizes equivalent to the structures being compared, whereas Stark et al (Stark, et al. 2003) introduced a model that is specific for the statistical significance of local similarities in protein structures. As in the case of sequence comparison, the smaller the number of considered amino acid, the higher the likelihood that a similarity occurs by chance (i.e. that the similarity is not biologically meaningful). Therefore, any score used to compare protein structures must be considered in the light of its statistical significance. On this basis, Stark et al developed a statistics to calculate the significance of RMSD between sets of residues in 3D space. To this aim, they built a geometrical model for estimating a priori the significance of a local similarity with a given RMSD, without requiring a background set of structures. In particular, the model provides a p-value for any RMSD observed for a query pattern. This p-value depends on a number of parameters (e.g. the number of residues composing the pattern, their abundance in the target database, etc.) and on some empirically determined constants.

Barker and Thornton (Barker, et al. 2003) derived an empirical measure for estimating the relative significance of hits obtained searching differing 3D motifs in a population of structures. In fact, if two or more 3D motifs with differing specificity and sensitivity are used, the score (e.g. RMSD) of their hits should be normalized in order to compare them and identify the best one. To this aim, these authors constructed a reference population of structures over which each of the 3D motifs is calibrated. The heuristic measure of significance for a match of a 3D motif at a given RMSD \( r \) is empirically calculated as the expected number of 3D sites found matching the motif within \( r \). Expectations are computed from the model distributions; these are obtained by fitting normal distributions to cumulative frequency data collected probing the reference population with 3D motifs. Moll et al (Moll, et al. 2010) use a nonparametric model to compute the statistical significance of a 3D motif match. This model, introduced by Fofanov and colleagues (Fofanov, et al. 2008), assumes that the matching algorithm returns for each target only the lowest RMSD complete match to a 3D motif and computes the exact p-value of matches with RMSD less than \( \varepsilon/(n)^{1/2} \), where \( \varepsilon \) is an RMSD cut-off parameter and \( n \) is the number of residues composing the 3D motif.

These and other existing approaches (see also (Davies, et al. 2007, Xie, et al. 2009)) introduced many interesting ideas in the field to face the problem of the lack of a rigorous statistics for protein structure comparison. However, the models introduced by these statistical approaches are calibrated taking into account specific issues such as the comparison algorithm, the 3D structure representation and structure datasets. Therefore, it is not clear whether they can be universally applied regardless of the specific comparison algorithm, 3D motif representation and dataset(s) used.

We can conclude that the problem of assessing the statistical significance of local structural similarity has not been yet completely solved and more systematic and solid new strategies are desirable.

7 Conclusions

One of the main goals of computational biology is the assignment of function to unknown biological entities, mostly proteins. This can often be achieved at the level of amino acid sequence, but the amino acid sequence of a protein is only a proxy for its structure which is the characteristic of a protein more closely related to function.
The success of sequence motif identification for function assignment has relied on the availability of organized collections, often manually validated, of known functionally important regions, on the ability of comparing sequence strings and on a reliable statistical analysis of the results. None of these aspects has so far been satisfactorily addressed as far as three-dimensional motifs are concerned. As we described in this chapter, in the latter case data are scarce and not well organized, the comparison problem is much more complex and the derivation of a baseline random model still problematic.

Nevertheless, several very useful tools are available and consistently used by computational biologists. Some specialized ones have also encountered the favour of experimental scientists, but an effort should be made in providing the whole collection in an organized and easy-to-use fashion, in integrating the diverse set of search and visualization tools and in providing users with more rigorous and intuitive measures of the reliability of their results.

Acknowledgement

This work was supported by FIRB 2004 ITALBIONET and Fondazione Roma.

References


