Fast and automated functional classification with MED-SuMo: an application on purine-binding proteins.

Olivia Doppelt-Azeroual, François Delfaud, Fabrice Moriaud, Alexandre de Brevern

Protein Science (2010) *in press*

Fast and automated functional classification with MED-SuMo: an application on purine binding proteins.

Olivia Doppelt-Azeroual [1,2,§], François Delfaud [2], Fabrice Moriaud [2] and Alexandre G. de Brevern [1]

[1] INSERM UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Paris Diderot – Paris 7, Institut National de la Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France.

[2] MEDIT SA, 2 rue du Belvédère, 91120, Palaiseau, France.

Short title: functional classification with MED-SuMo

[§]Corresponding author: Dr. Olivia Doppelt-Azeroual, MEDIT SA, 2 rue du Belvédère, 91120, Palaiseau, France.

Email addresses: olivia.doppelt@medit.fr, francois.delfaud@medit.fr, fabrice.moriaud@medit.fr, alexandre.debrevern@univ-paris-diderot.fr

Keywords: protein structures; ligand-protein interactions; binding sites; binding site classification; CDK2 and Aurora-A similarity; protein kinases.

Abstract

Ligand-protein interactions are essential for biological processes, and precise characterization of protein binding sites is crucial to understand protein functions. MED-SuMo is a powerful technology to localize similar local regions on protein surfaces. Its heuristic is based on a 3D representation of macromolecules using specific Surface Chemical Features associating chemical characteristics with geometrical properties. MED-SMA is an automated and fast method to classify binding sites. It is based on MED-SuMo technology, which builds a similarity graph, and it uses the Markov Clustering algorithm.

Purine binding sites are well studied as drug targets. Here, purine binding sites of the Protein DataBank (PDB) are classified. Proteins potentially inhibited or activated through the same mechanism are gathered. Results are analyzed according to PROSITE annotations and to carefully refined functional annotations extracted from the PDB.

As expected, binding sites associated with related mechanisms are gathered, *e.g.*, the Small GTPases. Nevertheless, protein kinases from different Kinome families are also found together, *e.g.*, Aurora-A and CDK2 proteins which are inhibited by the same drugs. Representative examples of different clusters are presented.

The effectiveness of the MED-SMA approach is demonstrated as it gathers binding sites of proteins with similar Structure-Activity Relationships. Moreover, an efficient new protocol associates structures absent of co-crystallized ligands to the purine clusters enabling those structures to be associated with a specific binding mechanism.

Applications of this classification by binding mode similarity include target based drug design and prediction of cross-reactivity and therefore potential toxic side effects.

Introduction

The use of the protein sequence is the simplest approach to infer *by analogy* a protein function, *e.g.*, PSI-BLAST [1]. In this research area, PROSITE is a recognized method, distinguishing protein family members from unrelated proteins [2, 3]. PROSITE patterns represent conserved motifs such as binding site regions. The PROSITE database scans sequences from the annotated UniProtKB / Swiss-Prot database [4, 5] and detects functional domains. Numerous other approaches exist, such as Pfam which uses a refined database of well-characterized protein domain families [6, 7]. The development of so many methods has led to the creation of meta-servers that measure consensus across multiple approaches, *e.g.*, JAFA [8].

Functional protein properties can also be characterized in terms of three-dimensional (3D) structural information. This provides valuable information for determining and understanding precise mechanisms of proteins implicated in diseases [9-11]. The combination of knowledge from 3D protein structures with hundreds of thousands of small-molecules can be used for structure and ligand-based drug design [12-17]. For example, Crespo and Fernandez used the protein structure of a imatinib-resistant mutant [18] to improve the anticancer drug by promoting stronger intermolecular non-bonded interactions than those bound by the original drug. In the same way, the resolution of the HIV-1 reverse transcriptase complex structure explained the potential application of anti-HIV drugs against resistance mutations. It also provided opportunities for understanding, with greater accuracy, inhibitor–

protein interactions and to determine reliably the structural effects of resistance mutations [19].

The Protein Data Bank (PDB [20, 21]) gathers today more than 59000 protein structures. About 3000 protein structures are not associated with a function. Proteins can be classified according to their folds [22-24], *e.g.*, SCOP (Structural Classification of Proteins) [25, 26]. From simple structural classification of protein, these methods have become useful tools to infer protein structures functions and to detect functional relationships, *e.g.* SCOP is now coupled with BLAST, while PSI-BLAST and RPS-BLAST are associated with Pfam domain search [1, 6, 27], even PROSITE motifs are now analyzed in 3D structures [28].

However, a limitation of these classifications is their use of complete protein folds or protein domains. Similarity of fold does not imply a direct similarity of function. For example, the TIM Barrel fold is an alternation of eight α-helices and eight parallel β-strands along with the peptide backbone. It is ancient [29] and shared by many different enzymes associated with at least 15 different functions [30]. In SCOP, all such proteins are associated with the same cluster.

It is now established that looking specifically at the protein interactions can clarify biological functions, *i.e.*, ligand-protein and protein-protein interactions. Ligand-protein interactions are at the basis of many fundamental biological processes. It is also known that the activity of a protein is mediated by a small, highly conserved set of residues within the binding site [31, 32]. Consequently, being able to detect and compare binding sites is valuable for the assignment of predicted structural functional annotations.

During recent years, various methods to compare binding sites have been elaborated, based on diverse types of descriptor. The general aim is to create

automated functional annotation methods independent from amino acid sequence or fold similarity.

Existing methods share common features. CavBase is based on the use of pseudo-centers, *i.e.*, 3D patterns corresponding to chemical properties of amino acids at the surface of proteins [33]. It detects related cavities using a clique detection algorithm. Site similarity ranking occurs according to property-based surface patches shared by the clique solutions. CavBase was used to predict unexpected drug cross-reactivity among functionally unrelated target proteins [34-36]. CavBase is restricted to cavity comparisons.

Like CavBase, SiteEngine [37, 38] also uses pseudo-centers. In SiteEngine, they are gathered into triangles which constitute vertices of graphs. The web version of the approach only enables the comparison of a single site versus another protein structure [39].

Other methods exist, including FLAP [40], CPASS [41] and eF-seek [42, 43]). Some enable the automatic creation of 3D motifs associated with binding sites for given type of ligands [44] or detect structural similarity to assign E.C. number [43]. While others, use the detection of conserved residues to characterise binding sites. In this field the Evolutionary Trace method is the most widespread [45-47]. For example, it was used to identify residue positions important in diverse GPCRs [48] and this method bypasses the need for experimental knowledge of the catalytic mechanism [49]. Thornton's group maintain the Catalytic Site Atlas (CSA), containing assigned catalytic residues, and gives an additional homologous set, with annotations inferred by PSI-BLAST and sequence alignment to the original set [31]. George and co-workers used the CSA to identify and segregate related proteins into those with a functional similarity and those where function differs [50]. ProFunc is a

metaserver using both sequence and structure prediction, although it does not provide any simple consensus output to use its results [51-54]. Roterman's approach is innovative. It detects regions of significantly irregular hydrophobicity distribution in proteins which appear to be associated with specific functions [55-58]. They propose a method to detect binding sites based on the hydrophobic distribution analysis in protein structures [59].

All methods cited above can be used to annotate protein structures. CPASS [41] and SiteEngine [37] were presented with examples of functional annotation for hypothetical protein structures while CavBase [33] and Gilbert and co-workers [44] illustrate their methods by classifying a protein structure dataset. Indeed, these types of classification are particularly relevant as they are not based on global fold alone. They provide structural functional classifications that can highlight links between proteins and that could be a good start for assigning predicted protein functions to hypothetical proteins.

In this research area, SuMo is a powerful technology to match similar local regions on protein surfaces [60]. Each chemical interactions of a amino acid residue is represented by a pseudo-center, named a Surface Chemical Feature (SCF) (see figure 1). These are gathered into triangles, the SuMo graph vertices. SCFs have heterogeneous geometrical properties, and these triangles have specific formation and superimposition rules (distance, angle), so the comparison heuristic is very fast. The comparison of a 3D motif against all binding sites of the PDB can be performed in a few minutes. The first demonstration of SuMo was the assignment of functional and non-functional lectins with a selectivity of 96% [61]. MED-SuMo is the latest evolution of the SuMo software developed by MEDIT [62] [63] [64]. Recent developments have improved the binding sites database, and functional annotation

functionalities are now integrated. Hydrophobic chemical features were added to the SCF dictionary and a cavity-detection algorithm included, improving MED-SuMo's effectiveness for detecting unknown binding sites.

Here, we use MED-SuMo to detect and characterize local similarities for the purpose of classify binding sites. This extension is known as the MED-SuMo Multi Approach (MED-SMA). A complete dataset of purine binding protein structures was collected and classified using MED-SMA. Different clusters were generated. Their characteristics and their distributions across different annotations, *e.g.*, PROSITE, ligand distribution, functional annotations, were analyzed. MED-SuMo is able to group protein binding sites with the same or with related functions, *i.e.*, sites which binds similarly related ligands. However, with purine binding sites, what is interesting is how drugs can be developed to activate or inhibit the protein function on which they are located.

In this study, we present the classification of all purine binding sites of the PDB and demonstrate a method to enrich the clusters with purine binding protein structures not co-crystallized with any ligands. Protein kinase distribution in the clusters is analyzed in the discussion section.

Results

Distinct functions can be underlined by the fact that two proteins interact differently with the same class of molecules. MED-SuMo can differentiate these proteins' binding sites. For example, no structural or functional similarities are detected between an actin protein bound to ATP (PDB code: 1S22) and a myosin protein bound to ATP (PDB code: 1FMW) [60, 61, 63, 64]. The classification tool MED-SMA was implemented to use this ability to classify datasets of binding sites

[63]. It operates through three main steps: (i) comparison of all the binding sites of a dataset using a pairwise comparison system, (ii) detection of matching regions in the binding sites to build a similarity graph, and (iii) classification of this graph with the Markov Clustering algorithm (MCL) [65]. This clustering algorithm detects densely populated regions of the similarity graph associated with highly scored matching regions and gathers the similar sub-sites into clusters. Figure 2 illustrates the overall procedure and specific details of the heuristic are provided in the Material and Methods section.

Purine binding sites classification.

Purine binding proteins structures were selected from the PDB. Ligands all contain either adenosine or guanosine: AXP (ATP, ADP, AMP, ANP and NAD) and GXP (GTP, GDP, GMP, and GNP) which only differ by two chemical groups while NAD and AXP differ on the phosphoryl side. The 2229 selected protein structures contain 2322 purine binding sites which were used to create the same number of MED-SuMo graphs. The classification required 4 hours on a bi-Xeon QuadCore 5335 machine, 3.5 hours for the pairwise comparison step and 30 minutes for the MCL classification. 247 clusters were created (see supplementary materials 1 and 2). These comprise 2115 binding sites, leaving 207 singletons associated with no binding site of the database. Each singleton is eliminated at the "patch merging" step represented on figure 2.2b of the MED-SMA procedure (see *Material & Method* part).

Global analysis. 60% (148) of the clusters have fewer than 5 binding sites, 25% (62) have between 5 and 10 binding sites, and 5% (14) have more than 30 binding

sites (see supplementary material 3a). Thus, small clusters are important in quantity, but there is a notable distribution of larger clusters.

Ligands within the classified binding sites were analyzed. In table 1a, the number of times a ligand is found at least once in a cluster is shown. Even if purine ligands are very similar, many clusters contain only one type. For instance, 82 clusters contain only ATP, 42 clusters have ATP and ADP and 8 have ATP and NAD. Thus for the AXP ligands, about 41% to 50% of their clusters contain just one type. These frequencies increase to ~60% for GTP and GMP clusters. Despite the fact that their structures are very similar, only 3 clusters contain both ATP and GTP. This highlights the specific binding modes of those ligands. For the clusters containing several ligand types, the nucleotide is often the same (Adenine or Guanine). Table 1b shows the same information as table 1a except the ligands are grouped into 3 classes: AXP, GXP and NAD. This table highlights the true diversity observed in the classification. 83% of the AXP clusters contain only AXP ligand, 71% of GXP clusters only GXP and 70% NAD clusters only NAD. The association of NAD and GXP is never observed whereas 30% of the NAD ligand clusters also contain AXP ligands. Nearly 17% of AXP clusters are found with NAD and GXP ligands.

PROSITE annotation in clusters. 296 different PROSITE patterns are associated with the proteins of the clusters. The pattern association is based on PDB identifiers. Therefore, as the chain corresponding to a pattern may not be the one where the purine binding site is located, manual checks were performed. 30% of the MED-SuMo clusters (74/247) are not associated with any defined pattern, while 28% (70/247) are associated with only one pattern and 3% (10/247) with more than five (see supplementary material 3c upper part). The last category includes two types: the large sized clusters, e.g., the clusters 40 and 157 which respectively contain 402 and

60 binding sites and are functionally very heterogeneous. The second type is clusters where structures are associated with many PROSITE patterns, *e.g.*, cluster 105 gathers 70 binding sites from actin structures, and is functionally a very homogeneous cluster. Indeed, 93% of the structures are associated with the three accession numbers of pattern PDOC00340 (PS00406, PS00432 and PS01132) while some structures are also associated with other patterns. For example an actin-DNase complex (PDB code 1ATN) [66] is associated with two accession numbers of pattern PDOC00711 (PS00919 and PS00918) due to the DNA complex chain. For most of the PROSITE annotated clusters, a common pattern is shared by a majority of binding sites. Some other patterns may be present but only because other chains are co-crystallized in the PDB structures.

More than 190 PROSITE accession numbers are specific to only one cluster (see supplementary material 3c, lower part). 61 are found in two clusters. Furthermore, seven accession numbers are found together in more than five clusters. The first three are from the protein kinase pattern PDOC00100; PS00107: a protein kinase ATP-binding region signature, PS00108: a serine/threonine protein kinase active-site signature and PS50011: protein kinase domain profile. In the classification, 61 of the protein kinase are associated with these 3 accession numbers which are in five different clusters. A careful analysis of the protein kinase in the different clusters is provided later in this discussion. A clear observation is that patterns associated with protein kinase are always found together in the clusters.

Functional annotation in clusters. Precisely defining the function(s) of protein structures is a complicated task. Using the MOLECULE field of the PDB files, functional annotations were extracted for all the structures of the dataset. These

assigned functions were subjected to extensive manual checks. The dataset collects 442 different functions. In this paragraph, a protein function is associated with a functional annotation extracted from the PDB files. The appraisal of the functional homogeneity in the clusters was made using the N_{eq} index (equivalent number of states, presented in [67] and demonstrated in [68]). This index is based on the Shannon entropy [69] and it gives an indication of the number of states (here functions) and their distribution within a cluster. For instance, if two functions each represent 50% of a cluster, the N_{eq} equals two, if one represents 90% and the other 10%, the N_{eq} is worth 1.13. The N_{eq} equals 1 if only one function is observed in a cluster.

163 clusters are associated with only one function. About one third contain at least two functions. 12% of the clusters (29) have a N_{eq} value equals or greater than 2 (see supplementary material 3d). Some clusters gather an important number of different functions (examples are presented below). As expected, the mean of the N_{eq} generally increases with the cluster size (see supplementary material 3e). However, there is no strong correlation. Small clusters have low N_{eq}, but several big ones also have low N_{eq} (see supplementary material 4). Eight clusters have a N_{eq} greater than 5, some are analysed in the following paragraphs.

MED-Sumo Cluster 4. The N_{eq} equals 14.92 indicating it is a very heterogeneous cluster. It contains 27 different functions. Mostly, they are epimerases, dehydratases and dehydrogrenases, *e.g.*, hydroxysteroid dehydrogrenases, the cluster binds NAD except protein Arna (PDB code 1Z7E [70]) which binds ATP. Figure 3 shows a 3D superimposition of a dTDP-D-glucose 4,6-dehydratase (PDB code 1KEP [71]) with four binding sites from cluster 4. The top-left hand figure; with UDP-Glucose-4-epimerase (PDB code 2P5Y), shows a very good and complete

superimposition. Pairs of SCFs are all around the ligands, *i.e*. the binding sites have a strong similarity. The top-right hand one represents the superimposition of the nicotinamide binding site of 1KEP with a GDP-Mannose-4,6-Dehydratase (PDB code 1RPN [72]). Interestingly, the ligand is globally well superimposed, with only a small shift. However, a more detailed analysis yields the fact that pairs of SCFs are only on the bottom left region of the binding sites. Figure 3 shows that only sub-sites are similar while the other side (top-right) is different. On the bottom left hand representation, the binding sites superimposition enables the overlap of the ligands NAD (from 1KEP) and ATP from the Arna protein binding site (PDB code 1Z7E) on their common regions (adenosine). The other side of the ligand is quite different as the left hand background of the binding site remains similar. This region of the binding site is highly conserved for all four proteins (1KEP, 2P5Y, 1RPN and 1Z7E).

Since this classification method is able to group binding site with similar sub pockets, we notice that this cluster is due to the left hand background sub pocket which is shared by functionally different proteins. This cluster also contains a protein annotated as "*hypothetical protein*" (code PDB 2D1Y). A recent study showed how MED-SuMo could help establish potential functions of proteins [64]. Here, a function corresponds better with a functional mechanism used by the protein to express its function. Superimposition of this protein's binding site with other binding sites from cluster 4 clearly illustrates a local similarity. The adjacency with the binding mode of the dTDP-D-glucose 4,6-dehydratase is illustrated in figure 3d (bottom-right hand picture). More than ten pairs of SCFs are detected whereas the best sequence match found by PSI-BLAST [1] using the SWISSPROT databank [73], has a sequence similarity score of 35%. This match is a 3-oxoacyl-[acyl-carrier-protein] reductase (SWISSPROT code Q9X248). Analysis of SCOP ids associated with each cluster

shows various behaviours. Here, all the proteins belong to SCOP family of Tyrosine-dependent oxidoreductases (c.2.1.2). Thus, proteins of this cluster have related folds; however, a weak sequence identity rate (15.4%) is calculated. It underlines clearly that our approach merges related local folds without regards to the sequence similarities, *i.e.*, the local 3D similarity of the protein interaction site directs the clustering.

MED-Sumo Cluster 33. The N_{eq} value is 8.53. However, the first observation is that the different functions of this cluster are linked to the transport through membrane, *e.g.*, cystic fibrosis transmembrane conductance, ABC (transmembrane) transporter. The co-crystallized ligands are homogeneous as only AXP ligands are present, mainly ATP and ADP. The common region is the phosphoryl part. Figure 4a shows the ligand superposition. It emphasizes that the binding similarities of the active sites cannot be around the nucleotide part as their positions are very different. The SCFs are mainly located around the phosphate groups. Some are also around the nucleotide part but as there are 25 superimposed SCFs pairs on the phosphoryl part, there are only 9 on the nucleotide part. Figure 4b shows the superimposition of 6 ligands from 6 proteins associated with different functions, *i.e.*, histidine permease (PDB code 1B0U [74]), maltose / maltodextrin transport ATP-binding (PDB code 1Q12 [75]), cystic fibrosis transmembrane transductance (PDB code 1R0X [76]), multidrug resistance-associated protein (PDB code 2CBZ [77]), α-hemolysin translocation ATP-binding protein HLYB (PDB code 2FF7 [78]) and peptide transporter TAP1 (PDB code 1JJ7 [79]). Four of them are co-crystallized with ATP and two with ADP. Figure 4b gives another example of how similarities detected by MED-SuMo can concern only a part of a binding site. The SCFs circled in white

highlight MED-SuMo's flexibility, *i.e.*, superimposition rules are loose enough to enable the superimposition in figure 4b; both SCFs represent the same chemical propriety and they are oriented in a very similar direction, however, the separation distance is large (0.5 Å).

The fact that many functions are in this cluster underlines that the binding of a phosphate ligand is not specific to one type of function. It is very common for the mechanism of transport through membrane proteins to require the energy of the phosphate transfer. Here we can say that despite the fact that there are many functions in that cluster, all proteins structures use a similar binding mode, characterized by MED-cluster 33. As for the previous presented cluster, here also, these proteins belong to the same SCOP family, ABC transporter ATPase domain-like (c.37.1.12). As previously observed, they also share a low sequence identity rate (26.2%). Figure 4 shows the methodology can highlight that the common binding part of this family is on the phosphate side of binding site, and is directly linked to function.

Other examples. MED-SuMo cluster 40 has the highest N_{eq}, 53.36. It is also the biggest cluster of the classification with 402 binding sites from 386 proteins. It includes 279 Small GTPase (72% of the cluster), in fact, all the Small GTPases of the dataset, 40 "Elongation Factor 2" (10%): all the Elongation Factor 2 of the dataset. The high N_{eq} comes from the 18% remaining; these proteins encompass a large number of functions which share the same mechanism, and interact with GXP ligands. This cluster is associated with a huge SCOP superfamily (c.37.1) corresponding to different families, mainly c.37.1.8 (80%), but also c.37.1.10, c.37.1.1, c.37.1.20 and c.37.1.4. Some proteins have no SCOP ids, but based on CATH classification, they can be considered as members of this superfamily. The important region of this

binding site is, as for MED-SuMo cluster 33, the phosphate side. Hence, our analysis underlines this similar behaviour for proteins from the same SCOP family. But, it also highlights that sites from MED-SuMo clusters 33 and 40 are clearly distinct, the first one corresponding to one superfamily and the second to four distinct superfamilies.

A few highly populated clusters also have low N_{eq} value. For instance, cluster 159 has a N_{eq} value of 1.74 and it contains 28 binding sites all from HSP70 proteins. In reality, its N_{eq} should be 1; but HSP70s are sometimes annotated differently. This verifies the fact that MED-SuMo is able to gather proteins with the same function. Cluster 105 has a N_{eq} equals to 1.07 whereas it contains 70 binding sites. This cluster includes all actins of the dataset. These proteins all bind AXP and have very specific binding modes.

In summary, MED-SMA, generates different types of cluster, some functionally very diverse while others functionally very homogeneous. Proteins with different functions that are in the same clusters; for example the small GTPase in cluster 40, can share related inhibition processes. Another example concerns the actin family, all actin proteins are in the same cluster; molecules with specific binding modes are needed to inhibit or activate them.

Links between clusters. Two clusters are linked if they both contain a site inherited from the same binding sites. A part of a binding site can be related to a protein family and another part to another protein family. Under a certain fraction, (here, parameter *covering_factor*: 60%) overlapping SCFs are considered as part of 2 separate sub-sites. They reflect binding site flexibility.

Figure 5 shows an interesting network of clusters where cluster 40, the biggest cluster of the classification, is in the centre (see supplementary material 2). This network implies eleven clusters and it can be divided into 2 parts:

The upper part involves 5 clusters, each related to the biggest, cluster 87 which is connected to cluster 40: cluster 70 (N_{eq}=1.6, size=5), cluster 69 (N_{eq}=1.0, size=2) and cluster 78 (N_{eq}=1.0, size=2) are adenylate kinase clusters. Adenylate kinases are phosphotransferase enzymes that catalyze the interconversion of adenine nucleotides. They play an important role in cellular energy homeostasis [80]. Cluster 87 (N_{eq}=6.5, size=56) also contains adenylate kinase, but also more diverse functions. Different nucleotide kinases are present, *e.g.*, thymidylate kinase or uridylate kinase. Nevertheless, all those structures are from enzymes that catalyze the phosphate transfer from ATPs to the 5' end of nucleotides. Cluster 176 (N_{eq}=2.3, size=16) contains other nucleotide kinases, *e.g.* deoxycytidine kinase (68%). Interestingly, these are not natural nucleotides. For example, the human deoxycytidine kinase is responsible for the phosphorylation of a number of clinically important nucleoside analogue pro-drugs [81].

The lower part of the network incorporates 6 clusters, all connected to cluster 40 except cluster 113. Cluster 138 (N_{eq}=1.0, size=2) is a GTPase cluster. Cluster 228 (N_{eq}=1.0, size=1) is a conserved active site with residues in the GTPase domains common to both signal recognition particle and conjugate receptor [82]. Cluster 28 (*N*eq=1.9, size=3) is small cluster of DNA polymerase III. The DNA polymerase III holoenzyme is the first enzyme complex involved in prokaryotic DNA replication [83]. Cluster 112 (N_{eq}= 1.0, size =15) gathers Heat shock locus (HSL) proteins (87%), a DNA polymerase III and CLP protease proteins. HSL and CLP have chaperone activities, being implicated in the formation of protein complexes. Cluster 245

(N_{eq}=5.9, size=33) is a heterogeneous cluster that gathers mostly F1 Atpase, with RecA proteins and even a myosin protein. More interestingly, a hypothetical protein from *Aquifex aeolicus* (O67745_AQUAE), is associated with this cluster [84].

The thirteen links are represented figure 5. Cluster 40 is connected to 5 clusters, (*e.g.*, cluster 87), cluster 70 to 3 clusters, 5 other clusters have two links and the 3 remaining, only one. All links can be illustrated with superimposition using the MED-SuMo 3D viewer. Figure 6 illustrates 2 links between 3 clusters: Cluster 245, 40 and 28. Hence, we selected 3 protein structures (PDB codes 1UM8 [85], 1SXJ [86] and 1XXI [87]). A very low sequence identity is found between the different sequences (4.4%). Thus, as expected, the global structures are quite different (see figure 6, left). However, figure 6 (right) shows that local similarity is important. The ligands, one ATP and two ADP, are closely superimposed. The bottom parts of the 3 binding sites are very similar which is highlighted by several SCFs. However, the reason why these proteins are not in the same cluster is that the similarities are only local; only SCFs on the bottom of the binding sites are well superimposed. This underlines a sub-pocket similarity which could lead to the fact that this part of the binding site could interact with the same binding modes.

Classification enrichment.

Nonetheless, a crucial question is if the protein structure has no purine ligand bound, can MED-SuMo still identify the purine ligand binding property of the protein? Using the ExPASy website for PROSITE, proteins with a purine binding patterns were selected. Since PROSITE highlights interesting regions of the protein sequences, binding sites are not always structurally defined with a ligand. The nine purine ligands were used as queries to get a protein structure list. 3515 structures were collected, of which only 880 were common to the classified PDB dataset. 1492 are co-

crystallized with non-purine ligands while 1143 are not co-crystallized with any. In these 3 subsets, we chose to associate the 1143 apo-structures to one of the 247 clusters. For this purpose, a particular MED-SuMo mode was used, it enables the comparison of whole protein surfaces to every purine binding site already classified (2115). Two filters were applied to ensure the quality of results from this strategy: (1) a high MED-SuMo score (value 5.5) and (2) a value of *covering_factor* equivalent to the one used in the MED-SMA merging step (see figure 2.2c) of 0.6. This value ensures at least 60 % of the SCFs are in common with the corresponding binding site of the cluster.

When applying the first filter only, clusters would be enriched by 1038 potential new binding sites associated with 567 of the 1143 structures without ligands (~50%). With the second filter, clusters are enriched by 203 potential binding sites issued from 196 protein structures. Here, seven structures are associated with more than one cluster. A single protein structure can have several purine-binding sites, and a protein structure can be associated with two linked clusters. For instance, the human tyrosine kinase c-Src (PDB code 1FMK [88]), not co-crystallized with any ligands, can be associated with two protein kinase clusters, clusters 157 and 211.

56 clusters are expanded in this protocol. Cluster 40, the biggest, gains 19 binding sites from 19 structures (from SCOP c.37.1.8 and c.37.1.10 families as the rest of the cluster). Protein kinase clusters 157 and 211 are, respectively, enriched by 26 and 9 of those apo-structures. It can be noted that at least 371 other structures have high similarities with purine binding sites, but were discarded by these very stringent parameters. A closer study could determine whether they should be included in the clusters or not, some being clearly positive hints.

Discussion

Methods to compare binding sites. The detection of functional sites on protein surface is important for the identification of biological activity. Most protein structures are implicated in, at least, one ligand-protein interaction, and they are implicated in the majority of critical biological processes. However, without known related sequences or structures their detection is difficult [89]. Innovative novel approaches have been proposed, *i.e.*, the use of hydrophobicity distribution on protein structures using the fuzzy oil drop model [59], the destabilization of limited protein regions [90], phylogenomic classification of protein sequences [91] or the classification of known protein catalytic sites [92]. Prediction of protein functional

sites is an important step in identifying small-molecule interactions for drug discovery [93] and to optimize the drugs targeting these sites [94]. Another valuable application is as a pre-processing step to reduce search space for rigorous computational docking algorithms.

Methods to compare binding sites have been developed using various kinds of structural descriptors, *e.g.*, CavBase uses pseudo-centers, and the strong hypothesis that chemical similarity and activity are linked [95]. In this field, MED-SuMo is an efficient approach based on Surface Chemical Features (SCFs). Each SCF represents a pertinent chemical property and is described with appropriate geometric rules. The search of equivalent binding sites is performed by detecting similar graphs where the vertices represent triangles of SCFs. The specific geometric rules of each SCF enable the heuristic to be fast. So, MED-SuMo offers an interesting and original approach to detect structural and functional similarities between protein binding sites.

Here, it is applied in a clustering approach where ligand environments are classified. An application to a particular protein family, the purinome, is presented.

AXP, NAD and GXP are simple ligands composed with related nucleotides and phosphoryl groups. Nevertheless, they are quite flexible and can adopt very different conformations within a binding site [96].

Direct comparison with other classification methods is difficult. Nebel and co-workers report a method to automatically generate 3D motifs from protein structure binding sites based on consensus atom positions and evaluate these with a set of adenine based ligands [44]. Their methodology was validated by generating automatically 18 different 3D patterns for the main adenine based ligands. Our study encompasses a larger set of proteins. The different classes presented in this study are found again by MED-SMA. Nonetheless, the classification has some differences. Hence, concerning the ADP4 pattern example (see figure 3 of [44] and associated text), three proteins of ADP4 out of five (PDB code 1EHI, 1E4E, 1GSA, 1KJQ and 1IAH) are associated with the same MED-SuMo cluster number 5 (PDB code 1EHI, 1E4E, 1KJQ). For two proteins (1GSA and 1IAH), they are associated with other clusters. In these cases, the common pattern is on the adenine binding region. The remainders of both datasets are quite different. Each cluster was systematically analysed in terms of PROSITE and SCOP ids distribution, and of sequence identities.

Classification of purine binding sites. The classification of all the co-crystallized binding sites of our dataset generates 247 distinct clusters. The clustering is quite robust as an average number of 48 SCFs are found in each sub-sites. They are mainly H-bond interactions SCF. An average of more than 80% of the binding sites' SCFs, are found in the sub-sites. Interestingly, numerous clusters consist of binding sites linked to various kinds of ligands. However, even if AXP and GXP only differ

by two chemical groups and NAD and AXP have a region strictly similar, most of the clusters are specific to one or another.

An extensive and difficult manual annotation of the protein structures was also undertaken, based on a specific PDB field. This functional annotation shows that one half of the clusters are associated with only one function and 12% with more than two. The main reason is that purine binding sites are not specific to the function of the protein but are related to an activation/inhibition mechanism. Binding sites in the same clusters could be targeted by similar drug molecules. We have presented and carefully analyzed examples of the eight clusters with high N_{eq}. Cluster 40 has the highest N_{eq} (53). It contains binding sites from 279 Small GTPase, and is associated with almost a hundred different functional annotations. However, interestingly it comprises all small GTPases of the dataset, all sharing the same functional (inhibition and activation) mechanism through the binding of phosphate side of the purine ligand and the P-loop. SCOP superfamily c.37.1 is well represented here by SCOP family c.37.1.8. The presentation of links between clusters emphasizes the complexity of this approach. Indeed, a link is represented by a binding site for which constituent sub-sites are found in two different clusters. A region of a binding site is found in one cluster and another region is in another one, potentially, slightly overlapping (the overlapping must be less than the *covering_factor* threshold). This property is not only a difference due to protein flexibility but it is a true distinction in protein binding sites which can highlight sub-pocket similarities.

PROSITE patterns. A pertinent result of this study is from the analysis of distribution of PROSITE patterns. Even when numerous clusters have some different PROSITE patterns, they may remain quite homogeneous as most have redundant

common patterns or related patterns. Our work also extends the creative works of Kasuya and Thornton made on the 3D-structure analysis of PROSITE patterns [97]. They found numerous PROSITE patterns with common three-dimensional structure characteristics which could be used to create templates defining 3D functional patterns. Wu and co-workers [98] recently improved on a previous study [99] showing that 3D information is significantly more relevant than PROSITE patterns. Our work suggests that common and distinct characteristics can be associated with a given pattern and that distinct patterns share common local features. In the same way, our analysis highlights the interest in enriching PROSITE annotations for related protein sequences and structures. Indeed one third of proteins from our dataset are not annotated with PROSITE patterns. We have also demonstrated that this binding site classification can be further enriched by apo-structures. Indeed, MED-SuMo can be used to first detect their binding sites and then SCFs signature can be compared to those within the clusters.

Protein kinase. They play a central role in cell regulation pathways in eukaryotes species [100]. As they represent the second largest drug target family for pharmaceutical companies, a chemogenomics concept, called kinomics [101] has been deeply explored. Although they essentially catalyze the same phosphoryl transfer reaction, they are involved with a remarkable number of different substrates, structures, and cell pathways. Analysis and classification of protein kinases have been made at the genomic sequence level with elegant approaches, such as KinG [102] which allowed the identification of novel kinases as in *Plasmodium vivax* genome leading to new classifications. Some kinase studies have combined sequence and

structure classification. Inhibition of kinase activities have been treated at the genomic level and analyzed with respect to these classifications [103].

In 2002, Manning et *al*. [104] established a standard protein kinase classification, the Kinome. This classification is sequence based and it highlights seven main families: TK: tyrosine kinase, TKL: tyrosine kinase-like, STE: Homologs of yeast Sterile, CK1; Casein Kinase 1, AGC; Protein kinase A, C, G, CAMK: Calcium/calmodulin-dependent protein kinase, CMGC: containing CDK, MAPK, GSK3, CLK families. An atypical kinase protein family was also described, containing all uncategorized kinase proteins. A point of interest in this analysis is the distance used in the classification which is solely based on local 3D similarity while the Kinome is based on complete sequences.

In the classification we present, protein kinases are present in seven clusters (46 (N_{eq}=2.14, size=11), 100 (N_{eq}=2, size=2), 121 (N_{eq}=2, size=2), 155 (N_{eq}=1.96, size=5), 157 (N_{eq}=17.71, size=60), 183 (N_{eq}=6.41, size=21), 211 (N_{eq}=11.29, size=23)). The analysis of the populated clusters (size>5) highlights particular aspects of the classification:

The most homogeneous cluster of this particular analysis is the MED-cluster 46. However, it does not contain only one type of kinase: 9 proteins are from the AGC and one from PTK family.

Cluster 211 has a high N_{eq} but a more detailed analysis shows that it is very pure cluster with respect to the Kinome classification. Almost all its binding sites are from two branches of the kinome tree, the PTK and CMGC. Only one other protein is from another small distinct branch between PTK and CMGC (PDB code 2A19 [105]). Other members of this kinome family are also found in cluster 157.

Cluster 157 has a high N_{eq} value, and 59 out of 60 binding sites are from protein kinase. However, they are from 3 different Kinome families (CMGC, TK and CAMK) and from one atypical. Even if their sequences are different (which is why they are in 3 distinct families), their ATP binding sites have strong local structural and functional similarities detected by MED-SMA. To understand the disparity of the protein kinase families in this cluster, the ATP binding site of a cell division kinase 2 (CDK2, PDB code 1B38 [106]) from the CMGC family was compared with the remainder of cluster 157 using MED-SuMo. Figure 7 represents the 35 first hits of this MED-SuMo analysis. The first observation is that all CDK2 of the dataset are found by MED-SuMo (dark blue). The second observation is that protein kinases from the same family (CMGC) are also in the hit list (light blue). With only 23.54 % sequence identity, CDK2 and glycogen synthase kinase 3-BETA (GSK3β) are gathered the same kinome family (CMGC). A SAR study on protein kinase structures available in the PDB in 2004 [101] found similarities between activities of proteins from those 2 families. MED-SuMo detects local structural and functional similarities and MED-SMA classifies them in the same cluster. It underlines the functional interest of our classification approach. The final point concerning this cluster derives from the observed presence of other colours shown in figure 7: PTK (green), CAMK (pink), TKL (grey) or atypical protein kinase (light red). The ANP binding site of a protein tyrosine kinase, *Aurora-A* (PDB code 2DWB) has higher score than some of the CDK2. Thus, the binding sites of 2DWB and 1B38 are more similar than 1B38 and other CDK2s (*e.g.* τ-protein kinase I: PDB code 1J1B [107]). This highlights surprising structural and functional similarities between a CDK2 and a aurora-A tyrosine kinase whereas they are not in the same kinome family. An experimental study realized by Pevarello *et.al.* show that CDK2 and Aurora-A activities can be

inhibited by the same molecule classes; 1,4,5,6-tetrahydropyrrolo[3,4-c]pyrazoles [108]. If they bind the same types of molecules, it implies that they are inhibited or activated by the same drugs, sharing binding modes of those molecules. MED-SMA clearly highlights this same fact by showing similar structural and functional properties gathered at those binding sites in the cluster.

The final remark on protein kinases concerns the enrichment protocol results. Many clusters have increased their number. For instance, cluster 157 and cluster 211 are respectively both enriched by 26 and 9 apo-protein structures. Hence, protein kinases sharing other types of ligand might also be associated with MED-clusters if the rest of the PROSITE dataset had been added to this enrichment protocol. Classification of datasets, with or without ligands, follow similar rules. Thus, the classification of protein kinases is often quite similar to the Kinome described by Manning and co-workers [104], which is logical as related sequences share functional similarities. Nonetheless, some striking exceptions are grouped in the same cluster, protein structures from different part of the Kinome. This is also logical as MED-SMA clusters local 3D surface similarities and not sequences. Moreover, experimental results support these associations, reflecting functional similarity across the Kinome.

MED-SMA utility. This type of relationship between families is very interesting and their detection is a direct application for MED-SMA. In this classification, we chose to fix a high MED-SuMo minimal score, (5.5 corresponding to at least 10 superimposable SCFs) in order to obtain functionally pure clusters. Other potential uses for this classification method are: deduction of enzymatic mechanism of poorly studied or newly discovered proteins, or in other cases, protein function deduction.

Thus, we can validate the assertion that functions can be assigned to unknown proteins by finding which cluster(s) are best matches for the concerned structures. Matching to clusters rather than single structures will diminish a significant amount of the noise. All described applications are based on a potential presence of better known binding sites in the same cluster. Other applications are planned, a complete protein kinase classification with no ligand type filter. We are also studying the results of a classification of all binding sites of the PDB which is a fairly substantial undertaking.

Link. One last interesting link is observed between cluster 56 (N_{eq}=1.76, size=12) and cluster 121 (N_{eq}=1, size=2). Cluster 56's main function is DNA topoisomerase II while cluster 120 contains a "histidine kinase". The link is due to the presence of a "histidine kinase" in cluster 56. A review [109] outlines the fact that similarities are found between diverse ATP binding proteins. In fact, they report that histidine kinases are related to the superfamily GHKL ATPase (Gyrase, Hsp90, Histidine Kinase and MutL). Other studies report that they are inhibited by the same drug, the radicicol. In a previous study [63], MED-SMA underlines these local similarities by collecting binding sites from these 4 families into a single cluster and illustrates them with a 3D view of their superimpositions around the drug radicicol (see figure 8 of [63]).

Conclusions

This approach is clearly embedded in the structural genomics field. It is fast and, as noted by Ferrè *et al.* [110] functional patches associated with a large collection of protein surface cavities can be used to provide functional clues to protein with unknown structures. This observation is relevant to the present study. Thus, MED-

SMA is an approach that may improve the efficiency and effectiveness of early stage drug discovery steps, involving the initial lead selection, improving poor leads, or, multivariate optimization, as it was used in a previous study [17]. This study demonstrates that MED-SuMo is a particularly well suited tool to both annotate protein structures and to enable structural functional classification. Finally, its effectiveness at dealing with the entire PDB shows that MED-SuMo is well-suited to large-scale applications.

Materials and methods

Protein structure database.

The dataset was built using the PDB [20]. X-ray protein structures co-crystallized with ATP, ADP, AMP, ANP, GTP, GDP, GMP, GNP or NAD were extracted. The final PDB dataset contains 2229 protein structures. To avoid a too large database, we chose to include only one binding site per type of ligand for each structure file. At the end, the MED-SuMo database contains 2,322 binding sites.

The PROSITE database [3] was also considered as it gathers protein domains, families or functional sites through more than 4300 sequence patterns or profiles. Each ligand name was used as a query to regroup related PROSITE patterns and profiles on the ExPASy website [111]. For example, 'ATP' is associated with the pattern PDOC00017. It corresponds to "ATP/GTP-binding site motif A (P-loop)". The PDB structures containing those patterns or profiles were used to gather a secondary dataset of 3,515 protein structures. As most of purine binding proteins are not co-crystallized with purine ligands, only 880 protein structures are in both datasets.

MED-SuMo Algorithm.

MED-SuMo is designed to localize similar regions associated with a defined function [16, 60, 61, 64]. Its main advantage is to detect binding sites with similar or related binding modes which could not be identified using rigid (or even flexible) superimposition approaches. Its heuristic is based on a 3D representation of macromolecule structures using precise Structural Chemical Features (SCFs). For MED-SuMo, a protein structure is represented by a set of functional groups: unbound hydrogen bond (Hbond) donors or acceptors, accessible sides of aromatic rings and carboxylate groups, primary amide, guanidinium, hydroxyl, imidazole, thioether and thiol groups. Each feature associates its chemical characteristics to precise geometrical properties. MED-SuMo comparison methodology (see figure 1) can be divided into two major steps:

(1) *The Graph Formation:* SCFs are displayed on the protein structure through a lexicographic analysis of the atoms in the PDB files, *i.e*., for each residue type, a list of predefined SCFs is specified (see figure 1a). For example, a phenylalanine is represented by two H-bond acceptors, one H-bond donor, one aromatic and three hydrophobics. Once all SCFs are assigned, their positions and orientations are filtered to discard those likely to be involved in intra protein interactions and those too buried to interact with a potential ligand (see figure 1). Remaining SCFs are assembled into triangles with specific geometric characteristics e.g. edge sizes, perimeter, angles (see figure 1c). The triangle network is represented as a graph data structure where triangles are vertices and edges connect adjacent triangles. All graphs are stored in the MED-SuMo database (see figure 1d).

(2) *The Graph Comparison*: To compare two graphs, MED-SuMo detects compatible triangles made of compatible SCFs (see figure 1e). Compatible triangles are called comparison “seeds”. When a seed is detected, MED-SuMo extends the

comparisons to the neighbourhood vertices, until no more similarities are found. This list of compatible triangles is then used to create a list of SCFs pairs which are called "*patches*" and which represent the MED-SuMo hits. Those hits are then organized according to their score [60] (see figure 1f). Angle tolerance between the pairs of compatible triangles enables MED-SuMo to include flexibility in the comparison.

Comparisons are usually made between a query and a database of precompiled graphs. Three kinds of MED-SuMo databases exist: the binding site database made of the small protein regions characterized by co-crystallized ligands and small peptides. The full surface database which contains whole surfaces of the protein structures and the MED-Portions database also containing small protein regions but where characterized by chemical fragments detected in the ligands or peptides from the PDB [16, 17].

The original version of SuMo is available on the internet [61] but the latest improvements are only included in the MED-SuMo software distributed by MEDIT SA [62]. These improvements concern the definition and conception of protein databases and as well as the heuristic itself. One of the most important new features is the Graphical User Interface (GUI). Indeed, the MED-SuMo GUI offers a simple, yet powerful, front-end to MEDIT's technology. To start a MED-SuMo search, the user loads its query protein in a 3D viewer. The binding sites are automatically detected by the presence of co-crystallized ligand or peptides. It is possible to either select one binding site or to define a manual selection. Once the selection is made, the user can launch the search on a one of the three available databases (binding sites, full surface or MED-Portions). The hits detected by MED-SuMo are displayed in a result table with columns containing the ligand 2D structure, the MED-SuMo score, the SCF signature, the RMSD of the corresponding SCF and other features (see figure 8). The

protein structures and the co-crystallized ligands can be superimposed in the 3D viewer thanks to the transformation matrix calculated by MED-SuMo for each hit.

Classification of protein binding sites.

As noted, MED-SuMo has an interesting and original approach to detect structural and functional similarities between protein binding sites. Its ability is now used to classify datasets of structures and this method is called MED-SuMo_Multi Approach (MED-SMA). It operates in three major steps: comparison of all the binding sites of a dataset using a pairwise comparison system. Detection of matching regions in the binding sites to build a similarity graph and finally, classification of this graph with the Markov Clustering algorithm (MCL [65]). Figure 2.1 illustrates the global procedure and Figure 2.2 depicts the 6 consecutive steps of the algorithm.

Algorithm Description. To begin, a list of proteins is selected (see protein database paragraph). To build the MED-SuMo database, two strategies can be adopted: (i) the database contains all binding sites of the selected proteins, (*i.e.*, binding sites where the co-crystallized ligands obey predefined rules including maximum (or minimum) number of atoms, number of residues if it is a small peptide; (ii) the database contains only specified binding sites, for example, only purine binding sites. Once the database is created, the comparison is launched using MED-SuMo pairwise comparison procedure (see figures 1 and 2.2a). The main parameter at this step is the minimal score tolerated by MED-SuMo *i.e.*, *score_min*. These comparisons highlight pairs of compatible SCFs between pairs of binding sites. At this step, if binding sites are isolated *i.e.* they do not match with any other binding sites of the dataset; they are considered as the singletons of the classification and thus won't be included in the clusters (see figure 2.2b). The detected SCF groups are called

patches (see figure 2.2c). A binding site can contain several patches which associates itself to several binding sites. For each binding site, all of its patches are parsed. If two patches from the same binding site share enough SCFs, they are merged in multipatch. If two merged patches are associated with two distinct binding sites, the multipatch (sub-sites) underlines common SCFs between the considered binding sites and the two concerned binding sites. The threshold that defines the number of common SCFs needed is the parameter *covering_factor*. This is set to 0.6, meaning that 60% of the SCFs must be shared. A multipatch is a set of SCFs common to several binding sites of the dataset; they are named sub-sites in this publication. They represent the true meaningful common regions of binding sites. They ensured two properties: (i) enough SCFs are in common, *i.e.*, binding sites are structurally and functionally related and (ii) they can underline sub-pocket similarity. To compute the similarity graph, the MED-SuMo score between matching sub-sites is calculated (see Figure 2.2e). At the end, MCL interprets the graph and classifies the protein binding site dataset into clusters of sub-sites (see figure 2.2f). A 2D plot of the clusters can be visualized using dedicated tools such as Biolayout [112, 113].

Classification Analysis.

A critical question when considering clustering methods is the quality of the data association within the clusters. For protein classification, anticipated results are different between sequence classifications, structural classifications, and different again with functional classifications [103, 114-116]. PDB files have many extractable annotations. The HEADER field, for example can give specific information about the protein function. However, we used the MOLECULE field as it gives more precise information regarding functional annotations of the protein structures. For example,

protein kinase (*e.g.*, PDB code 1B38, 1QMZ) are annotated as "TRANSFERASE" in the HEADER field, whereas the MOLECULE field specifies "CELL DIVISION PROTEIN KINASE 2". To evaluate the cluster homogeneity, an entropy-derived function is calculated for each cluster and then globally for the whole classification. This index is named N_{eq} for "equivalent number of states" [69]. It assesses the conditional equivalent number of the predicted states given the observed states. In our study, it is equivalent to the equivalent number of functions per cluster. First, the entropy of the cluster c, $H(c)$, is computed. Then, the N_{eq} is calculated, its expression is the exponential of Shannon entropy [117], $H(c)$.

$$N_{eq}(c) = e^{H(c)} \quad \text{with} \quad H(c) = -\sum_{i=1}^{F} p(i_c) \ln p(i_c) \quad (1)$$

Where $p(i_c)$ is the probability of the function i in the cluster c, and F is the count of observed functions. So, N_{eq}(c) varies between 1 (*i.e.*, only one function in the cluster) and F (*i.e.*, each structure function is different). The N_{eq} calculation is made on the MOLECULE fields of all the PDB files which were manually checked extensively and on which a few manual fixes were made.

Average sequence identity has been computed for each cluster thanks to CLUSTALW software [118].

Classification enrichment.

Our PROSITE dataset is also composed with purine binding protein structures. It contains 3,515 structures of three types: 1,492 are not co-crystallized with purine ligands, 880 are common to the PDB dataset, and are already included in our clusters and 1,143 are apo-structures (protein structures with no ligands). Apo-structure proteins are hard to study as they require the analysis of their whole surfaces. Moreover, MED-SMA has only been used with binding sites. However, an interesting functionality of MED-SuMo is that it can deal with whole surfaces and is able to

localize interesting binding regions on full protein surfaces [64]. To do so, MED-SuMo compared the full surface of a protein to a binding site database composed with experimentally defined binding sites from the PDB. As mentioned earlier, the classified MED-SuMo database is made with all purine binding sites from the PDB. In order to localize purine binding sites on those PROSITE apo-structures and to identify in which cluster they could belong to, we arranged the following enrichment protocol in three steps: (1) all full surfaces of the protein structures are compared to the purine binding sites database. (2) Results are filtered according to the MED-SuMo score (value 5.5, same as the parameter *score_min* in the classification protocol). (3) SCF signature analysis; a structure enriches a cluster only if it shares 60% of its SCFs with at least one binding site of the cluster. For computation reasons, huge PDB files were excluded. Based on 1143 structures, 1130 MED-SuMo runs were launched.

Implementation.

MED-SuMo server is written in OCaml. This language is suited for large-scale software engineering [117]. External libraries are used, including MLsqlite: a sqlite wrapper for OCaml; zmarshall, a compression file manager; findlib: a package management system for Ocaml.

MED-SuMo is a client-server application and uses a scripting language to process calculations requested by the remote interfaces. A Lua interpreter is embedded in the MED-SuMo code, using the Lua-ML library [119]. It enables the use of MED-SuMo's internal functions through very simple Lua scripts. The MED-SuMo_Multi module was added to MED-SuMo core and enables the classification of any binding site dataset. Lua scripts are used to create the database for the classifications. MED-SuMo jobs can be parallelized for several CPUs of multiprocessor computers and recent development at MEDIT has enabled MED-

SuMo to be distributed across a HPC (High-Performance Computing) cluster. Ongoing development is concerned with parallelization of the entire classification method.

Software Availability and requirements.

The standard MED-SuMo mode to query 3D interaction surfaces against binding sites databases or full surface databases is commercially available with the MED-SuMo Graphical User Interface. For the moment, MED-SMA is only available in a command line mode for integration with wider workflows. However, a web-based interface was developed to interactively explore the generated clusters. This will be available freely on the internet during the year 2010. MED-SuMo is commercial software, and further information is available at http://www.medit-pharma.com/. Parties interested in commercial evaluation of this technology can contact MEDIT SA to obtain free temporary licenses (info@medit.fr).

Researcher from the INSERM Institute UMR-S 665 has no financial interests in MEDIT SA and collaborates with this company only for the present project. Therefore, MEDIT SA has the exclusivity for MED-SuMo sales.

Acknowledgements.

The authors are indebted to S. Adcock for useful comments on the manuscript and to all researchers who deposit their structures in the PDB. This work was supported by National Institute of Health and Medical Research (INSERM), National Institute of Institute of Blood Transfusion (INTS), University Denis Diderot - Paris 7 and Ministère de la Recherche. ODA's PhD is financed by the French technical research association (ANRT) through a CIFRE grant. MEDIT SA retains all rights on the presented methodology.

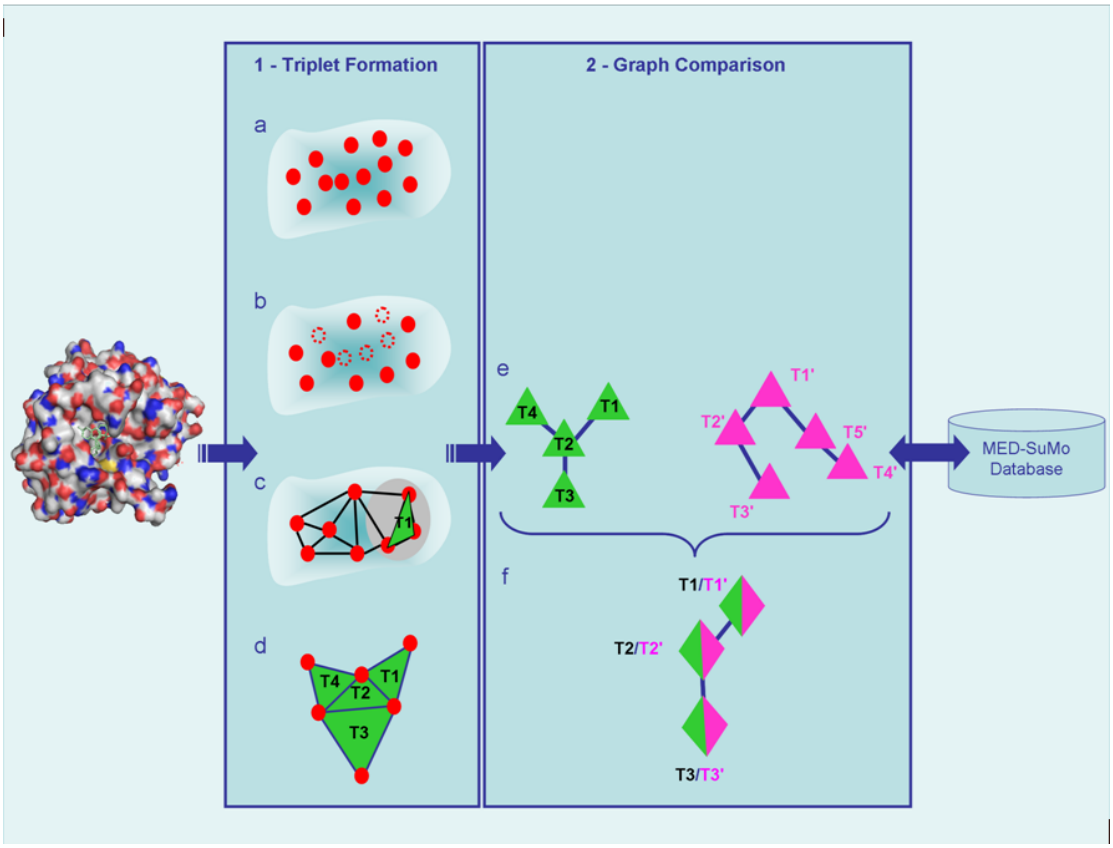


Figure 1 - *MED-SuMo Comparison Procedure*.
(1) Graph construction: (a) Surface Chemical Features (SCFs) are displayed on the protein structure through a lexicographic analysis of the PDB files. (b) Their positions and orientations are checked to discard SCF potentially involved in internal interactions or associated with buried atoms. (c) SCFs are gathered in triplets. (d) The triplet network is then stored as a graph data structure with the triplets as vertices and with edge connecting adjacent triplets.
(2) Graph Comparison: (e) The query graph (in green color) is compared to the database graphs (in pink color), compatible triplets are detected, *i.e.*, they are formed by compatible SCFs. (f) Corresponding graphs are hits found by MED-SuMo. See [60] and [63] for more details.

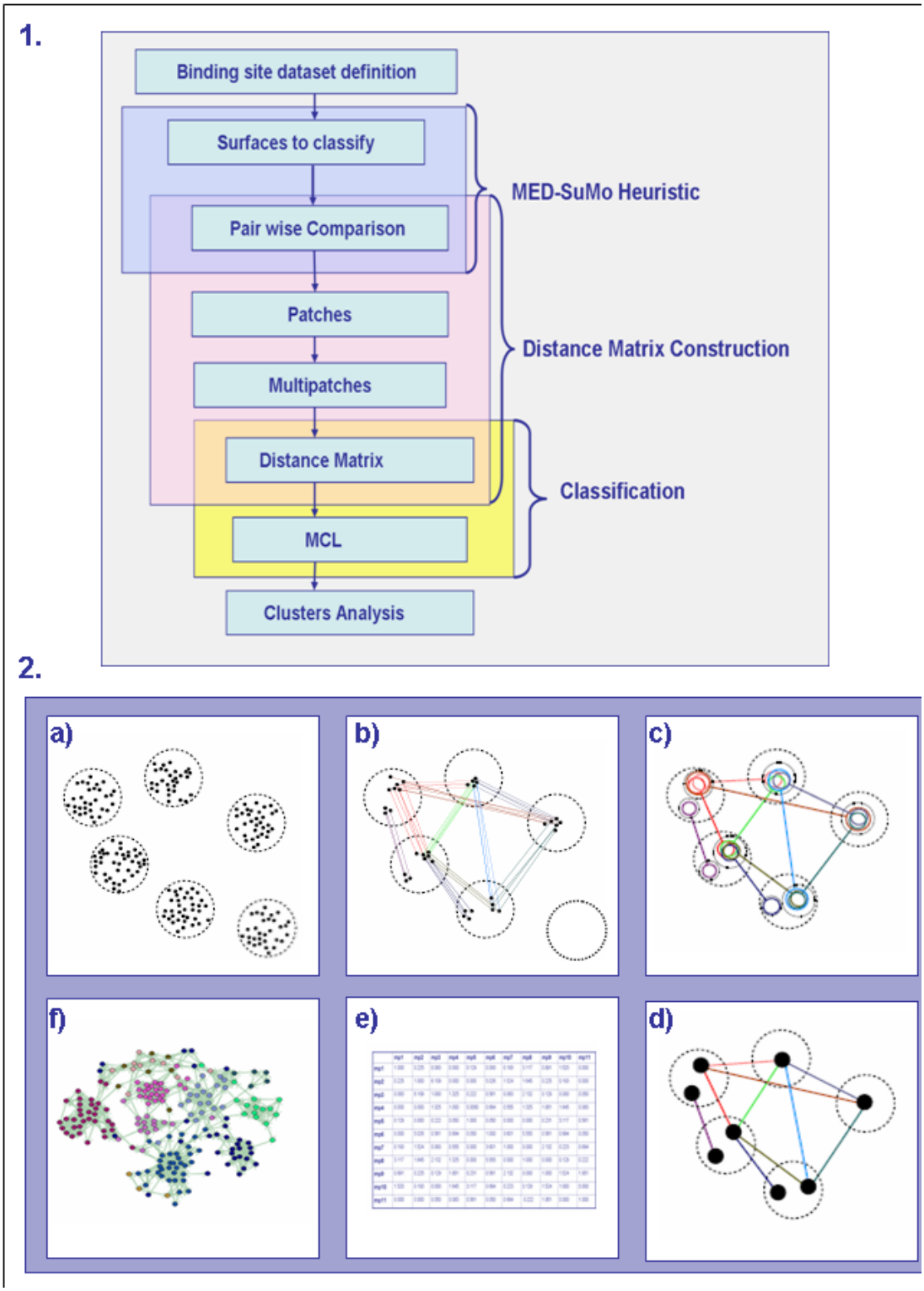


Figure 2 - *MED-SuMo Classification Procedure*.
Global steps of binding site classification heuristic. MED-SuMo_Multi approach (MED-SMA) can be divided in 3 major steps: Pairwise comparison (purple box), Similarity matrix construction (pink box) and the Markov Cluster Algorithm classification (Yellow box). **2.** *Six steps of the MED-SMA*. (a) Dataset construction, here, 5 binding sites are shown. The black dots represent SCFs. (b) Common SCFs detected by the pairwise comparisons. They are linked by a colored line; each color stands for matching sub-sites. (c) Matching SCFs between pairs of binding sites are

grouped in patch. A patch is a colored circle on the binding sites; some of them are overlapping. At this step, the network forms a graph structure where complex component are searched.(d) Parsing all patches from every component found, if overlapping patches have a certain amount of common SCFs (superior to a threshold value: parameter *covering_factor)*, they are merged in multipatches (black bold circles). (e) MED-Sumo scores between multipatches are calculated to create a similarity matrix. (f) The similarity matrix is used by Markov Clustering (MCL) algorithm to classify the dataset. Finally, Biolayout is used to visualize the cluster mainly in 2D.

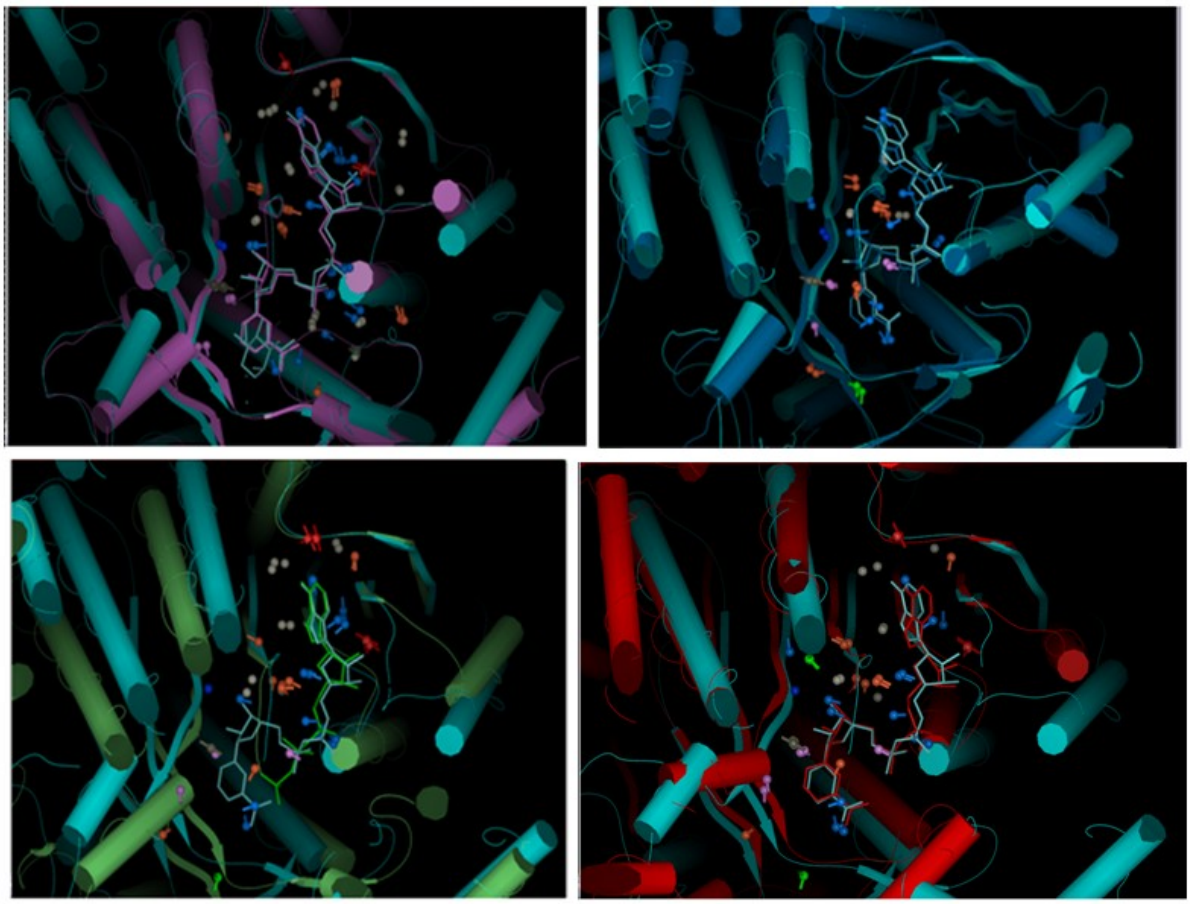


Figure 3 - *Example of binding site superimposition*.
dTDP-D-glucose 4,6-dehydratase (PDB code 1KEP) binding site is superimposed with 3 different binding sites of its cluster (MED-SuMo cluster 4). Each has different functions and the superimposition involves different parts of the binding site. (1) Global superimposition between 1KEP and UDP-Glucose-4-epimerase (PDB code 2P5Y). (2) "Left" side of the NAD with GDP-Mannose-4,6-Dehydratase (PDB code 1RPN) and (3) The nucleotide side of 1KEP binding site with the ATP of Arna protein binding site (PDB code 1Z7E). (4) Superimposition with a "hypothetical protein" (PDB code 2D1Y).

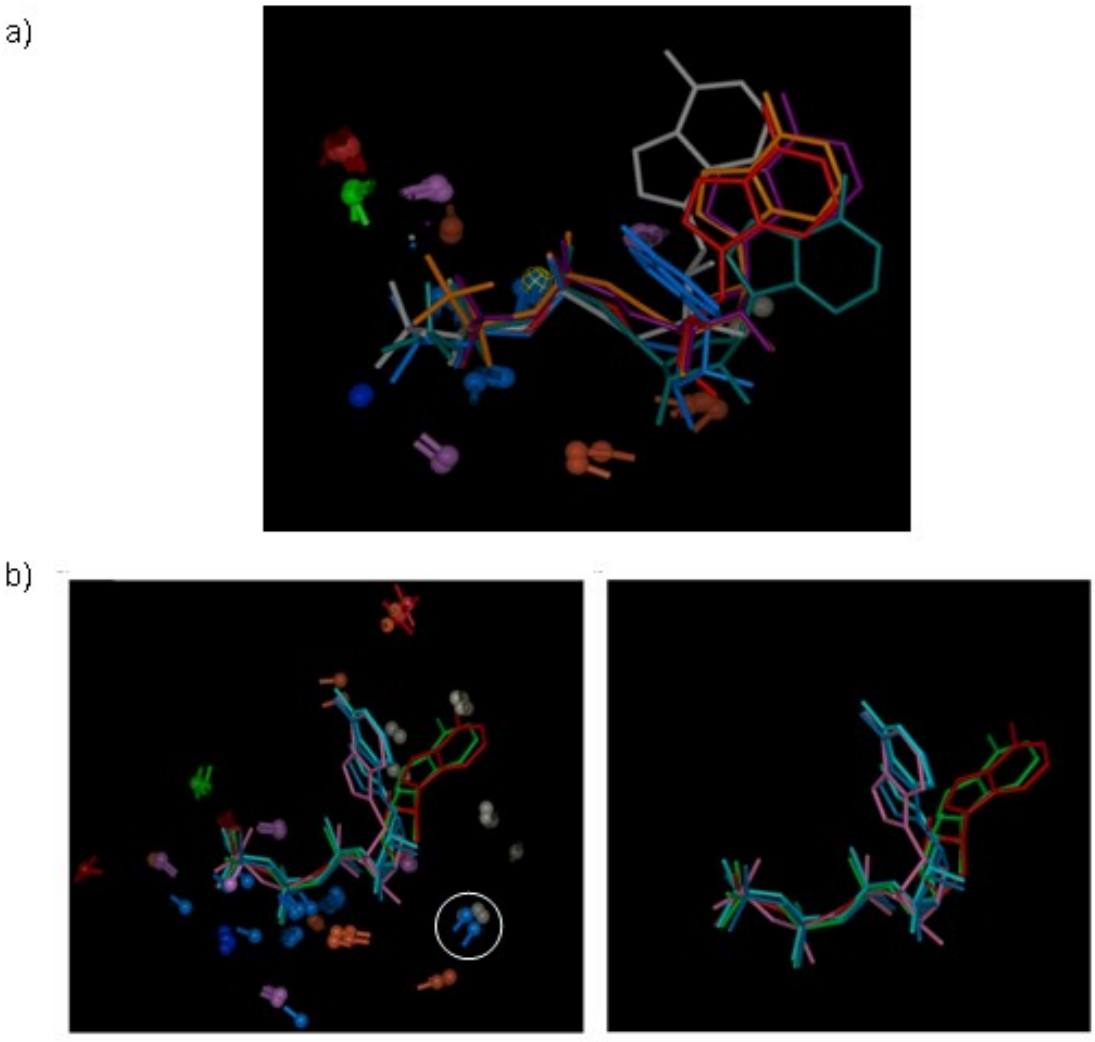


Figure 4 - *Superimposition of 6 ligands from cluster 33*.
Each ligand is taken from proteins with different functions. 4 ligands are ATP (taken from 2CBZ in grey color, 1B0U in orange color, 1R0X in blue, 1Q12 in green), and, 2 ADP (from 2FF7 in red and 1JJ7 in purple). b) Superimposition of ligands from cluster 33. (a) With and (b) without the SCFs. The phosphoryl groups are similarly arranged in each binding site, but the nucleotide region has two major conformations.

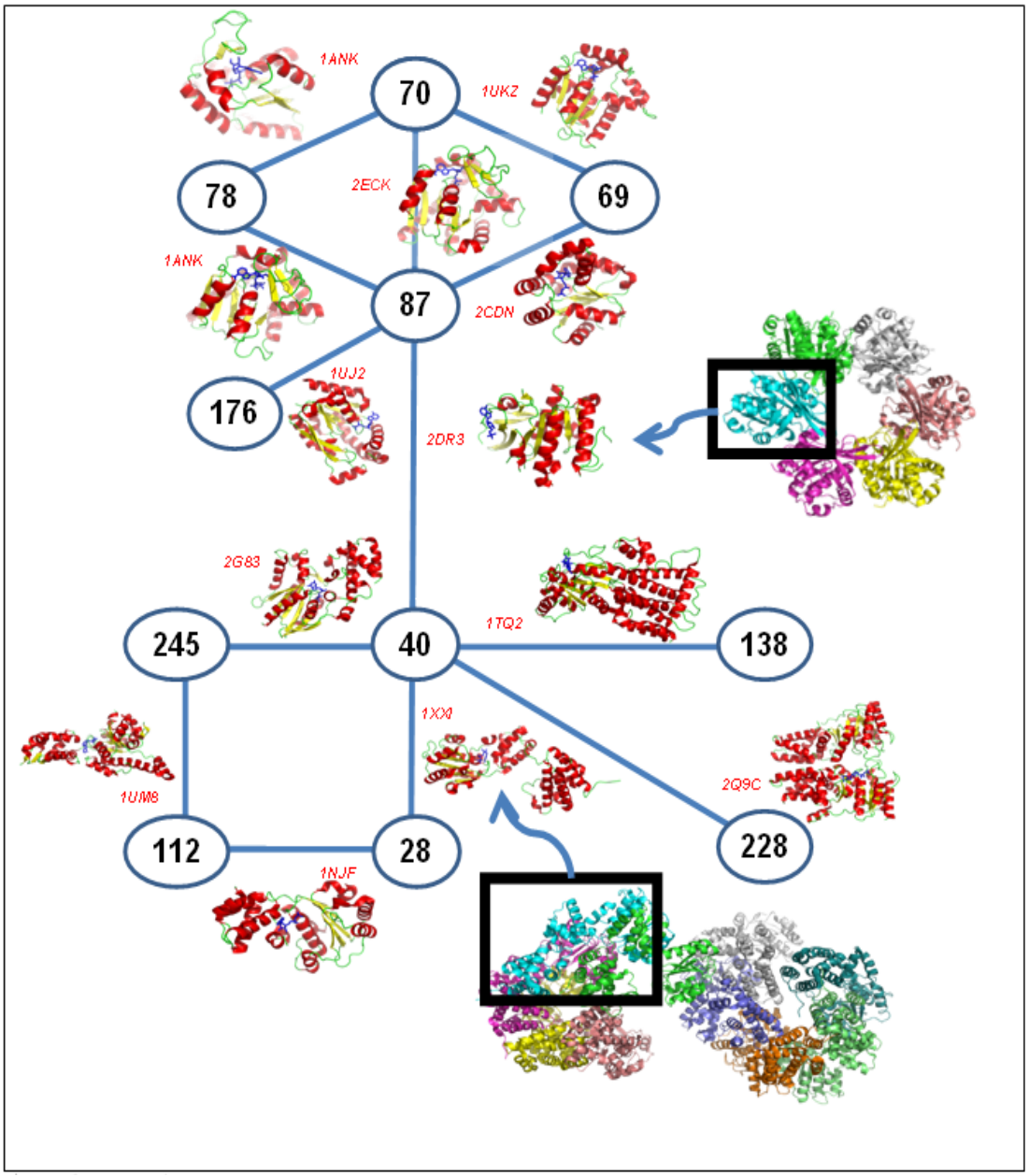


Figure 5 – *Representation of a network within the classification*.
Each circled numbers are cluster IDs. If two are connected, it means that they share a binding site. The protein structures represented on the lines are the ones containing the shared binding site. For example, 1XXI's ADP binding site is within cluster 40 and 28. Here 13 links are represented involving 11 clusters. Those links underline how MED-SMA highlights sub-pocket similarities (see Figure 6).

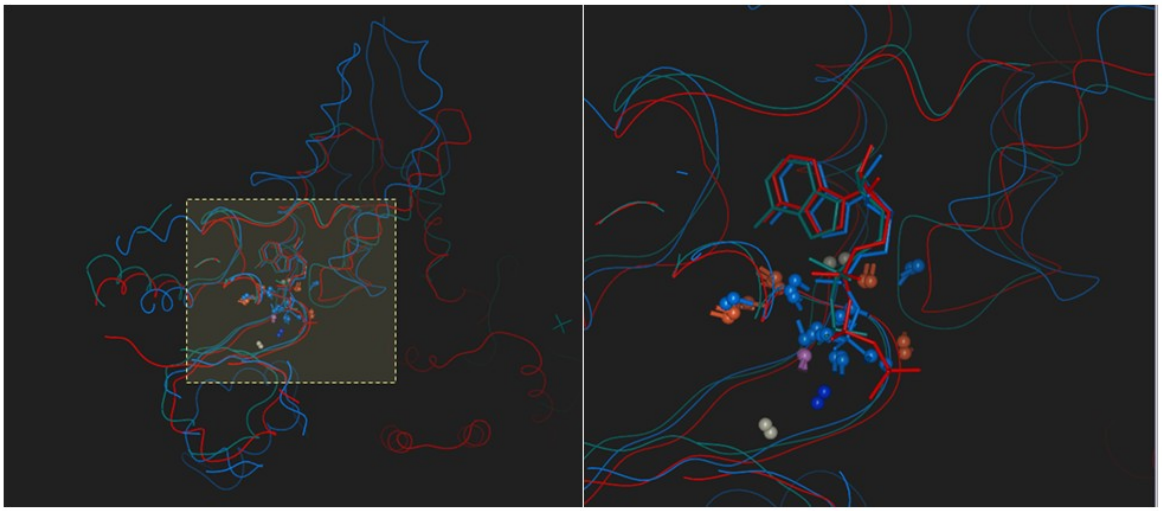
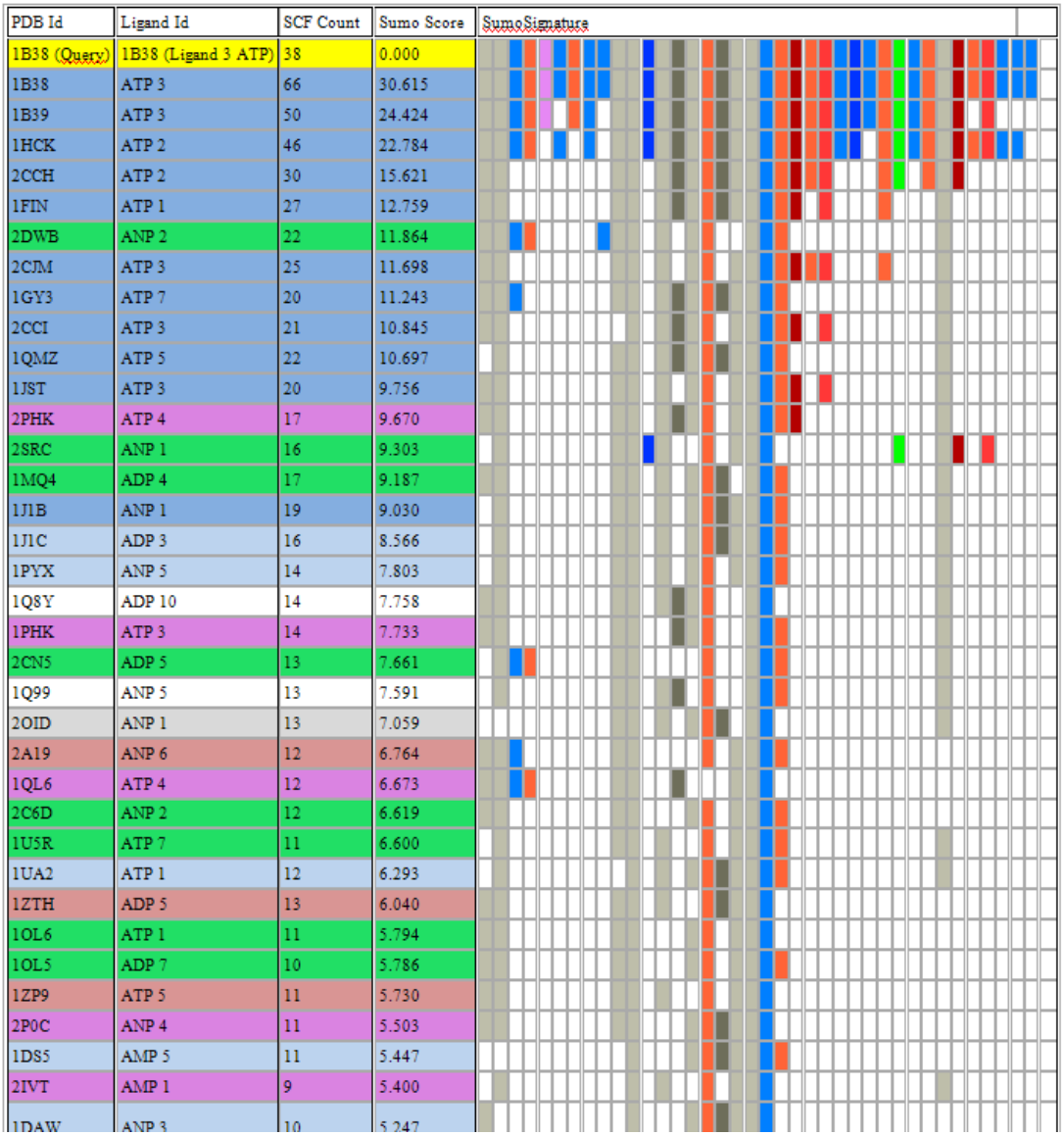


Figure 6 - *Illustration of 2 inter-clusters links***.**
On the left is represented a 3D superimposition of 3 structures from clusters 40, 112 and 245; PDB codes 1UM8 (blue), 1SXJ (red) and 1XXI (green). These proteins have a very low sequence identity, and they have very different fold. On the right, a closer view is represented, delimited by the yellow box on the left. Ligands, 1 ATP and 2 ADP are very well superimposed and the local similarities of the binding sites are highlighted by several SCFs and are very distinct. These proteins are not in the same cluster as their similarities are very local, only SCFs on the bottom of the binding sites are well superimposed. This is a typical example of sub-pocket similarity.



PDB Id	Ligand Id	SCF Count	Sumo Score
1B38 (Query)	1B38 (Ligand 3 ATP)	38	0.000
1B38	ATP 3	66	30.615
1B39	ATP 3	50	24.424
1HCK	ATP 2	46	22.784
2CCH	ATP 2	30	15.621
1FIN	ATP 1	27	12.759
2DWB	ANP 2	22	11.864
2CJM	ATP 3	25	11.698
1GY3	ATP 7	20	11.243
2CCI	ATP 3	21	10.845
1QMZ	ATP 5	22	10.697
1JST	ATP 3	20	9.756
2PHK	ATP 4	17	9.670
2SRC	ANP 1	16	9.303
1MQ4	ADP 4	17	9.187
1J1B	ANP 1	19	9.030
1J1C	ADP 3	16	8.566
1PYX	ANP 5	14	7.803
1Q8Y	ADP 10	14	7.758
1PHK	ATP 3	14	7.733
2CN5	ADP 5	13	7.661
1Q99	ANP 5	13	7.591
2OID	ANP 1	13	7.059
2A19	ANP 6	12	6.764
1QL6	ATP 4	12	6.673
2C6D	ANP 2	12	6.619
1U5R	ATP 7	11	6.600
1UA2	ATP 1	12	6.293
1ZTH	ADP 5	13	6.040
1OL6	ATP 1	11	5.794
1OL5	ADP 7	10	5.786
1ZP9	ATP 5	11	5.730
2P0C	ANP 4	11	5.503
1DS5	AMP 5	11	5.447
2IVT	AMP 1	9	5.400
1DAW	ANP 3	10	5.247

Figure 7 – *Comparison of a cyclin dependent kinase 2 to all binding sites of the classified dataset*.
Analysis of the MED-SuMo results of the comparison of the ATP binding site of a cyclin dependent kinase 2 (CDK2) (PDB code 1B38) to all purine binding sites of the dataset. The query line is yellow. Each colour corresponds to a kinase protein family: CMGC (dark blue are CDK2, light blue are other kind of CMGC); PTK (green); CAMK (pink), TKL (grey) or atypical protein kinase (light red). The white lines are non-human kinase proteins; they are not associated with any Kinome families.

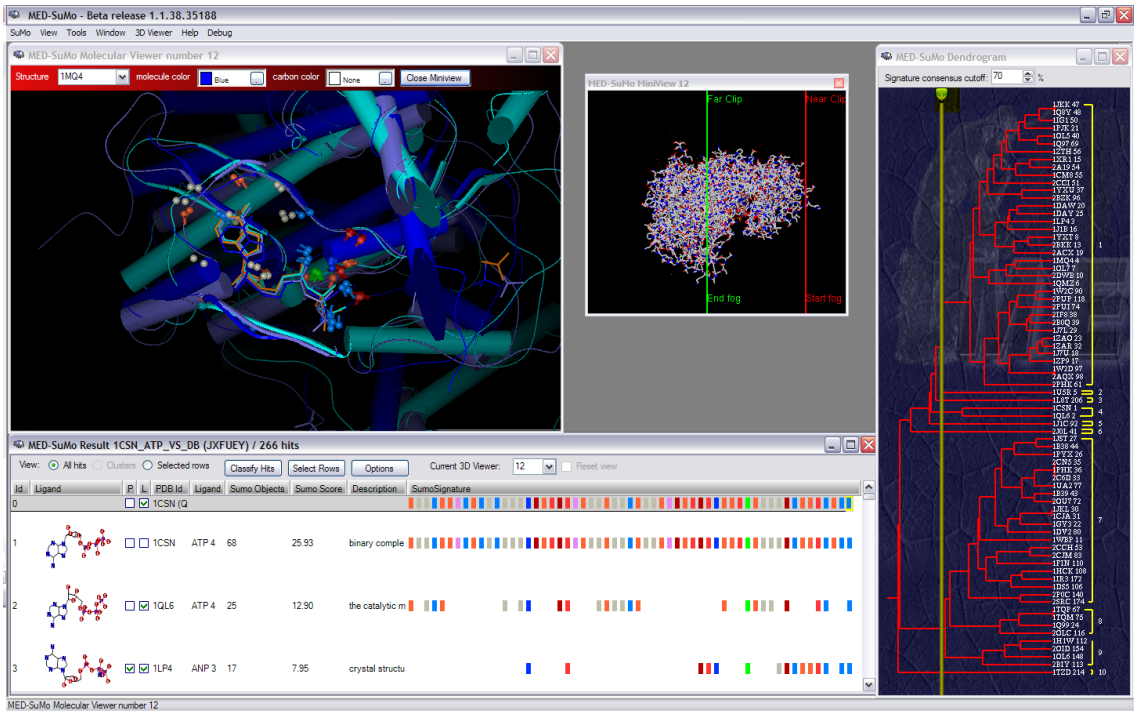


Figure 8 - *MED-SuMo Graphical User Interface*.
Four different windows are shown: (1) the 3D viewer is at the top left of the Figure, here 3 elements are superimposed: the protein binding sites, the corresponding co-crystallized ligands, represented in stick and finally, the SCFs that enabled that superimposition. The viewer is an ActiveX control that allows the user to move the structures e. g. rotate, translate to visualize the hits as desired. (2) The result table window: the first line corresponds to the query name and its corresponding SCF signature. All other lines correspond to hits found by MED-SuMo. They are originally sorted by decreasing MED-SuMo score (8^{th} column). Different elements are accessible, e.g. 2D representation of the co-crystallized ligand, MED-SuMo score, quantity of common SCFs, ligand name, structure header. The most important column contains the list of common SCF between the query and the hits: the SCF signature. Each SCF is represented by a colored rectangle e.g. light blue is for HBond donor, dark blue for positive charges. The most important characteristic of these SCFs is that they each stand for 3D functional similarities. (3) The mini-viewer window enables the user to set graphically the depth cueing and clipping. (4) The result table contains the common SCF signature between the query and the hits. MED-SuMo GUI enables the user to classify the hits according their SCF signatures. The 4^{th} window contains a dendrogram where all hits are present. The number of clusters can be selected by moving the yellow bar.

Table

a)

	ATP	ADP	AMP	ANP	GTP	GDP	GMP	GNP	NAD
ATP	80	42	21	28	3	3	0	2	8
ADP	42	104	14	47	1	6	1	3	8
AMP	21	14	56	8	1	1	5	0	6
ANP	28	47	8	61	0	2	0	3	0
GTP	3	1	1	0	14	4	0	0	0
GDP	3	6	1	2	4	21	3	1	0
GMP	0	1	5	0	0	3	13	0	0
GNP	2	3	0	3	0	1	0	6	0
NAD	8	8	6	0	0	0	0	0	53

b)

	AXP	GXP	NAD
AXP	303	31	22
GXP	31	76	0
NAD	22	0	53

Table 1 - *Confusion matrix of the ligand distribution within the clusters.*
(a) the confusion matrix is about each specific kind of the 9 ligands. (b) All adenine ligands are gathered in AXP, all guanine ligands in GXP whereas NAD remains NAD.

References

1. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
2. Sigrist, C.J., et al., *PROSITE: a documented database using patterns and profiles as motif descriptors.* Brief Bioinform, 2002. **3**(3): p. 265-74.
3. Hulo, N., et al., *The PROSITE database.* Nucleic Acids Res, 2006. **34**(Database issue): p. D227-30.
4. de Castro, E., et al., *ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W362-5.
5. Gattiker, A., E. Gasteiger, and A. Bairoch, *ScanProsite: a reference implementation of a PROSITE scanning tool.* Appl Bioinformatics, 2002. **1**(2): p. 107-8.
6. Sonnhammer, E.L., S.R. Eddy, and R. Durbin, *Pfam: a comprehensive database of protein domain families based on seed alignments.* Proteins, 1997. **28**(3): p. 405-20.
7. Finn, R.D., et al., *Pfam: clans, web tools and services.* Nucleic Acids Res, 2006. **34**(Database issue): p. D247-51.
8. Friedberg, I., T. Harder, and A. Godzik, *JAFA: a protein function annotation meta-server.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W379-81.
9. Todd, A.E., et al., *Progress of structural genomics initiatives: an analysis of solved target structures.* J Mol Biol, 2005. **348**(5): p. 1235-60.
10. Wendt, K.U., et al., *Structures and diseases.* Nat Struct Mol Biol, 2008. **15**(2): p. 117-20.
11. de Brevern, A.G., *New opportunities to fight against infectious diseases and to identify pertinent drug targets with novel methodologies.* Infect Disord Drug Targets, 2009. **9**(3): p. 246-7.
12. Guido, R.V., G. Oliva, and A.D. Andricopulo, *Virtual screening and its integration with modern drug design technologies.* Curr Med Chem, 2008. **15**(1): p. 37-46.
13. Rollinger, J.M., H. Stuppner, and T. Langer, *Virtual screening for the discovery of bioactive natural products.* Prog Drug Res, 2008. **65**: p. 211, 213-49.
14. Shaikh, S.A., et al., *From drug target to leads--sketching a physicochemical pathway for lead molecule design in silico.* Curr Pharm Des, 2007. **13**(34): p. 3454-70.
15. Waszkowycz, B., *Towards improving compound selection in structure-based virtual screening.* Drug Discov Today, 2008. **13**(5-6): p. 219-26.
16. Moriaud, F., et al., *Computational Fragment-Based Approach at PDB Scale by Protein Local Similarity.* J Chem Inf Model, 2009. **49**: p. 280-294.
17. Oguievetskaia, K., et al., *Computational fragment-based drug design to explore the hydrophobic sub-pocket of the mitotic kinesin Eg5 allosteric binding site.* J Comput Aided Mol Des, 2009.
18. Crespo, A. and A. Fernandez, *Induced Disorder in Protein-Ligand Complexes as a Drug-Design Strategy.* Mol Pharm, 2008.
19. Das, K., et al., *High-resolution structures of HIV-1 reverse transcriptase/TMC278 complexes: strategic flexibility explains potency against resistance mutations.* Proc Natl Acad Sci U S A, 2008. **105**(5): p. 1466-71.
20. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
21. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures.* J Mol Biol, 1977. **112**(3): p. 535-42.

22. Jefferson, E.R., T.P. Walsh, and G.J. Barton, *A comparison of SCOP and CATH with respect to domain-domain interactions.* Proteins, 2008. **70**(1): p. 54-62.
23. Hadley, C. and D.T. Jones, *A systematic comparison of protein structure classifications: SCOP, CATH and FSSP.* Structure, 1999. **7**(9): p. 1099-112.
24. Getz, G., et al., *Automated assignment of SCOP and CATH protein structure classifications from FSSP scores.* Proteins, 2002. **46**(4): p. 405-15.
25. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.
26. Andreeva, A., et al., *Data growth and its impact on the SCOP database: new developments.* Nucleic Acids Res, 2008. **36**(Database issue): p. D419-25.
27. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
28. Yoon, S., et al., *Clustering protein environments for function prediction: finding PROSITE motifs in 3D.* BMC Bioinformatics, 2007. **8 Suppl 4**: p. S10.
29. Ma, B.G., et al., *Characters of very ancient proteins.* Biochem Biophys Res Commun, 2008. **366**(3): p. 607-11.
30. Harrison, A., et al., *Quantifying the similarities within fold space.* J Mol Biol, 2002. **323**(5): p. 909-26.
31. Porter, C.T., G.J. Bartlett, and J.M. Thornton, *The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.* Nucleic Acids Res, 2004. **32**(Database issue): p. D129-33.
32. Bartlett, G.J., et al., *Analysis of catalytic residues in enzyme active sites.* J Mol Biol, 2002. **324**(1): p. 105-21.
33. Schmitt, S., D. Kuhn, and G. Klebe, *A new method to detect related function among proteins independent of sequence and fold homology.* J Mol Biol, 2002. **323**(2): p. 387-406.
34. Weber, A., et al., *Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition.* J Med Chem, 2004. **47**(3): p. 550-7.
35. Kuhn, D., et al., *Functional Classification of Protein Kinase Binding Sites Using Cavbase.* ChemMedChem, 2007. **2**(10): p. 1432-1447.
36. Kuhn, D., et al., *From the similarity analysis of protein cavities to the functional classification of protein families using cavbase.* J Mol Biol, 2006. **359**(4): p. 1023-44.
37. Shulman-Peleg, A., R. Nussinov, and H.J. Wolfson, *Recognition of functional sites in protein structures.* J Mol Biol, 2004. **339**(3): p. 607-33.
38. Mintz, S., et al., *Generation and analysis of a protein-protein interface data set with similar chemical and spatial patterns of interactions.* Proteins, 2005. **61**(1): p. 6-20.
39. Shulman-Peleg, A., R. Nussinov, and H.J. Wolfson, *SiteEngines: recognition and comparison of binding sites and protein-protein interfaces.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W337-41.
40. Baroni, M., et al., *A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application.* J Chem Inf Model, 2007. **47**(2): p. 279-94.
41. Powers, R., et al., *Comparison of protein active site structures for functional annotation of proteins and drug design.* Proteins, 2006. **65**(1): p. 124-35.
42. Kinoshita, K., Y. Murakami, and H. Nakamura, *eF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W398-402.
43. Standley, D.M., et al., *Protein structure databases with new web services for structural biology and biomedical research.* Brief Bioinform, 2008.

44. Nebel, J.C., P. Herzyk, and D.R. Gilbert, *Automatic generation of 3D motifs for classification of protein binding sites.* BMC Bioinformatics, 2007. **8**: p. 321.
45. Lichtarge, O., H.R. Bourne, and F.E. Cohen, *An evolutionary trace method defines binding surfaces common to protein families.* J Mol Biol, 1996. **257**(2): p. 342-58.
46. Mihalek, I., I. Res, and O. Lichtarge, *Evolutionary trace report_maker: a new type of service for comparative analysis of proteins.* Bioinformatics, 2006. **22**(13): p. 1656-7.
47. Morgan, D.H., et al., *ET viewer: an application for predicting and visualizing functional sites in protein structures.* Bioinformatics, 2006. **22**(16): p. 2049-50.
48. Madabushi, S., et al., *Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions.* J Biol Chem, 2004. **279**(9): p. 8126-32.
49. Kristensen, D.M., et al., *Prediction of enzyme function based on 3D templates of evolutionarily important amino acids.* BMC Bioinformatics, 2008. **9**: p. 17.
50. George, R.A., et al., *Effective function annotation through catalytic residue conservation.* Proc Natl Acad Sci U S A, 2005. **102**(35): p. 12299-304.
51. Laskowski, R.A., J.D. Watson, and J.M. Thornton, *From protein structure to biochemical function?* J Struct Funct Genomics, 2003. **4**(2-3): p. 167-77.
52. Laskowski, R.A., J.D. Watson, and J.M. Thornton, *ProFunc: a server for predicting protein function from 3D structure.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W89-93.
53. Morris, R.J., et al., *Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons.* Bioinformatics, 2005. **21**(10): p. 2347-55.
54. Watson, J.D., et al., *Towards fully automated structure-based function prediction in structural genomics: a case study.* J Mol Biol, 2007. **367**(5): p. 1511-22.
55. Brylinski, M., L. Konieczny, and I. Roterman, *Hydrophobic collapse in late-stage folding (in silico) of bovine pancreatic trypsin inhibitor.* Biochimie, 2006. **88**(9): p. 1229-39.
56. Konieczny, L., M. Brylinski, and I. Roterman, *Gauss-function-Based model of hydrophobicity density in proteins.* In Silico Biol, 2006. **6**(1-2): p. 15-22.
57. Brylinski, M., L. Konieczny, and I. Roterman, *Ligation site in proteins recognized in silico.* Bioinformation, 2006. **1**(4): p. 127-9.
58. Jurkowski, W., et al., *Conformational subspace in simulation of early-stage protein folding.* Proteins, 2004. **55**(1): p. 115-27.
59. Brylinski, M., et al., *Prediction of functional sites based on the fuzzy oil drop model.* PLoS Comput Biol, 2007. **3**(5): p. e94.
60. Jambon, M., et al., *A new bioinformatic approach to detect common 3D sites in protein structures.* Proteins, 2003. **52**(2): p. 137-45.
61. Jambon, M., et al., *The SuMo server: 3D search for protein functional sites.* Bioinformatics, 2005. **21**(20): p. 3929-30.
62. MEDIT-SA, *http://medit-pharma.com/*.
63. Doppelt-Azeroual, O., et al., *Analysis of HSP90 related folds with MED-SuMo classification approach.* Drug Design Development and Therapy, 2009. **3**: p. 59–72.
64. Doppelt, O., et al., *Functional annotation strategy for protein structures.* Bioinformation, 2007. **1**(9): p. 357-9.
65. van Dongen, S., *Graph Clustering by Flow Simulation*
2000, University of Utrecht.
66. Kabsch, W., et al., *Atomic structure of the actin:DNase I complex.* Nature, 1990. **347**(6288): p. 37-44.

67. de Brevern, A.G., C. Etchebest, and S. Hazout, *Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks.* Proteins, 2000. **41**(3): p. 271-87.
68. Hazout, S., *Entropy-derived measures for assessing the accuracy of N-state prediction algorithms.*, in *In Recent Advances in Structural Bioinformatics.* , A.G. de Brevern, Editor. 2007, Research signpost: Trivandrum, India. p. pp. 395-417.
69. Shannon, C., *A mathematical theory of communication.* Bell System Technical Journal, 1948. **27**: p. 379-423.
70. Gatzeva-Topalova, P.Z., May, A.P., Sousa, M.C. , *Structure and Mechanism of ArnA: Conformational Change Implies Ordered Dehydrogenase Mechanism in Key Enzyme for Polymyxin Resistance* Structure, 2005. **13**: p. 929-942.
71. Allard, S.T., et al., *Toward a structural understanding of the dehydratase mechanism.* Structure, 2002. **10**(1): p. 81-92.
72. Webb, N.A., Mulichak, A.M., Lam, J.S., Rocchetta, H.L., Garavito, R.M. , *Crystal structure of a tetrameric GDP-D-mannose 4,6-dehydratase from a bacterial GDP-D-rhamnose biosynthetic pathway.* Protein Sci, 2004. **13**(529-539).
73. Bairoch, A., Boeckmann B., Ferro S., Gasteiger E., *Swiss-Prot: Juggling between evolution and stability* Brief. Bioinform, 2004. **5**: p. 39-55.
74. Hung, L.W., Wang, I.X., Nikaido, K., Liu, P.Q., Ames, G.F., Kim, S.H., *Crystal structure of the ATP-binding subunit of an ABC transporter.* Nature, 1998. **396**: p. 703-707.
75. Chen, J., Lu, G., Lin, J., Davidson, A.L., Quiocho, F.A., *A tweezers-like motion of the ATP-binding cassette dimer in an ABC transport cycle.* Mol. Cell, 2003. **12**: p. 651-661
76. Lewis, H.A., et al., *Structure of nucleotide-binding domain 1 of the cystic fibrosis transmembrane conductance regulator.* EMBO J, 2004. **23**(2): p. 282-93.
77. Ramaen, O., Leulliot, N., Sizun, C., Ulryck, N., Pamlard, O., Lallemand, J.Y., Tilbeurgh, H., Jacquet, E., *Structure of the human multidrug resistance protein 1 nucleotide binding domain 1 bound to Mg2+/ATP reveals a non-productive catalytic site.* J.Mol.Biol. , 2006. **359**: p. 940-949
78. Zaitseva, J., et al., *A structural analysis of asymmetry required for catalytic activity of an ABC-ATPase domain dimer.* EMBO J, 2006. **25**(14): p. 3432-43.
79. Gaudet, R., Wiley, D.C., *Structure of the ABC ATPase domain of human TAP1, the transporter associated with antigen processing.* EMBO J., 2001. **20**: p. 4964-4972.
80. Beis, I. and E.A. Newsholme, *The contents of adenine nucleotides, phosphagens and some glycolytic intermediates in resting muscles from vertebrates and invertebrates.* Biochem J, 1975. **152**(1): p. 23-32.
81. Iyidogan, P. and S. Lutz, *Systematic exploration of active site mutations on human deoxycytidine kinase substrate specificity.* Biochemistry, 2008. **47**(16): p. 4711-20.
82. Reyes, C.L., et al., *X-ray structures of the signal recognition particle receptor reveal targeting cycle intermediates.* PLoS ONE, 2007. **2**(7): p. e607.
83. Kornberg, T. and M.L. Gefter, *Deoxyribonucleic acid synthesis in cell-free extracts. IV. Purification and catalytic properties of deoxyribonucleic acid polymerase III.* J Biol Chem, 1972. **247**(17): p. 5369-75.
84. Oganesyan, V., et al., *Structure of O67745_AQUAE, a hypothetical protein from Aquifex aeolicus.* Acta Crystallogr Sect F Struct Biol Cryst Commun, 2007. **63**(Pt 5): p. 369-74.
85. Kim, D.Y. and K.K. Kim, *Crystal structure of ClpX molecular chaperone from Helicobacter pylori.* J Biol Chem, 2003. **278**(50): p. 50664-70.
86. Bowman, G.D., M. O'Donnell, and J. Kuriyan, *Structural analysis of a eukaryotic sliding DNA clamp-clamp loader complex.* Nature, 2004. **429**(6993): p. 724-30.

87. Kazmirski, S.L., et al., *Structural analysis of the inactive state of the Escherichia coli DNA polymerase clamp-loader complex.* Proc Natl Acad Sci U S A, 2004. **101**(48): p. 16750-5.
88. Xu, W., S.C. Harrison, and M.J. Eck, *Three-dimensional structure of the tyrosine kinase c-Src.* Nature, 1997. **385**(6617): p. 595-602.
89. Goldsmith-Fischman, S. and B. Honig, *Structural genomics: computational methods for structure analysis.* Protein Sci, 2003. **12**(9): p. 1813-21.
90. Dessailly, B.H., M.F. Lensink, and S.J. Wodak, *Relating destabilizing regions to known functional sites in proteins.* BMC Bioinformatics, 2007. **8**: p. 141.
91. Brown, D.P., N. Krishnamurthy, and K. Sjolander, *Automated protein subfamily identification and classification.* PLoS Comput Biol, 2007. **3**(8): p. e160.
92. Mao, L., et al., *Molecular determinants for ATP-binding in proteins: a data mining and quantum chemical analysis.* J Mol Biol, 2004. **336**(3): p. 787-807.
93. Niefind, K., et al., *GTP plus water mimic ATP in the active site of protein kinase CK2.* Nat Struct Biol, 1999. **6**(12): p. 1100-3.
94. Yde, C.W., et al., *Inclining the purine base binding plane in protein kinase CK2 by exchanging the flanking side-chains generates a preference for ATP as a cosubstrate.* J Mol Biol, 2005. **347**(2): p. 399-414.
95. Sotriffer, C. and G. Klebe, *Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design.* Farmaco, 2002. **57**(3): p. 243-51.
96. Wei, L. and R.B. Altman, *Recognizing protein binding sites using statistical descriptions of their 3D environments.* Pac Symp Biocomput, 1998: p. 497-508.
97. Kasuya, A. and J.M. Thornton, *Three-dimensional structure analysis of PROSITE patterns.* J Mol Biol, 1999. **286**(5): p. 1673-91.
98. Wu, S., M.P. Liang, and R.B. Altman, *The SeqFEATURE library of 3D functional site models: comparison to existing methods and applications to protein function annotation.* Genome Biol, 2008. **9**(1): p. R8.
99. Halperin, I., et al., *The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications.* BMC Genomics, 2008. **9 Suppl 2**: p. S2.
100. Manly, C.J., et al., *Strategies and tactics for optimizing the Hit-to-Lead process and beyond--a computational chemistry perspective.* Drug Discov Today, 2008. **13**(3-4): p. 99-109.
101. Vieth, M., et al., *Kinomics-structural biology and chemogenomics of kinase inhibitors and targets.* Biochim Biophys Acta, 2004. **1697**(1-2): p. 243-57.
102. Krupa, A., K.R. Abhinandan, and N. Srinivasan, *KinG: a database of protein kinases in genomes.* Nucleic Acids Res, 2004. **32**(Database issue): p. D153-5.
103. Cheek, S., et al., *A comprehensive update of the sequence and structure classification of kinases.* BMC Struct Biol, 2005. **5**: p. 6.
104. Manning, G., et al., *The protein kinase complement of the human genome.* Science, 2002. **298**(5600): p. 1912-34.
105. Dar, A.C., T.E. Dever, and F. Sicheri, *Higher-order substrate recognition of eIF2alpha by the RNA-dependent protein kinase PKR.* Cell, 2005. **122**(6): p. 887-900.
106. Brown, N.R., et al., *Effects of phosphorylation of threonine 160 on cyclin-dependent kinase 2 structure and activity.* J Biol Chem, 1999. **274**(13): p. 8746-56.
107. Aoki, M., et al., *Structural insight into nucleotide recognition in tau-protein kinase I/glycogen synthase kinase 3 beta.* Acta Crystallogr D Biol Crystallogr, 2004. **60**(Pt 3): p. 439-46.

108. Pevarello, P., et al., *3-Amino-1,4,5,6-tetrahydropyrrolo[3,4-c]pyrazoles: a new class of CDK2 inhibitors.* Bioorg Med Chem Lett, 2006. **16**(4): p. 1084-90.
109. Dutta, R. and M. Inouye, *GHKL, an emergent ATPase/kinase superfamily.* Trends Biochem Sci, 2000. **25**(1): p. 24-8.
110. Ferre, F., et al., *Functional annotation by identification of local surface similarities: a novel tool for structural genomics.* BMC Bioinformatics, 2005. **6**: p. 194.
111. Gasteiger, E., et al., *ExPASy: The proteomics server for in-depth protein knowledge and analysis.* Nucleic Acids Res, 2003. **31**(13): p. 3784-8.
112. Enright, A.J. and C.A. Ouzounis, *BioLayout--an automatic graph layout algorithm for similarity visualization.* Bioinformatics, 2001. **17**(9): p. 853-4.
113. Goldovsky, L., et al., *BioLayout(Java): versatile network visualisation of structural and functional relationships.* Appl Bioinformatics, 2005. **4**(1): p. 71-4.
114. Cheek, S., H. Zhang, and N.V. Grishin, *Sequence and structure classification of kinases.* J Mol Biol, 2002. **320**(4): p. 855-81.
115. Panek, J., I. Eidhammer, and R. Aasland, *A new method for identification of protein (sub)families in a set of proteins based on hydropathy distribution in proteins.* Proteins, 2005. **58**(4): p. 923-34.
116. Henrick, K., et al., *Remediation of the protein data bank archive.* Nucleic Acids Res, 2008. **36**(Database issue): p. D426-33.
117. OCaml, http://caml.inria.fr/.
118. Higgins D., T.J., Gibson T., Thompson J.D., Higgins D.G., Gibson T.J., *CLUSTAL W: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice.* Nucleic Acids Research 1994. **22**: p. 4673-4680.
119. Lua-ML, http://caml.inria.fr/cgi-bin/hump.en.cgi?contrib=321.