Towards a Structural Classification of Phosphate Binding Sites in Protein–Nucleotide Complexes: An Automated All-Against-All Structural Comparison Using Geometric Matching

Andreas Brakoulias[1] **and Richard M. Jackson**[2*]
[1] *Department of Biochemistry & Molecular Biology, University College London, Gower Street, London, UK*
[2] *School of Biochemistry & Molecular Biology, University of Leeds, Leeds, UK*

ABSTRACT **A method is described for the rapid comparison of protein binding sites using geometric matching to detect similar three-dimensional structure. The geometric matching detects common atomic features through identification of the maximum common sub-graph or clique. These features are not necessarily evident from sequence or from global structural similarity giving additional insight into molecular recognition not evident from current sequence or structural classification schemes. Here we use the method to produce an all-against-all comparison of phosphate binding sites in a number of different nucleotide phosphate-binding proteins. The similarity search is combined with clustering of similar sites to allow a preliminary structural classification. Clustering by site similarity produces a classification of binding sites for the 476 representative local environments producing ten main clusters representing half of the representative environments. The similarities make sense in terms of both structural and functional classification schemes. The ten main clusters represent a very limited number of unique structural binding motifs for phosphate. These are the structural P-loop, di-nucleotide binding motif [FAD/NAD(P)-binding and Rossman-like fold] and FAD-binding motif. Similar classification schemes for nucleotide binding proteins have also been arrived at independently by others using different methods. Proteins 2004;56:250–260.** © **2004 Wiley-Liss, Inc.**

Key words: structural bioinformatics; structure comparison; P-loop; di-nucleotide binding motif

INTRODUCTION

The determination of protein function from sequence and structure is a major goal of bioinformatics techniques. In many cases sequence patterns are sufficient to identify function. The idea of using sequence patterns and profiles underlie the PROSITE,[1] PRINTS,[2] and Pfam[3] methodologies. However, structure can often provide a more powerful evolutionary link than sequence alone. The structural classification schemes of protein families (e.g., the CATH,[4] SCOP,[5] DALI[6] databases) have proved invaluable tools for structural biologists often allowing functional inferences to be made between members of a family even when sequence similarity is statistically low. Furthermore, the incorporation of both sequence and structure into functional pattern recognition has been used to identify the recurrence of short motifs (often in different folds) that infer function, such as the phosphate binding P-loop[7,8] or the DNA binding helix-hairpin-helix.[9,10] Indeed, Kasuya and Thornton identified maintenance of three-dimensional (3D) structure in PROSITE patterns, suggesting that many sequence motifs do infer common function.[11]

Function may also be predicted from recurrence of side-chain constellations.[12–14] These constellations usually contain residues that are of key importance in catalysis, binding or covalent attachment of enzyme cofactors. Currently work is in progress on a database of 3D motifs involved in enzyme catalysis.[15] This is being constructed from mechanistic knowledge of the residues involved in catalysis such that a sufficient 3D constellation of atoms can be used to search the protein databank for likely recurrences that will be functionally diagnostic. In other cases sequence conservation has been combined with structure mapping or profiles to generate tools to identify protein surface similarities for the prediction of protein functional sites.[16,17] Recently, conservation in a main-chain pattern in the canonical serine protease inhibitors has been identified where there is no sequence conservation or overall fold conservation.[18] This main-chain fragment was used to search the structural database for other potential canonical serine protease inhibitors. Clusters of hits were found in several extracellular proteins including hydrolases, toxins, cytokines, and viral proteins as well as the known canonical serine protease inhibitors.

One major limitation of all the (sequence, side-chain and main-chain) pattern recognition methods is that the pattern must be predefined. Prior knowledge of the pattern is required in all these methods. This may come from mul-

Grant sponsor: the Wellcome Trust.

*Correspondence to: Richard M. Jackson, School of Biochemistry & Molecular Biology, University of Leeds, Leeds, LS2 9JT, England. E-mail: jackson@bmb.leeds.ac.uk

Received 28 October 2003; Accepted 5 January 2004

Published online 2 April 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20123

tiple sequence alignment of proteins of known function or exploit knowledge of a common 3D motif (such as the phosphate bind P-loop or the Asp-His-Ser catalytic triad of the serine proteases or other catalytic residues identified in the literature). The methods are non-exploratory in as much as the patterns are already known. We can think of an alternative situation for a series of binding sites (in which sites are defined by residues in contact with ligand hetero-atoms or atoms of another protein) in which common structural features are not evident from sequence or global structural similarity. However, they share common atomic level structural features that confer specific ligand (or protein) binding properties yet are in unrelated binding sites. It would be advantageous to exploit these common atomic features of the protein to identify their shared mechanism of interaction with the ligand. With such a method no prior knowledge (i.e., sequence pattern, global structural similarity, or a known 3D motif) would be required to compare any two binding sites. This can be achieved by finding the maximum clique (largest 3D constellation of shared atom similarity) a mathematically defined and operationally independent definition of atomic level similarity.

Here we investigate the construction of a structural classification scheme for protein–ligand binding sites in the large class of nucleotide phosphate-binding proteins in an attempt to quantify common structural features that confer specific ligand binding properties in different binding sites. The method we have used exploits common atomic features through identification of the maximum clique. These features are not necessarily evident from sequence or from global structural similarity giving additional insight into molecular recognition not evident from current sequence or structural classification schemes. Another motivation for this work is that such a classification scheme could be used as a resource for creating structural templates that can be used to screen for binding sites in structurally determined yet functionally unknown hypothetical proteins that will arise from structural genomics initiatives.

Our aim here is to automate the identification of different nucleotide phosphate-binding sites and perform an all-against-all comparison. We have developed a method that uses geometric matching to identify similar 3D substructure in proteins at the atomic level. This produces an atom–atom similarity match for any two binding sites. We use cluster analysis of the resulting matrix of atom matches to define conformational classes for a large set of binding sites. The results allow the identification of distinct structural clusters involving common main-chain and/or side-chain patterns in an automated and unbiased way. One of the major challenges addressed here is the 3D matching of the ~3700 nucleotide phosphate-binding sites against each other. This all-against-all match equates to nearly seven million pairwise comparisons. Therefore the method must be sufficiently fast to allow computation in a reasonable time.

METHODS

Identification and Extraction of Phosphate Binding Sites in Protein–Nucleotide Complexes

All proteins containing a ligand with a phosphate moiety were extracted from the Protein Data Bank (PDB) along with their local protein environment. A program was written to identify structures that contain the PO_4 structural moiety in any ligand (HETATM records) and extract it along with the neighboring protein environment around the PO_4 moiety (ATOM records). There is no appropriate PDB annotation for the identification of a PO_4 moiety. Annotation can be unreliable even for the identification of bound inorganic phosphate (PO4) because the ligand descriptions in the PDB entries often use non-standard nomenclature or are incomplete. To identify the PO_4 moiety we use atomic coordinates. The distance between a ligand phosphorous atom and its neighboring oxygen atoms is calculated to identify covalent linkage. The equilibrium P—O distance is 1.5–1.6 Å for common molecular mechanics force fields used in protein refinement. We use a generous cut-off distance threshold of 2.0 Å for the atoms to be regarded as covalently bonded. A phosphorous atom that has four bonded neighboring oxygen atoms is deemed to be a PO_4 moiety. The local protein environment is defined as the spherical region of 7 Å surrounding the phosphorous atom of PO_4 the moiety (typically ~50–70 atoms). The program distinguishes the local environments according to the ligand the PO_4 moiety belongs to, by consulting the PDB annotation. We extract a subset of 3,737 PO_4 local environments from the binding sites of the nucleotides AMP, ADP, ATP, GDP, GTP, FAD, FMN, NAD, and NADP. These come from 896 protein complexes out of a total of 12,337 X-ray structures. This excludes the "LAYER 1" structures that are not fully validated. At this stage multiple nucleotide ligand binding sites for a particular protein and binding sites of homologous proteins were not removed (see below).

Atom-Level Comparison of Binding Sites Using Geometric Matching

We have written a program similar in concept to geometric hashing algorithms[19,20] from the field of computer vision for finding global structural similarity for two atom constellations. The measure of similarity is based purely on atom level similarity for two proteins. The algorithm proceeds by generating a series of possible triplets or "seed" matches consisting of three atoms forming a triangle. This is carried out for both of the two atomic level objects (the protein binding site in this study). The size and also the number of triangles is controlled by atom–atom distance cut-offs that define the upper and lower boundaries of the sides. This information is stored along with the information about the atom types. For the study here only distinction is made between atom types of carbon, nitrogen, oxygen, sulfur, and phosphorous atoms. For example, any type of carbon atom can be matched to any other. Then a discretized 3D image of the protein (binding site) is created for the object to be matched to (atom-set 1). This acts as a reference frame in which atom identities are stored at indices in a 3D Cartesian grid subject to them being within a certain atom-grid distance cut-off. The resolution of the protein image is controlled by the grid size. The atom-grid distance cut-off controls the focus (i.e., how smeared the atom is over adjacent grid points). Clearly the larger/smaller the grid size the lower/higher the resolution and the larger/smaller the atom-grid distance cut-off the lower/higher the level of focus. The algorithm proceeds by systematically matching all possible triplets between the two objects to be compared (taking into account possible symmetry). However, matching will only proceed given matching atom types at triangle vertices and if the corresponding triangle edge lengths differ by at most δ. These triangle matches are δ-compatible and a transformation is performed for each match using a least squares fitting routine.[21] This results in a translational and rotational matrix that maps one 3D triplet onto another. Unlike geometric hashing there is no pre-processing stage (where a hash table is used to store this information) matching is done "on the fly" and requires little storage. The transformation is applied to the object to be matched (atom-set 2) placing the coordinates in the scene of the discretized atom-set 1. Then for each transformed atom the x,y,z coordinates are converted to a nearest reference grid point which is in turn used as a reference to a one-dimensional array containing the identity of any atom type present at this location in the discretized atom-set 1. Each atom type match is given a score of unity, i.e., the score corresponds to the number of coincident atoms in the mapping. During the matching phase the transformation matrix for the mapping with the highest score is stored to memory. Parameters were chosen to find the optimal clique size (score) such that the root mean squared distance (RMSD) of the superposed atoms is < 1 Å. The all-against-all comparison of N different objects (protein binding sites) produces a matrix with N*(N-1)/2 pair-wise scores. For the 3737 PO_4 local environments this comprises nearly seven million pair-wise comparisons.

In addition, we implemented the maximum clique detection algorithm of Bron and Kerbosch[22] for comparison (result not shown). We optimized the Bron-Kerbosch algorithm for speed of pair-wise comparison for two atom-sets, however, the geometric matching algorithm is two orders of magnitude faster in our hands making it more suited to compute all-against-all comparisons for a large number of binding sites.

Cluster Analysis

The agglomerative means linkage method (UMPG) was used to cluster the matrix of pair-wise similarity scores.[23] Mojena's stopping rules[24] were used to decide at what level the clustering is stopped for ease of analysis. More relaxed criteria (higher K-values) will lead to a few large clusters. We have performed an all-against-all comparison of nucleotide phosphate ligand binding sites involving over 3737 local PO4 environments. Using very strict clustering criteria reduces this to 476 representative environments. The notation used in the results to describe clusters is of the form cluster *X_Y*. This relates to the number of clusters present, *X*, at a particular agglomeration level (e.g., 6 represents the merging of the 476 representatives to six clusters) and the cluster number *Y*. This notation indicates how "early" the cluster is formed in the sense of successive divisions or branching from a common union. Note that this is the reverse of how the agglomerative UMPG algorithm actually works (it performs a series of successive mergers).

Analysis of Binding Site Clusters Using Geometric Matching

Further analysis was carried out for all significant clusters identified by atom-level comparison and cluster analysis. A modified version of the geometric matching algorithm was used. The method performs geometric matching as described above to determine the most representative local environment. This is the one with the highest level similarity as determined by the combined score with all other environments defined in the cluster. This is used as a template on which all other environments are then mapped. Cumulative scores are maintained for the mapping of atom identities to those in the template. The user can apply a percentage cut-off or "focus" value to restrict the atomic output for members of a cluster. A "fuzzy" cluster can be cleaned-up by only allowing atoms which have a corresponding atom identity within say 60% of the representatives of the cluster. This would represent a "focus" value of 60%. A 100% focus value implies that all the representatives of the cluster have the corresponding template atom identity. The aim of this analysis is to clean up the "fuzzy" image where no focus value is applied. This highlights similarities shared across a range of different local environments allowing easy graphical identification of common structural elements in the cluster. Focusing can be used to create a structural template for a given cluster representing the common structural elements. This can be used to define a structural classification scheme or in principle the template can be used to search for binding sites in new protein structures.

RESULTS

We have performed an all-against-all comparison of nucleotide–phosphate ligand-binding sites involving over 3737 local phosphate (PO_4) environments. Using very strict clustering criteria[24] reduces this to 476 representative environments eliminating highly similar local environments (e.g., those belonging to different chains of the same protein or highly similar homologues). The environments were annotated according to the dominant domain involved in the interaction according to the SCOP fold classification[5] and also according to any PROSITE pattern[1] present as predicted with PPSearch.[25] In the later case residues of the pattern must be actually present in binding site environment for the pattern to be counted. At relaxed clustering values fewer clusters are produced and more distant evolutionary relationships become evident. For example, this gives rise to clusters for which the SCOP family may be different. However, there are members of

TABLE I. List of Ten Main Clusters Resulting From the Clustering of 476 Representative Local Environments of Phosphate Binding Sites in Protein-Nucleotide Complexes

Cluster	Environments number (%)	Prevalent SCOP fold (% of cluster)	PROSITE patterns	Ligand
P-loop 4_3	32 (6.7%)	P-loop containing nucleotide triphosphate (97%)	PS00017 (91%)	ADP (47%)
		Phosphoenolpyruvate carboxykinase (3%)	PS00692 (3%)	ATP (31%)
			PS00411 (3%)	GDP (16%)
				GDP (16%)
P-loop 6_6	22 (4.6%)	P-loop containing nucleotide triphosphate hydrolases (95%)	PS00017 (95%)	ATP (82%)
		Phosphoenolpyruvate carboxykinase (51%)		GTP (14%)
				ADP (4%)
P-loop 48_4	24 (5.0%)	P-loop containing nucleotide triphosphate hydrolases (54%)	PS00017 (37%)	ADP (42%)
		GroEL-like chaperones, ATPase domain (8%)	PS00103 (8%)	ATP (21%)
		PRTase-like (8%)	PS00012 (8%)	AMP (17%)
				FMN (8%)
FAD/NAD(P)-binding 4_1	62 (13%)	FAD/NAD(P)-binding domain (81%)	PS00504 (6%)	FAD (89%)
		Nucleotide-binding domain (16%)	PS00225 (6%)	NAD (5%)
			PS00076 (24%)	NAP (3%)
			PS00623 (6%)	
FAD-binding 14_12	13 (2.7%)	FAD-binding domain (85%)	PS00435 (15%)	FAD (85%)
				ADP (15%)
Rossmann fold 29_24	15 (3.1%)	NAD(P)-binding Rossmann-fold (100%)	—	NAP (60%)
				NAD (40%)
Rossmann fold 29_29	21 (4.4%)	NAD(P)-binding Rossmann-fold (90%)	—	NAD (90%)
				NAP (10%)
Rossmann fold 48_30	19 (4.0%)	NAD(P)-binding Rossmann-fold (95%)	—	NAD (79%)
				NAP (21%)
Rossmann fold 48_35	13 (2.7%)	NAD(P)-binding Rossmann-fold (77%)	—	NAD (69%)
				NAP (15%)
Ribonuclease H-like 48_11	11 (2.3%)	Ribonuclease H-like motif (82%)	PS00329 (36%)	ATP (64%)
			PS00297 (27%)	ADP (36%)

the same cluster that belong to different folds or even classes of protein. Where there is significant structural similarity these later similarities may represent fundamental chemical constraints on the recognition of functional groups in protein binding. Clustering of the 476 representative local environments set produces ten main clusters (see Table I) representing 49% (232) of the total number of representative atom-sets. We broadly categorize those clusters according to their main (fold-level) structural characteristics, but focus on the atomic-level similarities of the binding sites themselves. The folds that represent the largest clusters are the P-loop, FAD/NAD(P)-binding, FAD-binding, Rossmann and Ribonuclease H-like folds.

P-loop

There are three main clusters displaying P-loop fold topologies. These are clusters 4_3, 6_6 and 48_4 with 32, 22, and 24 members respectively, totalling 78 local environments. Clusters are named according to the number of clusters at a particular agglomeration level followed by the cluster number (see Cluster Analysis section in Methods).

Cluster 4_3

Cluster 4_3 is the largest cluster and is very well-defined [see Fig. 1(a)], predominantly covering ATP/ADP binding sites, from eight different SCOP families (see Table II). All but one environment belong to the SCOP fold-level: P-loop containing nucleotide triphosphate hydrolases. The single exception is an environment from a kinase (1aq2) which has the SCOP fold Phosphoenolpyruvate carboxykinase (ATP-oxaloacetate carboxy-liase). It superposes closely to the other members of the cluster and contains the predominant "P-loop" PROSITE pattern PS00017 [see Fig. 1(b)].

The majority (29/32) of the environments contain no less than 75% of the residues of the "ATP/GTP-binding site motif A" (P-loop) PROSITE pattern (PS00017), described as [AG]-x(4)-G-K-[ST]. The remaining three environments do not contain this sequence pattern. However, they show a close structural overlap and differ only in one residue position at the start or end of the sequence pattern. They belong to structural families of the P-loop fold that contain the PS00017 PROSITE pattern. The environments are centered around the beta-phosphate, P-β (22/32) or alpha-phosphate, P-α (10/32) of the mono-di-tri nucleotides. The close structural overlap is also reflected in the placement of the magnesium ions that are complexed with the ligand (22/32).

Cluster 6_6

Cluster 6_6 also covers ATP/GTP binding sites [see Fig. 1(c)]. All the folds present in the cluster are also represented in cluster 4_3 with atom-sets from seven families

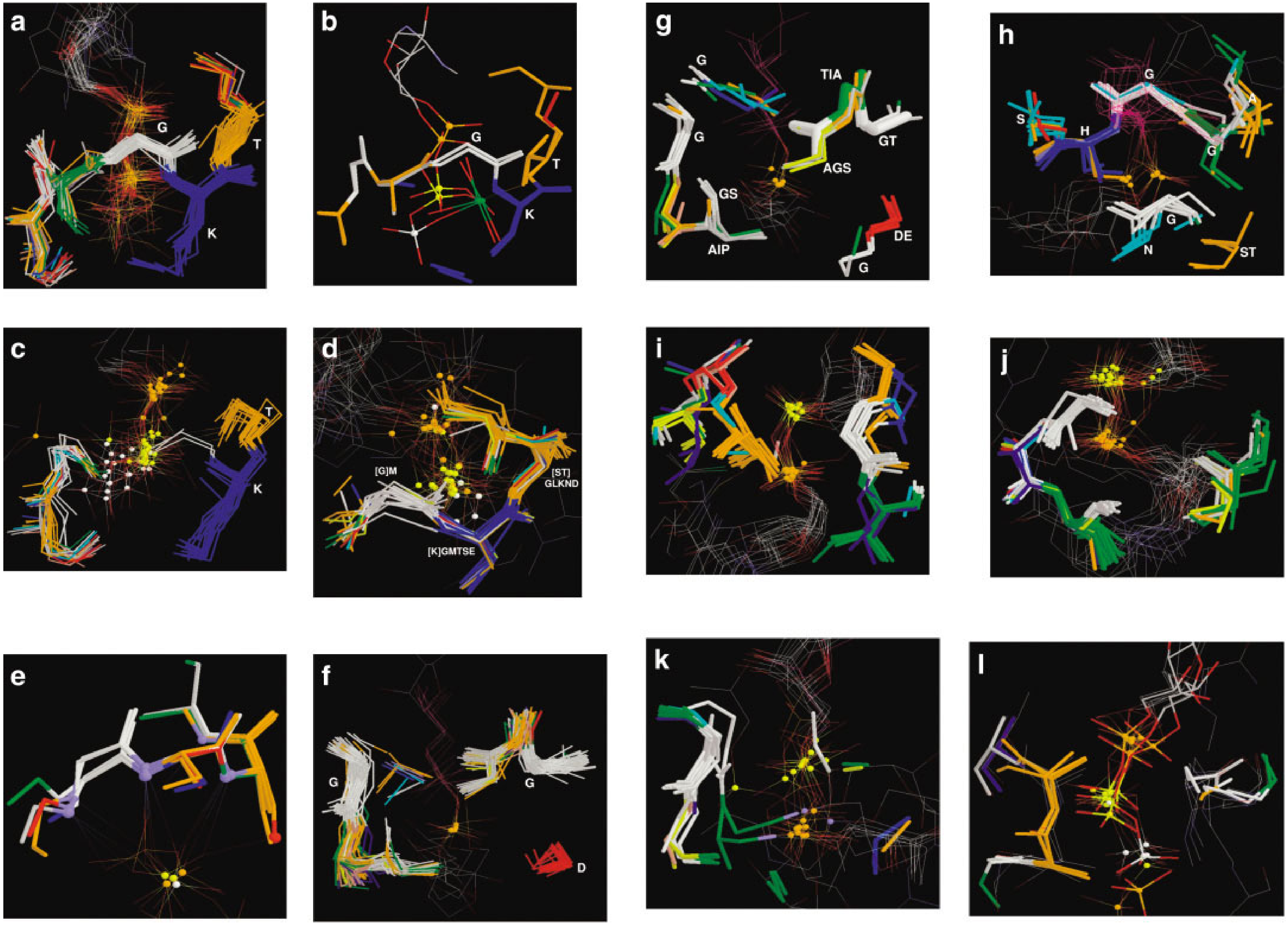
a
G
T
K
b
G
T
K
g
G
TIA
G
GT
AGS
GS
AIP
DE
G
h
G
S
H
A
G
G
N
ST
c
T
K
d
[G]M
[ST]
GLKND
[K]GMTSE
i
j
e
f
G
G
D
k
l

Figure 1.

and two folds. The center for the cut-off sphere that defines the environment is the gamma phosphate, P-γ group (the third phosphate). Hence all but three ligands are nucleotide tri-phosphates (ATP or GTP). Only three environments are centered about the P-β phosphate in this cluster. The main difference with the first cluster is that the structural P-loop is interrupted in the middle of its length with part of the backbone lying outside the local environment distance limit. Like the first cluster, there is a single atom-set from a protein (1ayl, a kinase) with a different fold (Phosphoenolpyruvate carboxykinase, ATP-oxaloacetate carboxy-liase) that superposes very well with the rest of the atom-sets that come from proteins exhibiting the fold "P-loop containing nucleotide triphosphate hydrolases." The cluster is essentially a satellite cluster of 4_3 representing the P-loop as viewed from the P-γ phosphate environment. It is noteworthy that there are no P-γ phosphate centered environments in cluster 4_3.

Cluster 46_4

Cluster 46_4 is highly diverse in its distribution of folds and represents a more diffuse overlap, [see Fig. 1(d)]. It contains 24 local environments that represent five different classes according to the SCOP classification, nine folds and 14 families (see Table II). The classes include: all-alpha proteins (α), all-beta proteins (β), alpha and beta proteins—mainly parallel beta sheets (α/β), alpha and beta proteins—mainly antiparallel beta sheets (α + α), and multi-domain proteins (α and β). Despite such heterogeneity, the cluster still displays significant structural similarity in the signature G-K-[S/T] part of the P-loop. However, many of the environments now lack the PROSITE sequence motif. The cluster is best described as the structural P-loop; glycine being the most conserved (23/24), then lysine (16/24). However, there is little sequence similarity beyond this.

The structural P-loop consists of an N-terminal helix capped by a glycine residue. Several studies have commented on the significance of this arrangement noting the significance of the helix dipole in stabilizing the binding of negative charge.[7,26,27] Seemingly the only sequence requirement is for the glycine residue in the G-K-[S/T] motif. The importance of the strict conservation of glycine has been questioned,[27] given that the potential Cβ position is not necessarily occupied by one of the phosphate atoms and the glycine (Φ,Ψ) angles are scattered over the Ramachandran (Φ,Ψ) map. This later point would appear to be in disagreement with our findings. Even within the most diverse P-loop cluster 46_4, the G-K-[S/T] glycine (Φ,Ψ) angles do appear to be centred around (0,90) which is not in one of the favoured Ramachandran regions (A,B,L) but bounds the additionally allowed (l) and generously allowed regions (~l). This arrangement allows the four backbone NH groups of the residues at G-K-[S/T]-X to coordinate the phosphate oxygens through hydrogen bonds [see Fig. 1(e)]. Visual inspection shows that this is mostly via two (or more) h-bonds to oxygens of each adjacent phosphate, i.e., it is a di-phosphate binding motif. Both possible di-phosphate binding modes (P-γ + P-β and P-α + P-β phosphate groups) are present in the structural data for di- and tri-phosphate nucleotides. However, it should be noted that not all four possible hydrogen bonds are always present.

The dominant sequence pattern of the structural P-loop is, expectedly, that of the "ATP/GTP-binding site motif A" PROSITE pattern (PS00017), described as [AG]-x(4)-G-K-[ST]. Extending it so as to include all the environments of the P-loop clusters 4_3 and 6_6 results in the pattern [AGT]-x(4)-G-K-[STG]. However, there is much greater sequence diversity in cluster 48_4 [see Fig. 1(d)], therefore a single sequence pattern describing the structural P-loop is less appropriate. Furthermore, this local structural pattern appears in proteins with very different overall folds (up to the class level). All the conserved loops are in fact part of the first turn of an α-helix preceded by a Gly residue.

FAD/NAD(P)-Binding

This constitutes the single largest cluster with 62 environments that represent five families from two folds (see Tables I and II). The most prevalent ligand is FAD (89% of the environments) bound in an extended conformation and the cluster is dominated by the "FAD/NAD(P)-binding domain" SCOP fold from the alpha and beta (a/b) class.

This cluster is surprisingly compact given its size. It is dominated by three families of the "FAD/NAD(P)-binding domain" fold and also has representatives of two families of the "Nucleotide-binding domain" fold. At low focus values (30%) it displays a cage-like backbone frame (of four segments 2–4 residues long). The arrangement is distinct from the structural P-loop. At a 50% focus value, what is mainly retained is two three-residue long loops and a conserved Asp residue [see Fig. 1(f)].

Within the different families that comprise the cluster there is sequence conservation that is reflected in the

Fig. 1. **a**: The P-loop cluster 4_3 at a focus value of 50%. The protein atoms are colored according to their RasMol[40] residue types, illustrating the high degree of atomic similarity between their sequences. The white, blue, and orange colored residues correspond to the characteristic G-K-[ST] of the PROSITE sequence pattern PS00017. The ligands in the background are the hetero atoms of the ligand which are included for reference and were not included in the geometric matching. These are colored by RasMol atom type. **b**: A superposition between representatives of the two SCOP folds that are present in the 4_3 cluster, at 50% focus. Both are complexed with AD(T)P through a magnesium ion (colored green). The phosphate groups of the ligands are highlighted and colored according to their position along the ligand, P-α (orange), P-β (yellow), P-γ (white). **c**: The P-loop cluster 6_6 at a focus value of 50%. **d**: Cluster 48_4 at a focus value of 50%. The sequence variability is indicated. The bracketed residues are the ones that comply with the PROSITE pattern PS00017. **e**: Superposition of the representatives of the five SCOP classes of P-loop cluster 48_4 at a focus value of 50%. The main-chain amide nitrogen atoms involved in hydrogen bonding the structurally conserved phosphate atoms are shown (as blue spheres). **f**: Cluster 4_1 at a focus value of 50%. **g**: Superposition of the representatives of the five families (from two folds) of cluster 4_1 at a focus value of 50%. The Residue positions with highest sequence conservation are labelled. The PS00225 PROSITE pattern is shown as a thicker wire-frame representation. **h**: Cluster 14_12 at a focus value of 50%. The two compact constellations of ligand phosphates belong to each of the two families in the cluster. **i**: Cluster 29_24 at a focus value of 50%. **j**: Cluster 29_29 at a focus value of 50%. **k**: Cluster 48_35 at a focus value of 50%. **l**: Cluster 48_11 at a focus value of 50% with three family representatives in bold.

TABLE II. Distribution of SCOP Families for Each Fold in the Ten Main Clusters

Cluster	SCOP Fold and Class		SCOP family
P-loop 4_3	P-loop containing nucleotide triphosphate hydrolases	a/b	Nucleotide and nucleoside kinases
			G proteins
			Motor proteins
			Nitrogenase iron protein-like
			RecA protein-like (ATPase-domain)
			ABC transporter ATPase domain-like
			Extended AAA-ATPase domain
	Phosphoenolpyruvate carboxykinase (ATP-oxaloacetate carboxy-liase)	a/b	Phosphoenolpyruvate carboxykinase (ATP-oxaloacetate carboxy-liase)
P-loop 6_6	P-loop containing nucleotide triphosphate hydrolases	a/b	G proteins
			Motor proteins
			Nitrogenase iron protein-like
			RecA protein-like (ATPase-domain)
			ABC transporter ATPase domain-like
			Extended AAA-ATPase domain
	Phosphoenolpyruvate carboxykinase (ATP-oxaloacetate carboxy-liase)	a/b	Phosphoenolpyruvate carboxykinase (ATP-oxaloacetate carboxy-liase)
P-loop 48_4	P-loop containing nucleotide triphosphate hydrolases	a/b	Nucleotide and nucleoside kinases
			Chloramphenicol phosphotransferase
			G proteins
			Nitrogenase iron protein-like
			RecA protein-like (ATPase-domain)
	PRTase-like	a/b	Phosphoribosyltransferases (PRTases)
	Ribokinase-like	a/b	MurD/MurF
	GroEL-like chaperones, ATPase domain	All alpha	GroEL
			Group II chaperonin (CCT, TRIC)
	Reductase/isomerase/elongation factor common domain	All beta	Phthalate dioxygenase reductase
	FMN-binding split barrel	All beta	PNP-oxidase like
	Ribosomal protein S5 domain 2-like	a + b	Homoserine kinase, N-terminal domain
	Ferredoxin-like	a + b	NAD-binding domain of HMG-CoA reductase
	Sugar phosphatases	a and b	Sugar phosphatases
FAD binding 4_1	FAD/NAD(P)-binding domain	a/b	FAD-linked reductases, N-terminal domain
			Succinate dehydrogenase/fumarate reductase N-terminal domain
			FAD/NAD-linked reductases, N-terminal and central domains
	Nucleotide-binding domain	a/b	N-terminal domain of adrenodoxin reductase-like
			D-amino acid oxidase, N-terminal domain
FAD binding 4_12	FAD-binding domain	a + b	FAD-linked oxidases, N-terminal domain
			CO dehydrogenase flavoprotein N-terminal domain-like
Rossmann fold 29_24	NAD(P)-binding Rossmann-fold	a/b	Tyrosine-dependent oxidoreductases c.2.1.2
Rossmann fold 29_29	NAD(P)-binding Rossmann-fold	a/b	Alcohol/glucose dehydrogenases, C-terminal domain
			Tyrosine-dependent oxidoreductases
			Glyceraldehyde-3-phosphate dehydrogenase-like, N-terminal domain
			Formate/glycerate dehydrogenases, NAD-domain
			Lactate & malate dehydrogenases, N-terminal domain
			6-phosphogluconate dehydrogenase-like, N-terminal domain
			Amino-acid dehydrogenase-like, C-terminal domain
Rossmann fold 48_30	NAD(P)-binding Rossmann-fold	a/b	Alcohol/glucose dehydrogenases, C-terminal domain
			Tyrosine-dependent oxidoreductases
			Glyceraldehyde-3-phosphate dehydrogenase-like, N-terminal domain
			Formate/glycerate dehydrogenases, NAD-domain
			Lactate & malate dehydrogenases, N-terminal domain
			6-phosphogluconate dehydrogenase-like, N-terminal domain
			Amino-acid dehydrogenase-like, C-terminal domain
Rossmann fold 48_35	NAD(P)-binding Rossmann-fold	a/b	Alcohol/glucose dehydrogenases, C-terminal domain
			Formate/glycerate dehydrogenases, NAD-domain
			6-phosphogluconate dehydrogenase-like, N terminal domain
	FAD/NAD(P)-binding domain	a/b	FAD-linked reductases, N-terminal domain
			FAD-linked reductases, N-terminal and central domains
	Phosphoglycerate kinase	a/b	Phosphoglycerate kinase
Ribonuclease 48_11	Ribonuclease H-like motif	a/b	Actin/HSP70
			Glycerol kinase

presence of five PROSITE patterns, parts of which (< 40%) are included in the local environments. However, most of these patterns are not in the structurally conserved regions. Only PROSITE patterns, PS00504 (Fumarate reductase/succinate dehydrogenase FAD-binding site) and PS00225 (Crystallins beta and gamma "Greek key" motif signature) overlap with part of the structural consensus [Fig. 1(g)]. These patterns however are present in just 9% of the cluster atomsets (20% of the ones that have any PROSITE pattern). Furthermore only 3–4 of the residues of these patterns overlap with the conserved structure and they are not always the unambiguous residues of the pattern. This contrasts with the P-loop clusters where the PROSITE pattern has extensive overlap whenever it is present. The overall sequence conservation of the structural consensus in this cluster is limited to a few key residues. These are mainly glycine residues that make close contact (van der Waals and h-bonding) and form a glycine-cage around the nucleotide di-phosphate moiety. Typically this consists of four or five glycine residues [see Fig. 1(g)] present on the four separate segments of the protein chain. However, the core conservation of this cluster are two glycine residues, and an aspartate residue. One is the second glycine within the well known di-nucleotide binding motif (characteristic of the Rossmann-like fold and also found in both the "Nucleotide-binding domain" and "NAD(P)-binding Rossmann-fold" SCOP folds) containing the consensus sequence G-x-**G-x-x**-G, where bold represents structural conservation. The highly conserved aspartate forms a hydrogen bond with the O3 group of the flavin moiety of FAD and constitutes one of the residues within a FAD-binding sequence fingerprint[28] T-x(4)-h-ϕ-h-h-G-**D** (where h, a non-polar residue; ϕ, is any aromatic residue).

FAD-Binding

Cluster 14_12 contains 13 local environments from two families of the alpha + beta (a + b) class, "FAD-binding domain" fold of covalent flavoproteins.[29] It mostly binds FAD (in 85% of the environments), which displays a kinked conformation. For the only two environments that bind ADP the ligands superpose well with FAD, having a similar orientation and positioning of the adenene ring. All 13 local environments belong to two SCOP families of the "FAD-binding domain" superfamily. They differ from the FAD/NAD(P)-binding cluster at the class level (see Table II). There is a distinct structural "signature" for the cluster. It contains a structurally conserved semi-circular 7–9 residue loop that accommodates the nucleotide di-phosphate moiety and two independent 2-residue long segments [see Fig. 1(h)].

There is a degree of residue conservation on visual inspection, but no conserved sequence pattern. However, many of the environments contain two consensus glycine residues at the same locations within the 7–9 residue loop and another conserved glycine in an adjacent 2-residue segment. Like the FAD/NAD(P)-binding cluster this constitutes a glycine "cage" around the di-phosphate moiety. However, they do not contain the di-nucleotide binding sequence motif. There is only one instance of a PROSITE pattern present and this is the "Peroxidase_1" ("Peroxidases proximal heme–ligand signature", PS00435). Interestingly the PROSITE pattern matches the full length of the 7–9 residue loop, however, the match is a false positive hit in the PROSITE database. The structural superposition results in a poorly defined sequence pattern probably of limited use for sequence searching.

The two predominantly FAD binding clusters (FAD/NAD(P)-binding and FAD-binding) do not have many similarities on a superficial level. Their dominant folds belong to different classes, their ligands have different conformations, they do not share any appreciable sequence similarity and there are distinct structural features in the local environments. However, both contain conserved yet different glycine "cage" structures that although not completely conserved in all members do represent consensus structural motifs with conserved main-chain conformations.

Rossmann Fold

There are four clusters that are dominated by the Rossmann fold and contain 15, 21, 19, and 13 members respectively totalling 68 local environments.

In the first cluster (29_24) all 15 atom-sets are members of the "Tyrosine-dependent oxidoreductases" SCOP family, of the "NAD(P)-binding Rossmann-fold domains" fold. The structurally conserved region is mainly along two 3–4-residue loops which display a very compact superposition [see Fig. 1(i)]. The cluster has no PROSITE sequence patterns and it binds both NADP and NAD with very little difference between the two ligand environments in this cluster. There is sequence conservation of glycine-isoleucine-glycine residues with the isoleucine pointing toward the nicotinamide ring. This occurs in the common NAD(P) phosphate binding motif G-x-x-**x-G-I-G** characteristic of the Rossmann fold (bold represents structural conservation) as well as a threonine residue in an adjacent segment.

The second cluster (29_29) includes seven families from the "NAD(P)-binding Rossmann-fold domains" fold including "Tyrosine-dependent oxidoreductases" SCOP family (see Table II). It has no associated PROSITE pattern and predominantly binds NAD. It has a very compact four-residue loop containing the di-nucleotide binding motif (also characteristic of the Rossmann-like fold) with a strong sequence preference G-x-**G-x-V-G** with the valine pointing toward the nicotinamide ring. There is a separate much more poorly defined 2–3-residue stretch on an adjacent segment [see Fig. 1(j)]. The cluster is centered around the nicotinomide phosphate of NAD(P).

The third Rossmann fold cluster 48_30 matches cluster 29_29 in most respects. It includes the same seven families from the "NAD(P)-binding Rossmann-fold domains" superfamily (see Table II). It predominantly binds NAD and has no PROSITE patterns associated with it. The main difference is that the cluster is centered around the adenine phosphate of NAD(P) therefore there is a different structural consensus again involving the di-nucleotide binding motif G-**x**-**G**-**x**-**V**-G. There is also a separate very poorly defined 2–3-residue stretch on an adjacent segment. The cluster displays similar local structure to that of 29_29 and is a case of clusters that diverged from the same branch.

The last cluster, 48_35, differs in that it includes six families from three different folds (see Table II). It displays a local environment that is analogous to the 29_29 and 48_30 clusters [see Fig. 1(k)], however, is structurally much more sparsely conserved. It includes part of the di-nucleotide binding motif G-**x**-**G**-**x**-x-G also conserved in the Rossmann fold and the Rossmann-like fold of the FAD/NAD(P)-binding domain. This more distant structural relationship is not surprising, given the relatively small extent of structural overlap. However, structural similarity with (3pgk) the ATP binding Rossman-like alpha/beta 3-Layer(aba) sandwich of the Phosphoglycerate kinase (PGK) fold is even more remote (residues 334–336, **N-G-P**). The PGK fold is also known to include the structural P-loop (residues 370–372, G-G-D),[27] however, only the aspartate residue is present in this particular phosphate environment.

Ribonuclease H-Like

The single cluster 48_11 includes 11 atom-sets of three families from a single superfamily. All members have the Ribonuclease H-like fold and have two distinct loops and PROSITE patterns that map to structurally conserved regions in both [see Fig. 1(l)]. We should note that this cluster diverges from the Rossmann fold clusters. It contains the structural di-nucleotide binding motif G-**x**-**G**-**x**-x-G, however, only the second glycine residue is conserved (e.g., 1kax, 70kDa heat shock cognate protein (HSP70), residues 200–202, **L-G-G**). This maps to the signature HSP70_2 PROSITE pattern, P00329, (in 1kax, residues 197-210, IFD**LGG**GTFDVSIL). The other larger loop is only characteristic of the Ribonuclease H-like fold containing the HSP70_1 PROSITE pattern, P00297. The distinctive structurally conserved five-residue loop maps to this PROSITE pattern [IV]-D-**L-G-T-[ST]-x**-[SC]. However, it should be noted that seven of the atom-sets of the cluster contain neither PROSITE pattern, reflecting the higher level of structural conservation.

DISCUSSION

We have described a method for an automated all-against-all structural comparison of phosphate binding sites in protein–nucleotide complexes using geometric matching. Clustering by site similarity produces a classification of binding sites for the 476 representative local environments producing ten main clusters representing 49% of the representative atom-sets. The similarities make sense in terms of both structural and functional classification schemes. We have categorized the clusters with atomic similarity according to their main (fold-level) structural characteristics. In many of the clusters sequence conservation is evident in many members but is rarely conserved throughout the structurally conserved binding sites. Half of the phosphate groups in protein–nucleotide complexes cluster into one of the ten main clusters and these in turn represent a very limited number of unique structural binding motifs for phosphate. These are the structural P-loop, di-nucleotide binding motif (FAD/NAD(P)-binding and Rossman-like fold) and FAD-binding motif. Similar classification schemes to the one reported here have also been arrived at for nucleotide binding proteins independently by others.[27,30–32] The results of the all-against-all automated classification here are in agreement with these classification schemes, although detailed comparison is not possible. The characteristics of the di-nucleotide binding motif and mononucleotide-binding motif (or P-loop) are well documented.[30] The FAD/NAD(P)-binding and Rossman folds employ the same intrinsic di-nucleotide recognition motif with a consensus sequence G-x-G-x-x-G where the first two glycines contact the pyrophosphate moiety of the di-nucleotide. These two glycines are highly conserved in the structural binding motif and have main-chain dihedral angles that require glycine residues. The structural P-loop motif is ubiquitous in phosphate binding proteins with a consensus sequence G-K-[S/T], again glycine plays an important role in facilitating the close approach of phosphate to the protein main-chain. Indeed glycine is highly conserved in the binding loops of the FAD-binding glycine "cage" and the distinctive structurally conserved 5-residue loop in the ribonuclease H-like fold. Glycine appears to perform a unique role in phosphate binding that is linked to facilitating the close approach of phosphate oxygen atoms to the amide atoms of the protein main-chain. In some cases this is a result of allowing main-chain dihedral angles that require glycine residues. In others the absence of the steric bulk of a side-chain Cβ atom is a contributing factor. Interestingly, and in agreement with our findings, Lupas et al. have shown that the P-loop and FAD/NAD(P)-binding motifs represent two examples (among several others) of instances of local sequence and structural similarity in different protein folds.[33] They suggest that together with evidence from proteins containing sequence (and structural) repeats this argues in favor of the evolution of modern protein domains from ancient short peptides (antecedent domain segments). Such a mechanism would explain why these local motifs are ubiquitous in nucleotide binding proteins with different folds, however, other hypotheses such as evolutionary convergence or divergence are also possible.[33]

Some issues arise out of trying to produce an automated classification of phosphate binding sites in protein–nucleotide complexes. The 7-Å radius around the PO_4 phosphate is not enough to cover the whole ligand environment, resulting in the production of shifted views of the same environment through the perspective of the different ligand PO_4 positions e.g., the mono-, di- or tri-nucleotide phosphates. These correctly group into different "daughter" clusters. In such a case we accept the descriptive level of their single parental cluster (as in cluster 4_1) or we consider them collectively in retrospect (as it is the case for the P-loop clusters 4_3 & 6_6). It would be misleading to consider the daughter clusters as "distinct" biological entities, since they refer, ultimately, to the same nucleotide ligand binding site. A modification of the definition of the local environment could remedy this in some cases. However, similarities are expected for ligand environments defined by both the whole nucleotide ligand and/or by smaller defined moieties such as the PO_4 functional group analyzed here. Secondly, there are instances of local environments of the same fold that have been clustered separately in other smaller clusters. Again, this is an effect of the local nature of the overlap (radial cut-off from the phosphorous atom), and a consequence of the clustering method used here.

This analysis does not provide a comprehensive classification of all phosphate binding sites in protein–nucleotide complexes. We have detailed only the top ten clusters. In this sense the current study provides a "top-down" classification of binding sites representing the largest, most structurally diverse yet visually coherent clusters. Many of the other clusters are visually less coherent than the ten detailed here when clustered according to pair-wise similarity scores using the UMPG method and have not been analyzed in detail. Others produce smaller coherent clusters. Also, this means that visually coherent clusters with clear atomic level similarity, e.g., the Rossmann fold and Ribonuclease H-like fold clusters (which both contain the dinucleotide binding motif) coalesce, often with several other clusters to form less visually coherent clusters. This is an unfortunate yet inherent consequence of using the current means linkage clustering method. A purely numerical similarity threshold[27] or even better, a statistical significance score[34] would provide a better basis for structural classification. A further limitation of the current method relates to the initial similarity clustering to create the 476 representative local environments. A number of highly similar binding sites can sometimes be condensed to one or two representatives. This has the effect of condensing some SCOP superfamilies of related sites to small clusters that appear to under-represent them in the 476 representatives (at least in terms of their biological prevalence). This is in some cases further exacerbated by nucleotide binding proteins that are predominantly complexed with synthetic nucleotide phosphate derivatives (e.g., ANP, GNP etc.) that were not included in this study. Therefore, families such as the Protein Kinases appear under represented and do not constitute a significantly large cluster in the present study. Lastly we have not considered solvent molecules in this study, however, conserved structural water molecules can sometimes also play an important role in ligand binding.[35]

Recently, other new methods for comparing and searching for catalytic/binding sites have been described.[36–39] These address the problem of doing pair-wise comparison at either the site or whole protein level to reveal local atomic similarities within tertiary structure or alternatively to search a predefined structural template against a the protein database. The detection of local structural similarity in proteins using these methods or those described here could provide further insights into protein function particularly where sequence and fold similarity are absent.

ACKNOWLEDGMENTS

This work was supported by funding from the Wellcome Trust and BBSRC (R.M.J.) and in the form of a studentship from the Onassis foundation (A.B.). We would like to thank Janet Thornton for her support and helpful discussions.

REFERENCES

1. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A. The PROSITE database, its status in 2002. Nucleic Acids Res 2002;30:235–238.
2. Attwood TK, Blythe MJ, Flower DR, Gaulton A, Mabey JE, Maudling N, McGregor L, Mitchell AL, Moulton G, Paine K, Scordis P. PRINTS and PRINTS-S shed light on protein ancestry. Nucleic Acids Res 2002;30:239–241.
3. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2002;30:276–280.
4. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
5. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
6. Holm L, Sander C. Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res 1997;25:231–234.
7. Swindells MB. Classification of doubly wound nucleotide binding topologies using automated loop searches. Protein Sci 1993;2:2146–2153.
8. Via A, Ferre F, Brannetti B, Valencia A, Helmer-Citterich M. Three-dimensional view of the surface motif associated with the P-loop structure: cis and trans cases of convergent evolution. J Mol Biol 2000;303:455–465.
9. Doherty AJ, Serpell LC, Ponting CP. The helix-hairpin-helix DNA-binding motif: a structural basis for non-sequence-specific recognition of DNA. Nucleic Acids Res 1996;24:2488–2497.
10. Jones S, Barker JA, Nobeli I, Thornton JM. Using structural motif templates to identify proteins with DNA binding function. Nucleic Acids Res 2003;31:2811–2823.
11. Kasuya A, Thornton JM. Three-dimensional structure analysis of PROSITE patterns. J Mol Biol 1999;286:1673–1691.
12. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J Mol Biol 1994;243:327–344.
13. Wallace AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 1997;6:2308–2323.
14. Russell RB. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 1998;279:1211–1227.
15. Wallace A, Thornton J. PROCAT, a database of 3D enzyme active site templates. http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html.
16. Aloy P, Querol E, Aviles FX, Sternberg MJ. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 2001;311:395–408.
17. de Rinaldis M, Ausiello G, Cesareni G, Helmer-Citterich M. Three-dimensional profiles: a new tool to identify protein surface similarities. J Mol Biol 1998;284:1211–1221.
18. Jackson RM, Russell RB. The serine protease inhibitor canonical loop conformation: examples found in extracellular hydrolases, toxins, cytokines and viral proteins. J Mol Biol 2000;296:325–334.
19. Bachar O, Fischer D, Nussinov R, Wolfson H. A computer vision based technique for 3-D sequence-independent structural comparison of proteins. Protein Eng 1993;6:279–288.
20. Pennec X, Ayache N. A geometric algorithm to find small but highly similar 3D substructures in proteins. Bioinformatics 1998; 14:516–522.

21. McLachlan AD. Gene duplications in the structural evolution of chymotrypsin. J Mol Biol 1979;128:49–79.
22. Bron C, Kerbosch J. Algorithm 457: finding all cliques of an undirected graph. Commun ACM 1973;16:575–577.
23. Martin AC, Thornton JM. Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. J Mol Biol 1996;263:800–815.
24. Mojena R. Hierarchical grouping methods and stopping rules: an evaluation. The Computer Journal 1977;20:359–363.
25. Fuchs R. Predicting protein function: a versatile tool for the Apple Macintosh. Comput Appl Biosci 1994;10:171–178.
26. Copley RR, Barton GJ. A structural analysis of phosphate and sulphate binding sites in proteins. Estimation of propensities for binding and conservation of phosphate binding sites. J Mol Biol 1994;242:321–329.
27. Kinoshita K, Sadanami K, Kidera A, Go N. Structural motif of phosphate-binding site common to various protein superfamilies: all-against-all structural comparison of protein-mononucleotide complexes. Protein Eng 1999;12:11–14.
28. Eppink MH, Schreuder HA, Van Berkel WJ. Identification of a novel conserved sequence motif in flavoprotein hydroxylases with a putative dual function in FAD/NAD(P)H binding. Protein Sci 1997;6:2454–2458.
29. Murzin AG. Structural classification of proteins: new superfamilies. Curr Opin Struct Biol 1996;6:386–394.
30. Schulz G. Binding of nulceotides by proteins. Curr Opin Struct Biol 1992;2:61–67.
31. Traut TW. The functions and consensus motifs of nine types of peptide segments that form different types of nucleotide-binding sites. Eur J Biochem 1994;222:9–19.
32. Denessiouk KA, Rantanen VV, Johnson MS. Adenine recognition: a motif present in ATP-, CoA-, NAD-, NADP-, and FAD-dependent proteins. Proteins 2001;44:282–291.
33. Lupas AN, Ponting CP, Russell RB. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol 2001;134:191–203.
34. Stark A, Sunyaev S, Russell RB. A model for statistical significance of local similarities in structure. J Mol Biol 2003;326:1307–1316.
35. Bottoms CA, Smith PE, Tanner JJ. A structurally conserved water molecule in Rossmann dinucleotide-binding domains. Protein Sci 2002;11:2125–2137.
36. Lehtonen JV, Denessiouk K, May AC, Johnson MS. Finding local structural similarities among families of unrelated protein structures: a generic non-linear alignment algorithm. Proteins 1999;34:341–355.
37. Oldfield TJ. Data mining the protein data bank: residue interactions. Proteins 2002;49:510–528.
38. Stark A, Russell RB. Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. Nucleic Acids Res 2003;31:3341–3344.
39. Jambon M, Imberty A, Deleage G, Geourjon C. A new bioinformatic approach to detect common 3D sites in protein structures. Proteins 2003;52:137–145.
40. Sayle RA, Milner-White EJ. RASMOL: biomolecular graphics for all. Trends Biochem Sci 1995;20:374.