Analysis of HSP90-related folds with MED-SuMo classification approach

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

[1]MEDIT SA, Palaiseau, France;
[2]INSERM UMR-S 726, Equipe de Bioinformatique Génomique and Moléculaire, (EBGM), DSIMB, Institut National de Transfusion Sanguine (INTS), Université Paris Diderot, Paris, France

Abstract: Three-dimensional structural information is critical for understanding functional protein properties and the precise mechanisms of protein functions implicated in physiological and pathological processes. Comparison and detection of protein binding sites are key steps for annotating structures with functional predictions and are extremely valuable steps in a drug design process. In this research area, MED-SuMo is a powerful technology to detect and characterize similar local regions on protein surfaces. Each amino acid residue's potential chemical interactions are represented by specific surface chemical features (SCFs). The MED-SuMo heuristic is based on the representation of binding sites by a graph structure suitable for exploration by an efficient comparison algorithm. We use this approach to analyze one particular SCOP superfamily which includes HSP90 chaperone, MutL/DNA topoisomerase, histidine kinases, and α-ketoacid dehydrogenase kinase C (BCK). They share a common fold and a common region for ATP-binding. To analyze both similar and differing features of this fold, we use a novel classification method, the MED-SuMo multi approach (MED-SMA). We highlight common and distinct features of these proteins. The different clusters created by MED-SMA yield interesting observations. For instance, one cluster gathers three types of proteins (HSP90, topoisomerase VI, and BCK) which all bind the drug radicicol.

Keywords: functional classification, surface similarity, protein surface chemical feature, radicicol binding

Introduction

Protein three-dimensional (3D) structural information help to understand functional protein properties and the precise mechanisms of proteins implicated in physiological and pathological processes.[1] Knowledge of 3D protein structures linked to small molecules can be used for structure- and ligand-based drug design approaches.[2,3] It also gives direct hints to the protein functional mechanisms. A protein's activity often depends on a small, highly conserved set of residues within the binding site.[4,5] Comparison and detection of protein binding sites are key steps for annotating structures with functional predictions. In this field, Structural Genomics consortia have radically changed mankind's base of protein structural knowledge. Their endeavors have permitted the resolution of numerous structures characterized as "Unknown function", and multiple functional sites are not associated with any known binding partner.[6] Consequently, the development of computational methods to functionally annotate protein structures has become a major research area.

The simplest approaches are based on sequence *analogy*, eg, PSI-BLAST,[7] or on the characterization of functional patterns or profiles, eg, PROSITE.[8] They help to draw on knowledge and assumptions of protein functions in assigning predicted functions. However, they cannot embrace the complexity of local 3D folds. During the past years, various methods to compare and detect binding sites have been elaborated; they use diverse types of descriptors. Their general purpose is often to create automated functional annotation methods independent from amino acid sequence or from

Correspondence: Olivia Doppelt-Azeroual
MEDIT SA, 2 rue du Belvédère, 91120,
Palaiseau, France
Tel +33160148743
Fax +33160148473
Email olivia.doppelt@univ-paris-diderot.fr

global fold similarity, eg, CavBase,[9] SiteEngine,[10] FLAP,[11] CPASS,[12] or eF-seek.[13]

Some of these approaches share gross features but they also have notable distinctions. For instance, SiteEngine and CavBase both associate physico-chemical properties to structural characteristics. However, SiteEngine allows the comparison of entire protein surfaces to a binding site database, whereas CavBase is restricted to cavity comparisons. The web-based version of SiteEngine is restricted to the comparison of a single site versus one protein structure.[10] CavBase detects related cavities based on a clique detection algorithm[9] while CPASS comparison uses an alignment of binding site pairs through a root–mean–square–difference (RMSD) scoring function.[12] Roterman has developed an innovative methodology based on irregular hydrophobicity distribution.[14] A few other methods are based on the detection of conserved residues to characterize binding sites, eg, evolutionary trace method[15–17] or sequence alignment with a dedicated dataset as Catalytic Site Atlas (CSA).[4]

In this research area, SuMo is a powerful technology to localize similar local regions on protein surfaces ie, binding sites.[18] Each chemical property, or interaction, of an amino acid residue is represented by a specific surface chemical feature (SCF). These are gathered in triangles to constitute a SuMo graph vertex. Since each SCF is associated with heterogeneous geometrical properties, and that triplets have specific superimposition rules (distance, angle), the comparison heuristic is extremely rapid. The comparison of a 3D pattern against all the binding sites of the PDB can be performed in a few minutes.[19] MED-SuMo is the latest evolution of SuMo software developed by MEDIT-SA (see http://www.medit.pharma.com/). Recent developments have improved its binding site database, and have included novel functional annotation tools as presented in a recent study.[20]

Proteins are also classified according to their folds,[21] eg, SCOP (Structural Classification of Proteins),[22,23] that provides a manually refined classification with detailed and comprehensive descriptions of the structural and evolutionary relationships of the known protein structure.[22,23] However, a critical limitation of these fold-based classifications is the use of complete protein folds or protein domains. Similarity of fold does not necessarily correspond to a similarity of function. In this paper, we focus on an interesting SCOP superfamily which includes the heat shock protein 90 SCOP family (HSP90, see Figure 1).

HSP90 is one of the most abundant proteins. Its different forms exhibit mainly chaperone functions associated to protein folding, cell survival,[24] apoptosis and tumor repression.[25] It binds ATP (see Figures 2a and 2b) and is the target of some innovative drugs including geldanamycin which has enabled 50% reduction of tumor growth,[26] and celasterol which disrupts interactions between HSP90 and Cdc37 in pancreatic cancer cells.[27] Some recent research focussed on a new potential drug, radicicol. This molecule has a very high affinity for HSP90 (20 nM).[28] Figure 3 shows the association of the drug with the HSP90 at the binding site normally filled with a natural ligand.[28] However, radicicol is not specific to HSP90 as it binds bacterial Sensor Kinase PhoQ,[29] and topoisomerase VI.[30] An interesting detail is that HSP90 chaperone, MutL/DNA topoisomerase or histidine kinases share (see Figure 1) a common fold and that a common region of ATP-binding has been detected (see Figures 2c and 2d).

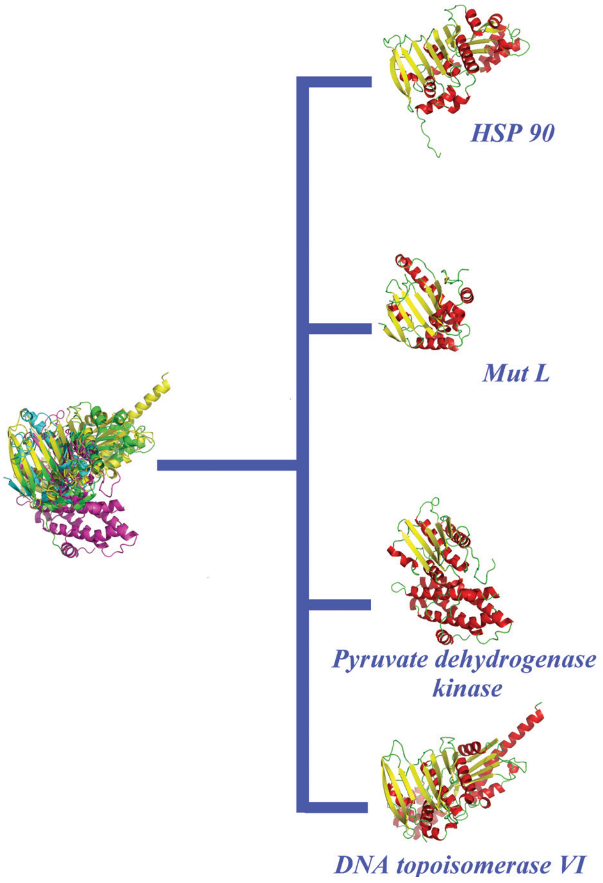


Figure 1 Heat shock protein 90 (HSP90) SCOP superfamily: GHKL: HSP90, MutL proteins, pyruvate dehydrogenase kinase and DNA topoisomerase VI all share this fold.

To analyze the similar and different features of this fold, we use a novel classification method, MED-SuMo Multi approach (MED-SMA), based on the MED-SuMo technology. In this work, binding sites from the SCOP superfamily ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase proteins are gathered in a dataset, compared

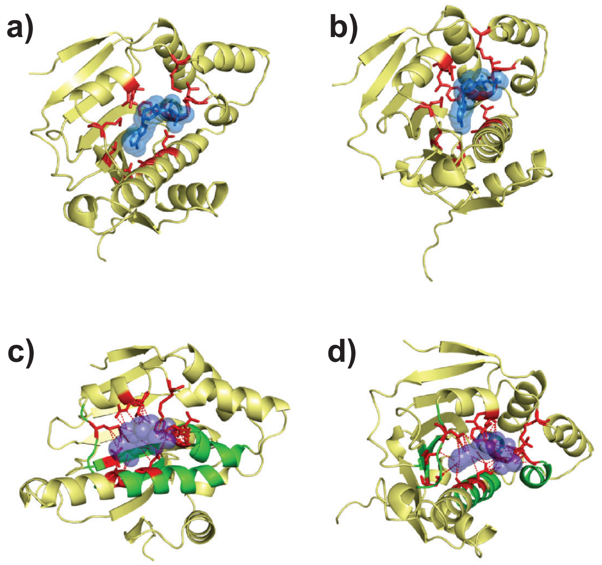


Figure 2 An example of heat shock protein 90 (HSP90) bound to its natural ligand. The protein shown is an HSP90 of *Saccharomyces cerevisiae* (PDB code 1AMW). **a**–**b**) underlines the close contacts (in red) of the ADP (in blue). **c**–**d**) underlines in green the common binding region of this SCOP superfamily.

pairwise and classified using the Markov Cluster Algorithm (MCL).[31] Results from this method highlight common and distinct functional features between the analyzed proteins.

Materials and methods

Protein structure database

SCOP web site provides the list of proteins associated to a selected fold.[23] The *"ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase"* superfamily contains 116 PDB structures (see http://scop.berkeley.edu/data/scop.b.e.ccg.A.html). The protein binding sites were selected to perform the classification.

MED-SuMo algorithm

MED-SuMo is designed to localize similar regions associated to a defined function.[18–20] A key advantage is its ability to detect binding site similarities even when local flexibility is observed. Its heuristic is based on a 3D representation of macromolecules using precise SCFs. For MED-SuMo, a protein structure is represented by a set of functional groups including, for example, unbound hydrogen bond (Hbond) donors or acceptors, accessible sides of aromatic rings and carboxylate, charges, hydroxyl groups. Each feature encodes its chemical characteristics with precise geometrical properties. The overall MED-SuMo comparison methodology is presented in Figure 4. SCFs are displayed on the protein structure through a lexicographic analysis of the atoms in the PDB files, ie, a residue is represented by a set of representative SCFs (cf. Figures 4a, 4b). Their positions and orientations are filtered as shown in Figure 4c. Remaining SCFs are assembled into triplets with specific geometric characteristics, eg, edge size, perimeter, angles (cf. Figure 4d). The full triplet network is stored in the MED-SuMo database as a graph data structure where triplets are the vertices and edges connect adjacent triangles (ie, those sharing at least two SCFs).

To compare graphs, MED-SuMo looks for compatible triplets; composed of compatible SCFs (cf. Figure 4e). These triplets are called comparison "seeds". When a seed is detected, MED-SuMo extends the comparisons to the vertices of the neighbourhood, until no more similarities are found. This process enables the formation of similar patches (common groups of SCFs) between two graphs, weighted up by the MED-SuMo score.[18] These comparisons are usually performed between a query and a database of precompiled graphs. Two kinds of MED-SuMo database are commonly used: the binding site database that is composed from the SCFs around co-crystallized ligands and the full surface database, composed from SCFs covering the whole surface of each studied protein, typically the entire PDB. The database characteristics are defined by three essential parameters: the size of the ligand environment taken into account

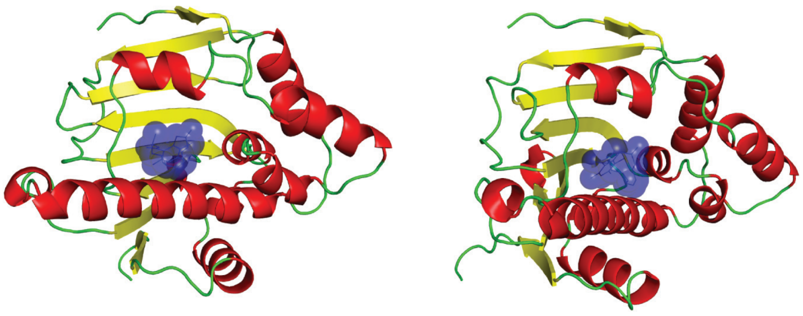


Figure 3 An example of heat shock protein 90 (HSP90) bound to radicicol. Both views represent an HSP90 of *Saccharomyces cerevisiae* (PDB code 1BGQ) bound to the drug radicicol shown in blue (see Figure 2 to compare with the natural ligand of HSP90).

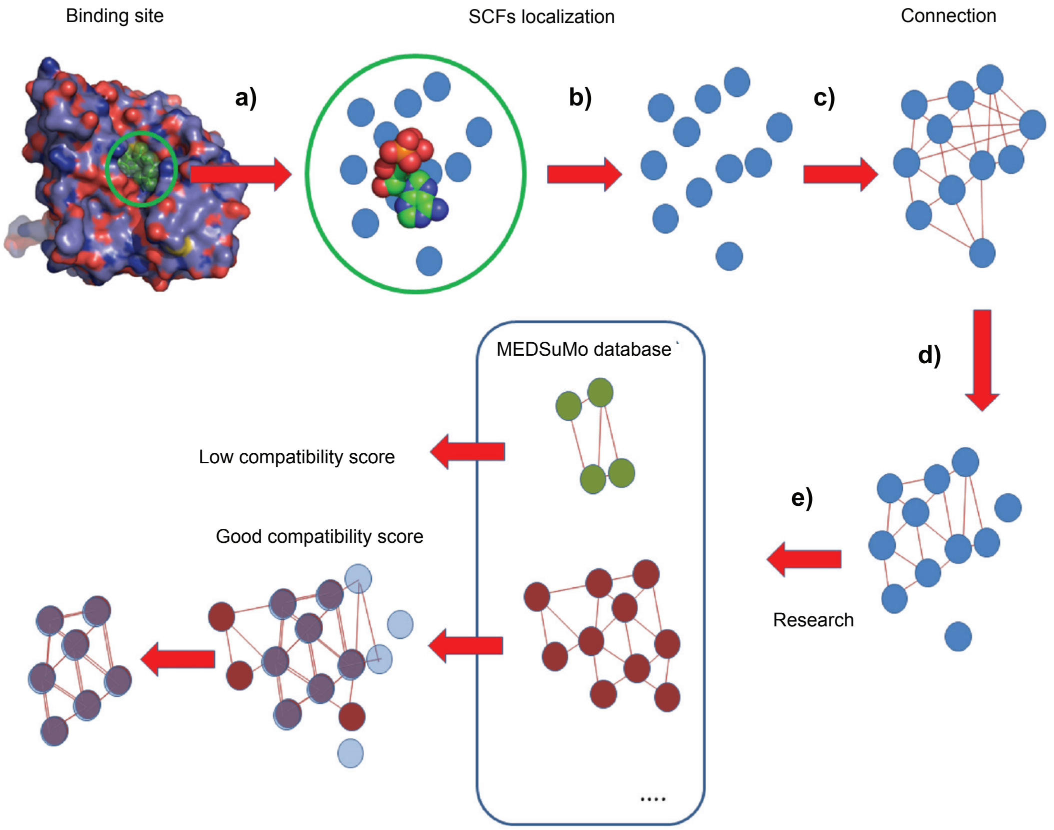


Figure 4 MED-SuMo comparison procedure. **a**) Localization of an interesting part of the protein surface often characterized by the presence of a co-crystallized ligand. **b**) Surface chemical features (SCFs) are displayed on the protein structure through a lexicographic analysis of the PDB files. **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. **e**) The query graph (in blue) is compared to the database graphs (in green and brown); they usually represent all binding sites of the PDB. Compatible triplets are detected, ie, they are formed by compatible SCFs. At last, the corresponding graphs (hits) are ranked in regard to their compatibility score.

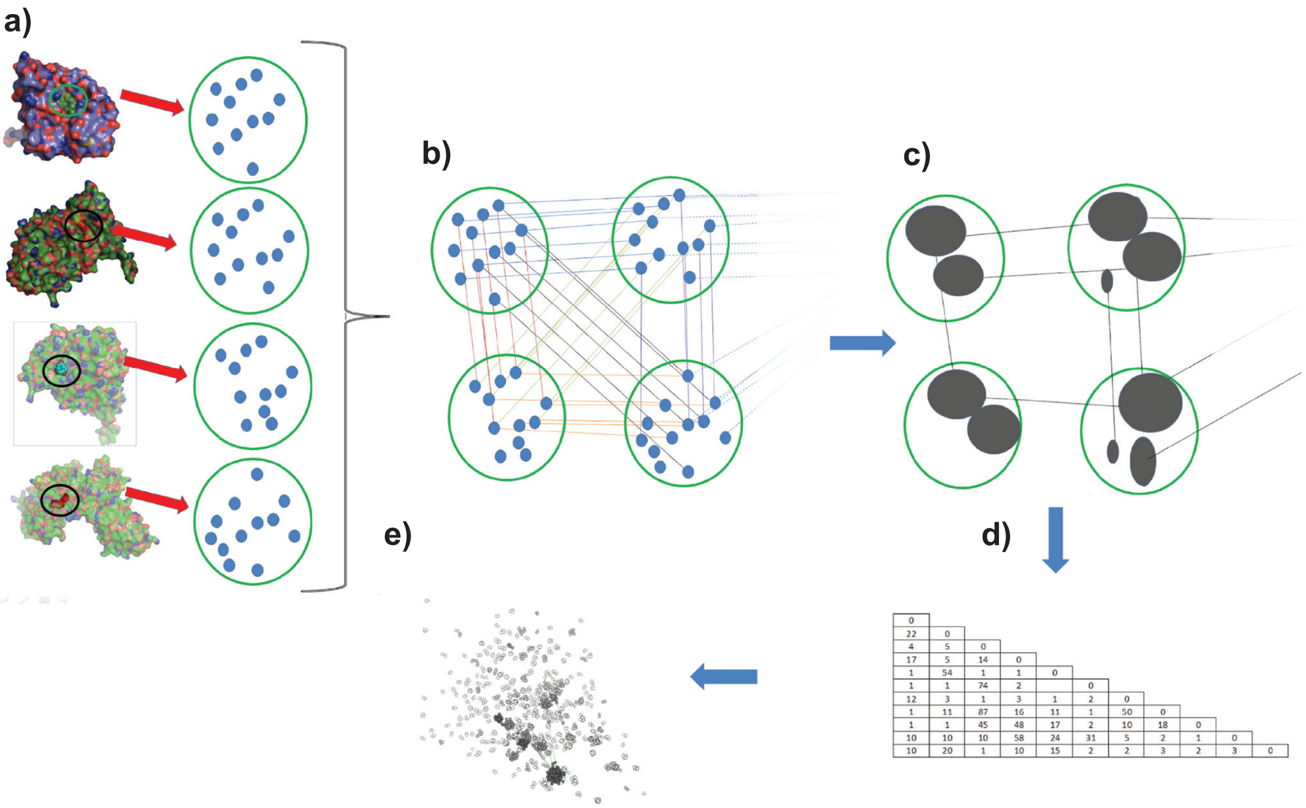
by MED-SuMo (named *ligand_radius* and only concerning the binding site database), the maximal distance between two SCFs to be included in a triplet (named *edge_max*) and the maximal perimeter for a triangle (named *max_edge_sum*).

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.[18–20] We decided to apply this approach to classify defined sets of structures. This new method, named MED-SuMo Multi Approach (MED-SMA), enables the comparison of all binding sites from a set of proteins using a pairwise comparison system. Matching regions are found in the binding sites to derive a similarity graph. This graph is classified with the MCL[31]. Figure 5 illustrates the overall procedure. For this work, MED-SMA is only applied on the MED-SuMo binding sites database.

To begin, a set of proteins is selected (see previous paragraph, cf. Figure 5a). Ligands' characteristics are used to decide which binding sites to include in the MED-SuMo database. Once the ligands parameters are set, the database is created and the pairwise comparison is launched using the standard MED-SuMo comparison procedure.

These comparisons highlight similar regions between pairs of binding sites (cf. Figure 5b) represented by groups of SCFs called patches. Only comparisons with a MED-SuMo score higher than a fixed cut-off (parameter *score_min*) are accepted. Patches associated to the same binding sites are analyzed: if two patches share enough SCFs (defined by a threshold parameter named *covering_factor*), they are merged in a multipatch (cf. Figure 5c). A multipatch is a set of SCFs common to several binding sites of the protein set; they can also be called sub-sites. They represent the true meaningful common regions of binding sites. They have two properties: (i) enough SCFs are in common, such that binding sites are structurally and chemically similar, and (ii) they can provide a measure of sub-pocket similarity. These measures are used to compute a similarity matrix. For this matrix, the MED-SuMo score between matching multipatches is calculated (cf. Figure 5d). MCL is used to interpret the matrix through



0										
22	0									
4	5	0								
17	5	14	0							
1	54	1	1	0						
1	1	74	2		0					
12	3	1	3	1	2	0				
1	11	87	16	11	1	50	0			
1	1	45	48	17	2	10	18	0		
10	10	10	58	24	31	5	2	1	0	
10	20	1	10	15	2	2	3	2	3	0

Figure 5 Global steps of binding site classification heuristic. MED-SuMo Multi approach (MED-SMA) can be divided in 5 steps: **a**) Database construction: all selected binding sites are stored as graph in the MED-SuMo database. **b**) Pairwise comparisons: all binding sites are compared to each other to detect similarities between pairs (lines with different colors). These similarities are called patches **c**) If overlapping patches have a certain amount of common SCFs (more than a threshold value: parameter covering_factor), they are merged in multipatches (grey circles). **d**) MED-SuMo scores between pairs of multipatches are calculated and used to create a similarity matrix which is classified by MCL (Markov Cluster Algorithm) to create clusters of binding sites. **e**) Biolayout 2D view of the MED-SMA clusters.

classification of the protein binding site set into clusters of sub-sites (cf. Figure 5e). A 2D plot of the clusters can be visualized using tools such as Biolayout.[32,33]

Results

MED-SMA classification

To generate the MED-SuMo database, only binding sites co-crystallized with ligands with more than ten atoms are selected. Of the originally selected 116 PDB structures, 101 satisfy this filter. This yields a total of 146 binding sites in the final database. Several kinds of ligands are present, purines, eg, adenosine tri-phosphate or N-ethyl-5'-carboxamido adenosine, or potential drugs, eg, Radicicol or Novobiocin. Of these 146 binding sites, 78 are from HSP90, 38 from topoisomerase/MutL, 26 are from histidine kinase, and four are from α-keto-acid dehydrogenase kinase C (BCK). The database parameters are set to a ligand radius of 6.0 Å and triangle parameters of 13 Å and 39 Å (respectively *edge_max* and *max_edge_sum*). To classify this dataset, MED-SMA takes around two minutes on a four CPU machine. The classification parameters are set to a minimal compatibility score (*score_min*) of 4.0 and a *covering_factor* of 0.6.

Here, the MED-SMA approach produces five clusters. The distribution of these clusters in regards to the SCOP families is shown in Table 1 and the composition of each cluster is available in Supplementary data 1.

Two types of MED-SMA clusters are seen. Three clusters are homogeneous as they contain only proteins from a unique SCOP family (MED-SMA clusters 1, 3, and 5). Two clusters are heterogeneous as they contain at least two SCOP families (MED-SMA clusters 2 and 4). MED-SMA clusters 1 and 3 are specific to topoisomerase/MutL while cluster 5 is specific to histidine kinase. MED-SMA cluster 2 contains binding sites from two families (ie, BCK and histidine kinase) and MED-SMA cluster 4's binding sites are from three of the four families (HSP90, topoisomerase/MutL, and BCK).

MED-SMA clusters 1 and 3

MED-SMA clusters 1 and 3 contain 22 and 6 binding sites of the 38 proteins of the topoisomerase/MutL/DNA gyrase

Table 1 Confusion matrix of the SCOP families within the clusters. The MED-SuMo clusters are arranged vertically whereas the SCOP families are arranged horizontally. MED-SuMo clusters #1, #3 and #5 are homogeneous clusters, they only contain protein from: SCOP DNA gyrase/MutL family (for #1 and #3) and histidine kinase, respectively. MED-SuMo clusters #2 and #4 are heterogeneous

SCOP fam MED-Clusters	HSP90	DNA gyrase MutL	Histidine kinase	A-ketoacid dehydrogenase kinase C
1	0	22	0	0
2	0	0	15	3
3	0	6	0	0
4	78	10	0	1
5	0	0	11	0

family, respectively. The two forms of topoisomerases IV structures of *Escherichia coli* (PDB code 1S14 and 1S16) share 99.5% sequence identity except for a 23 residue insertion in 1S16. These two proteins are separated by MED-SMA. A precise look at their ATP-binding sites highlights structural similarities but, above all, some strong distinctions. Figure 6 shows a 3D superimposition of these proteins. The region noted (1) on Figure 6 shows an excellent superimposition of several β-sheets and 2 α-helixes. Moreover a part of the binding sites is also similar, with a set of five SCFs well superimposed (noted [2] on Figure 6). Conversely, the other side of the binding site (noted [3] on Figure 6) is quite diverse. Ligands of these two topoisomerases are novobiocin for 1S14 and phosphoaminophosphonic acid-adenylate ester (ANP) for 1S16. They are not located at the same spatial position and their overlap is small (~10 atoms) compared to their respective sizes (44 atoms for novobiocin and 31 atoms for ANP). Furthermore, novobiocin can not fit at all in the 1S16 binding site, otherwise a steric clash appears with 1S16's α helixes (noted [4] on Figure 6). Thus, binding sites from MED-SMA clusters 1 and 3 do not share sufficient similarities to be gathered by MED-SMA, neither can they bind the same kind of molecules. Interestingly, the two forms are very close but the residue insertion causes strongly diverging affinities to ligands of this class.[34] So, our results reinforce the study of Bellon and colleauges. Moreover, it characterizes with elegance the fact that these two distinct local conformations are found in different related proteins.

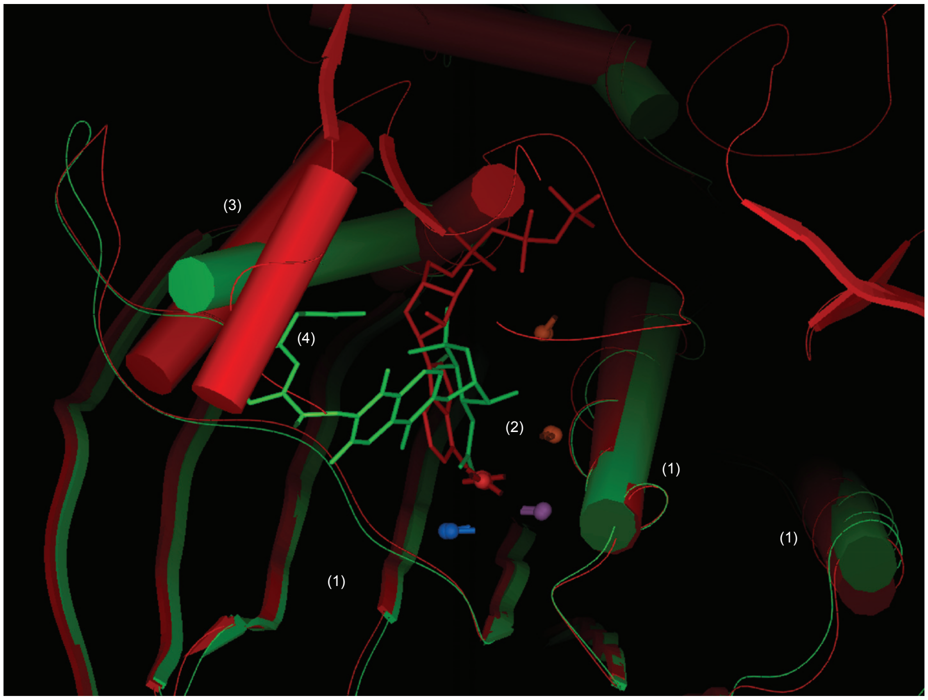


Figure 6 Superimposition of two topoisomerase VI separated by MED-SMA. PDB codes 1S16 (red) and 1S14 (green) are superimposed. They are both topoisomerase but their binding sites do not share enough similarity to be grouped in the same cluster. This figure is divided by several numbered regions: **1**) Protein structure similarities. two α helixes and several β-sheets are common to both structures. **2**) Low similarity in binding sites underlined by five SCFs. **3**) Difference between the two structures on the other side of the binding site. **4**) Potential clash between the query ligand and the hit protein structure.

MED-SMA cluster 4

As mentioned earlier, MED-SMA cluster 4 gathers three different SCOP families. It is the largest cluster, containing 89 binding sites. All HSP90s of the dataset are present (78 binding sites), 10 from mutL/DNA topoisomerase family (with one topoisomerase VI, five MutL, and four PMS2) and one from BCK family. Only the histidine kinase family is not represented in this MED-SMA cluster. The ligands are highly diverse with 48 unique ligands found.

Binding sites in this MED-SMA cluster share a common set of SCFs. Figure 7 shows a global superimposition of one structure of each family. The white rectangles show similarities whereas the remainder is very different as represented in the global superimposition of all the protein families in Figure 1. Figure 8 shows a close view around the radicicol. The eight labelled SCFs (circled in yellow) are shared by all superimposed structures in Figure 7. They are located all around the ligand meaning that the similarities concern the whole binding site.

The fact that MED-SMA gathers the binding sites from three different SCOP families implies a high probability that the binding modes are related. Considering the nonspecific drug radicicol which binds HSP90 and topoisomerase VI,[30] we could easily make the hypothesis that this drug would also bind the different proteins included in this MED-SMA cluster.

MED-SMA clusters 2 and 5

MED-SMA clusters 2 and 5 mostly consist of histidine kinase. MED-SMA cluster 2 is heterogeneous while MED-SMA cluster 5 is homogeneous. Cluster 5 is very worthwhile because it is pure and that the dimensions of its binding sites are very similar as they all bind purine ligands. Since the binding sites gathered by MED-SMA share binding modes to ligands, this type of cluster could be used to search for specific drugs; here, drugs to inhibit histidine kinase CheA action.

Interestingly, MED-SMA cluster 2 also contains two histidine kinase CheA (PDB codes 2CH4 and 1I5D). The separation of proteins from the same family in two different clusters is due to differences between their binding sites. When 1I5D's binding site is compared to histidine kinase CheA from cluster 5, the MED-SuMo score is less than 4.0 (which is the cut-off we chose for the pairwise comparison step). So, a drug designed to inhibit binding sites of cluster 5 would not bind (or not with the same affinity) the two excluded histidine kinase CheA binding sites.

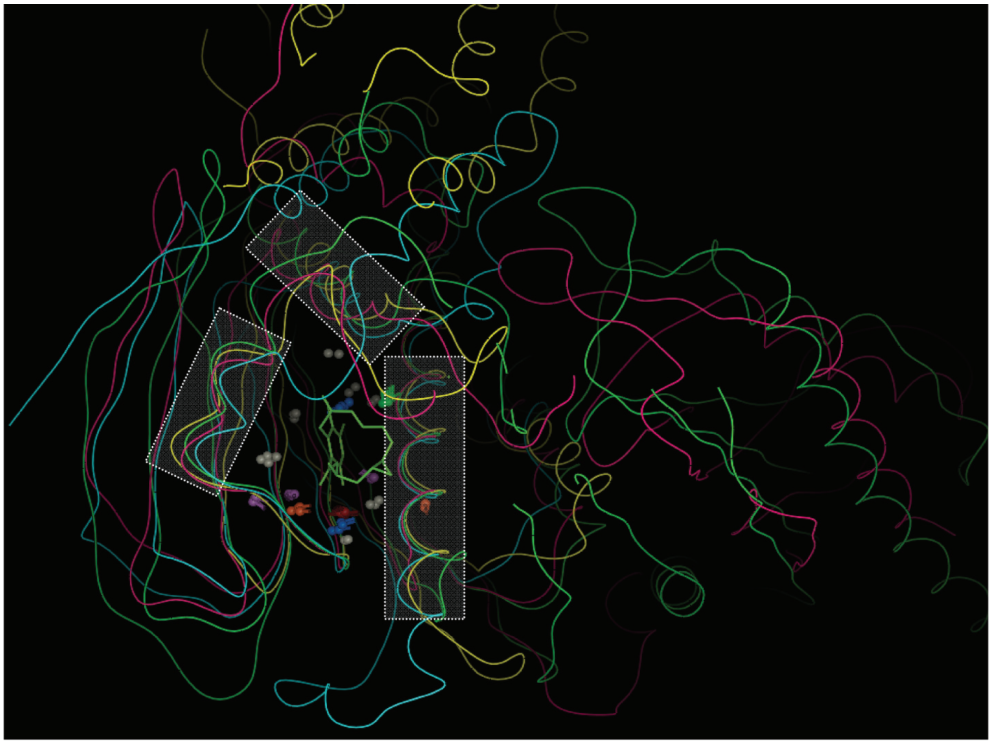


Figure 7 Superimposition of four proteins from three distinct SCOP families but gathered in the same cluster by MED-SMA. (PDB codes 2HKJ [green], 2CCT [cyan], 1B63 [pink] 1JM6 [yellow]). The white rectangles show similarities around the ligands and also the helices from the Bergerat fold. The rest of the superimposition is quite messy, as protein global folds are very different.

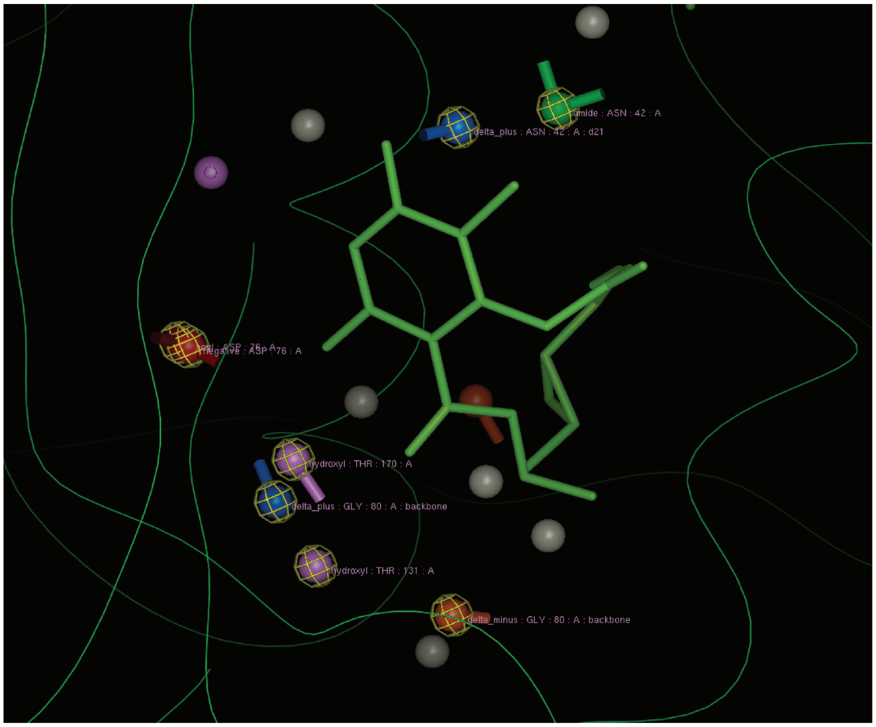


Figure 8 A close view around the radicicol ligand. The eight labelled SCFs (circled in yellow) are shared by all superimposed structures in Figure 7. They are located all around the ligand, which means that the similarities concern the whole binding site.

Another interesting point on MED-SMA cluster 2 is that it contains both BCK and anti-sigma factor spoIIab. These two proteins are inhibited like HSP90 by the radicicol. However, as they are not associated to MED-SMA cluster 4, it may reflect a specific binding mode.

Discussion

The detection of functional sites on protein surfaces is important for the identification of biological activity. Ligand-protein interactions occur for the majority of protein structures and they are implicated in major biological processes. However, with no help from known related sequences or structures their detection is difficult.[14] Several innovative approaches have been proposed, ie, the use of hydrophobicity distribution on protein structures based on the fuzzy oil drop model,[35] the destabilization of limited protein regions,[36] phylogenomic classification of protein sequences[37] or the classification of known protein catalytic sites.[38] Prediction of protein functional sites is an important step to identify small-molecule interactions for drug discovery[39] and it can be very useful to optimize drug design.[40] Another valuable application is as a pre-processing step to reduce the search space for rigorous computational docking algorithms.

Methods to compare binding sites have been developed using various kinds of structural descriptors, eg, CavBase uses pseudocenters,[41] and the strong hypothesis that chemical similarity and activity are linked. In this field, MED-SuMo has an interesting approach using SCFs. Each SCF represents a pertinent chemical property and is described with specific geometric rules. The search for equivalent binding sites is performed by detection of similar graphs.[42] The specific geometric rules of each SCF enable the heuristic to be quite fast. So, MED-SuMo provides an interesting and original method to detect structural and functional similarities between protein binding sites. Unlike MED-SuMo, very few methods enable functional classification of sets of binding sites[43] and specific binding sites are usually chosen (protein kinase) for the published work. Comparing our protocol with others is quite difficult.

Here, it is applied in a new clustering approach where the ligand environment is classified. An application to a particular protein fold, the Bergerat ATP-binding fold characterized as the ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase SCOP superfamily is described here. The constituent families are quite different but their ATP binding sites appear quite alike. MED-SMA detects five different clusters. Three out of five are specific to a single family. These three MED-SMA clusters highlight the specificity of the binding sites; for example; no molecule binding to cluster 1's binding site would also bind MED-SMA cluster 2 sites with the same interactions. The fact that the ligands are similar in MED-SMA cluster 1 and 2 (eg, ADP) emphasizes the previous observation. The ligands are the same whereas the binding modes are different. Oppositely, MED-SMA cluster 4 gathers three different families. The 3D superimposition from MED-SuMo, points out the difference of the global fold whereas the Bergerat fold can be observed (white rectangle on Figure 7). Interestingly, SCFs can be found all around the query ligand (cf. Figure 7), meaning that there is a global similarity of the binding sites from the three SCOP families. Moreover, this result is consistent with the experimental data as the proteins from these three SCOP families all bind radicicol.[28–30,44]

These different results demonstrate the ability of the method to gather binding sites with related binding modes. This kind of relationship between families is very interesting and their identification is a direct application for MED-SMA. Moreover, with this kind of association, we can validate the assertion that functions can be assigned to unknown proteins by associating them to a specific best matching cluster. Matching clusters rather than single structures overcomes most of the noise in both the assignments and in the functions of those assigned matches. Other applications are planned, for example, a more general kinase classification using MED-SMA is under investigation.

Conclusions

This example clearly shows that our approach is well suited for finding common and distinct characteristics of ligand binding pockets. Thus, close proteins can have different local binding modes, while more distant ones can share common binding features ie, a potential cross-reaction may be possible. For instance, proteins associated to radicicol are found in the same MED-SMA clusters. This approach is clearly applicable to structural genomics research. As noted by Ferrè and colleagues, functional patches associated to a large collection of protein surface cavities can be used to provide functional clues for protein with unknown structures.[45] This observation is shared from our study. Thus, MED-SuMo is an approach that may improve the efficiency and effectiveness of early steps along the drug discovery path, improving early lead choices, enhancing poor leads, or aiding multivariate optimizations. This study further demonstrates that MED-SuMo is appropriate for both annotating protein structures and for deriving structural functional classifications.

Finally, with its effectiveness at dealing with the entire PDB, and the parallelisation of the computational process in course, MED-SuMo is well-suited to large-scale applications. In fact it is currently used to resolve the big challenge of the POPS project (see http://www.pops-systematic.org/) in classifying every binding site represented in the PDB.

Software licensing

Commercial information regarding MED-SuMo is available at http://www.medit.fr/. Questions about MED-SuMo licensing should be addressed to info@medit.fr. Researcher from the Inserm Institute UMR-S 726 has no financial interests in MEDIT and collaborates with this company only for the present project. Therefore, MEDIT SA has the exclusivity for MED-SuMo sales.

Acknowledgments

This work was supported by French Institute for Health and Medical Care (INSERM) and University Denis Diderot Paris 7. ODA's PhD is financed by the French technical research association (ANRT) through a CIFRE grant. MEDIT holds all the rights on the presented methodology. The authors are indebted to S. Adcock for useful comments on the manuscript.

References

1. Wendt KU, Weiss MS, Cramer P, Heinz DW. Structures and diseases. *Nat Struct Mol Biol*. 2008;15:117–120.
2. Guido RV, Oliva G, Andricopulo AD. Virtual screening and its integration with modern drug design technologies. *Curr Med Chem*. 2008;15:37–46.
3. Waszkowycz B. Towards improving compound selection in structure-based virtual screening. *Drug Discov Today*. 2008;13:219–226.
4. Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*. 2004;32:D129–133.
5. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. *J Mol Biol*. 2002;324:105–121.
6. Fox BG, Goulding C, Malkowski MG, Stewart L, Deacon A. Structural genomics: from genes to structures with valuable materials and many questions in between. *Nat Methods*. 2008;5:129–132.
7. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389–3402.
8. Bairoch A. PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res*. 1991;19(Suppl):2241–2245.
9. Schmitt S, Kuhn D, Klebe G. A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol*. 2002;323:387–406.
10. Shulman-Peleg A, Nussinov R, Wolfson HJ. SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res*. 2005;33:W337–41.
11. Baroni M, Cruciani G, Sciabola S, Perruccio F, Mason JS. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *J Chem Inf Model*. 2007;47:279–294.
12. Powers R, Copeland JC, Germer K, Mercier KA, Ramanathan V, Revesz P. Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins*. 2006;65:124–135.
13. Standley DM, Kinjo AR, Kinoshita K, Nakamura H. Protein structure databases with new web services for structural biology and biomedical research. *Brief Bioinform*. 2008;9:276–285.

14. Brylinski M, Prymula K, Jurkowski W, et al. Prediction of functional sites based on the fuzzy oil drop model. *PLoS Comput Biol*. 2007;3:e94.
15. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*. 1996;257:342–358.
16. Mihalek I, Res I, Lichtarge O. Evolutionary trace report_maker: a new type of service for comparative analysis of proteins. *Bioinformatics*. 2006;22:1656–1657.
17. Morgan DH, Kristensen DM, Mittelman D, Lichtarge O. ET viewer: an application for predicting and visualizing functional sites in protein structures. *Bioinformatics*. 2006;22:2049–2050.
18. Jambon M, Imberty A, Deleage G, Geourjon C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins*. 2003;52:137–145.
19. Jambon M, Andrieu O, Combet C, Deleage G, Delfaud F, Geourjon C. The SuMo server: 3D search for protein functional sites. *Bioinformatics*. 2005;21:3929–2930.
20. Doppelt O, Moriaud F, Bornot A, de Brevern AG. Functional annotation strategy for protein structures. *Bioinformation*. 2007;1:357–359.
21. Jefferson ER, Walsh TP, Barton GJ. A comparison of SCOP and CATH with respect to domain-domain interactions. *Proteins*. 2008;70:54–62.
22. Andreeva A, Howorth D, Chandonia JM, et al. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res*. 2008;36:D419–425.
23. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 1995;247:536–540.
24. Picard D. Heat-shock protein 90, a chaperone for folding and regulation. *Cell Mol Life Sci*. 2002;59:1640–1648.
25. Whitesell L, Lindquist SL. HSP90 and the chaperoning of cancer. *Nat Rev Cancer*. 2005;5:761–772.
26. Goetz MP, Toft DO, Ames MM, Erlichman C. The Hsp90 chaperone complex as a novel target for cancer therapy. *Ann Oncol*. 2003;14: 1169–1176.
27. Zhang T, Hamza A, Cao X, Wang B, Yu S, Zhan CG, Sun D. A novel Hsp90 inhibitor to disrupt Hsp90/Cdc37 complex against pancreatic cancer cells. *Mol Cancer Ther*. 2008;7:162–170.
28. Roe SM, Prodromou C, O'Brien R, Ladbury JE, Piper PW, Pearl LH. Structural basis for inhibition of the Hsp90 molecular chaperone by the antitumor antibiotics radicicol and geldanamycin. *J Med Chem*. 1999;42:260–266.
29. Guarnieri MT, Zhang L, Shen J, Zhao R. The Hsp90 inhibitor radicicol interacts with the ATP-binding pocket of bacterial sensor kinase PhoQ. *J Mol Biol*. 2008;379:82–93.
30. Corbett KD, Berger JM. Structural basis for topoisomerase VI inhibition by the anti-Hsp90 drug radicicol. *Nucleic Acids Res*. 2006;34: 4269–4277.
31. van Dongen S. *Graph Clustering by Flow Simulation*. PhD thesis. Utrecht, The Netherlands: University of Utrecht; 2000.
32. Enright AJ, Ouzounis CA. BioLayout–an automatic graph layout algorithm for similarity visualization. *Bioinformatics*. 2001;17:853–854.
33. Goldovsky L, Cases I, Enright AJ, Ouzounis CA. BioLayout(Java): versatile network visualisation of structural and functional relationships. *Appl Bioinformatics*. 2005;4:71–74.
34. Bellon S, Parsons JD, Wei Y, et al. Crystal structures of *Escherichia coli* topoisomerase IV ParE subunit (24 and 43 kilodaltons): a single residue dictates differences in novobiocin potency against topoisomerase IV and DNA gyrase. *Antimicrob Agents Chemother*. 2004;48:1856–1864.
35. Dessailly BH, Lensink MF, Wodak SJ. Relating destabilizing regions to known functional sites in proteins. *BMC Bioinformatics*. 2007;8:141.
36. Brown DP, Krishnamurthy N, Sjolander K. Automated protein subfamily identification and classification. *PLoS Comput Biol*. 2007;3:e160.
37. Ramensky V, Sobol A, Zaitseva N, Rubinov A, Zosimov V. A novel approach to local similarity of protein binding sites substantially improves computational drug design results. *Proteins*. 2007; 69:349–357.
38. Mao L, Wang Y, Liu Y, Hu X. Molecular determinants for ATP-binding in proteins: a data mining and quantum chemical analysis. *J Mol Biol*. 2004;336:787–807.
39. Niefind K, Putter M, Guerra B, Issinger OG, Schomburg D. GTP plus water mimic ATP in the active site of protein kinase CK2. *Nat Struct Biol*. 1999;6:1100–1103.
40. Yde CW, Ermakova I, Issinger OG, Niefind K. 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;347:399–414.
41. Nebel JC, Herzyk P, Gilbert DR. Automatic generation of 3D motifs for classification of protein binding sites. *BMC Bioinformatics*. 2007;8:321.
42. Wu S, Liang MP, Altman RB. The SeqFEATURE library of 3D functional site models: comparison to existing methods and applications to protein function annotation. *Genome Biol*. 2008;9:R8.
43. Kuhn D, Weskamp N, Hullermeier E, Klebe G. Functional classification of protein kinase binding sites using Cavbase. *Chem Med Chem*. 2007;2:1432–1447.
44. Besant PG, Lasker MV, Bui CD, Turck CW. Inhibition of branched-chain alpha-keto acid dehydrogenase kinase and Sln1 yeast histidine kinase by the antifungal antibiotic radicicol. *Mol Pharmacol*. 2002;62:289–296.
45. Ferre F, Ausiello G, Zanzoni A, Helmer-Citterich M. Functional annotation by identification of local surface similarities: a novel tool for structural genomics. *BMC Bioinformatics*. 2005;6:194.

Supplementary data

Supplementary table 1

MED-SMA cluster ID	PDB_LIG_ID	Ligand name	SCOP family
CL_1	1EI1_1_92	ANP	DNA_Gyrase_B_EColi
CL_1	1EI1_2_90	ANP	DNA_Gyrase_B_EColi
CL_1	1MX0_1_31	ANP	TOPO_VI
CL_1	1MX0_2_29	ANP	TOPO_VI
CL_1	1MX0_3_28	ANP	TOPO_VI
CL_1	1MX0_4_25	ANP	TOPO_VI
CL_1	1MX0_5_22	ANP	TOPO_VI
CL_1	1MX0_6_21	ANP	TOPO_VI
CL_1	1PVG_1_1	ANP	DNA_TOPO_II_Byeast
CL_1	1PVG_2_0	ANP	DNA_TOPO_II_Byeast
CL_1	1QZR_1_117	CDX	DNA_TOPO_II_Byeast
CL_1	1QZR_2_113	ANP	DNA_TOPO_II_Byeast
CL_1	1QZR_3_111	ANP	DNA_TOPO_II_Byeast
CL_1	1S16_1_102	ANP	TOPO_IV
CL_1	1S16_2_100	ANP	TOPO_IV
CL_1	1Z59_1_17	ADP	TOPO_VI
CL_1	1Z5A_1_11	ADP	TOPO_VI
CL_1	1Z5A_2_8	ADP	TOPO_VI
CL_1	1Z5B_1_86	ADP	TOPO_VI
CL_1	1Z5B_2_84	ADP	TOPO_VI
CL_1	1Z5C_1_9	ADP	TOPO_VI
CL_1	1Z5C_2_5	ADP	TOPO_VI
CL_2	1GJV_1_112	SAP	alpha-ketoacid_dehydrogenase_kinase
CL_2	1GKZ_1_33	ADP	alpha-ketoacid_dehydrogenase_kinase
CL_2	1I5D_1_118	128	Histidine_Kinase_CheA
CL_2	1ID0_1_10	ANP	Histidine_Kinase_PhoQ
CL_2	1JM6_2_61	ADP	Pyruvate_dehydrogenase_kinase
CL_2	1L0O_1_75	ADP	Anti-sigma_factor_spoIIab
CL_2	1L0O_2_73	ADP	Anti-sigma_factor_spoIIab
CL_2	1TH8_1_123	ADP	Anti-sigma_factor_spoIIab
CL_2	1TH8_1_124	ADP	Anti-sigma_factor_spoIIab
CL_2	1THN_1_104	ADP	Anti-sigma_factor_spoIIab
CL_2	1THN_2_101	ADP	Anti-sigma_factor_spoIIab
CL_2	1TID_1_35	ATP	Anti-sigma_factor_spoIIab
CL_2	1TID_2_32	ATP	Anti-sigma_factor_spoIIab
CL_2	1TIL_1_27	ATP	Anti-sigma_factor_spoIIab
CL_2	1TIL_2_24	ATP	Anti-sigma_factor_spoIIab
CL_2	1TIL_3_23	ATP	Anti-sigma_factor_spoIIab
CL_2	2C2A_1_120	ADP	Sensor_histidine_kinase_TM0853
CL_2	2CH4_1_56	ANP	Histidine_Kinase_CheA
CL_3	1AJ6_1_76	NOV	DNA_GYRASE_B_EColi
CL_3	1KIJ_1_66	NOV	DNA_GYRASE_B_TT
CL_3	1KIJ_2_64	NOV	DNA_GYRASE_B_TT
CL_3	1KZN_1_52	CBN	DNA_GYRASE_B_EColi

(*Continued*)

Supplementary table 1 (*Continued*)

MED-SMA cluster ID	PDB_LIG_ID	Ligand name	SCOP family
CL_3	1S14_1_105	NOV	TOPO_IV
CL_3	1S14_2_103	NOV	TOPO_IV
CL_4	1A4H_1_62	GMY	HSP90_Yeast
CL_4	1AM1_1_37	ADP	HSP90_Yeast
CL_4	1AMW_1_7	ADP	HSP90_Yeast
CL_4	1B62_1_16	ADP	MulL
CL_4	1B63_1_91	ANP	MulL
CL_4	1BGQ_1_55	RDC	HSP90_Yeast
CL_4	1BYQ_1_99	ADP	HSP90_Human
CL_4	1EA6_1_110	ADP	PMS2
CL_4	1EA6_2_109	ADP	PMS2
CL_4	1H7U_1_46	ATG	PMS2
CL_4	1H7U_2_44	ATG	PMS2
CL_4	1JM6_1_63	ADP	Pyruvate_dehydrogenase_kinase
CL_4	1NHH_1_43	ANP	MulL
CL_4	1NHI_1_108	ANP	MulL
CL_4	1NHJ_1_42	ANP	MulL
CL_4	1OSF_1_83	KOS	HSP90_Human
CL_4	1QY5_1_13	NEC	HSP90_Dog
CL_4	1QY8_1_87	RDI	HSP90_Dog
CL_4	1QYE_1_143	CDY	HSP90_Dog
CL_4	1TBW_1_107	AMP	HSP90_Dog
CL_4	1TBW_2_106	AMP	HSP90_Dog
CL_4	1TC0_1_125	ATP	HSP90_Dog
CL_4	1TC0_2_121	ATP	HSP90_Dog
CL_4	1TC6_1_116	ADP	HSP90_Dog
CL_4	1TC6_2_114	ADP	HSP90_Dog
CL_4	1U0Y_1_26	PA7	HSP90_Dog
CL_4	1U0Z_1_95	RDC	HSP90_Dog
CL_4	1U0Z_6_93	RDC	HSP90_Dog
CL_4	1U2O_1_3	NEC	HSP90_Dog
CL_4	1U2O_2_2	NEC	HSP90_Dog
CL_4	1UY6_1_88	PU3	HSP90_Human
CL_4	1UY7_1_6	PU4	HSP90_Human
CL_4	1UY8_1_82	PU5	HSP90_Human
CL_4	1UY9_1_4	PU6	HSP90_Human
CL_4	1UYC_1_144	PU7	HSP90_Human
CL_4	1UYD_1_74	PU8	HSP90_Human
CL_4	1UYE_1_141	PU9	HSP90_Human
CL_4	1UYF_1_71	PU1	HSP90_Human
CL_4	1UYG_1_138	PU2	HSP90_Human
CL_4	1UYH_1_68	PU0	HSP90_Human
CL_4	1UYI_1_135	PUZ	HSP90_Human
CL_4	1UYK_1_133	PUX	HSP90_Human
CL_4	1UYM_1_132	PU3	HSP90_Human
CL_4	1YC1_1_15	4BC	HSP90_Human
CL_4	1YC3_1_14	4BC	HSP90_Human

(*Continued*)

Supplementary table 1 (*Continued*)

MED-SMA cluster ID	PDB_LIG_ID	Ligand name	SCOP family
CL_4	1YC4_1_89	43P	HSP90_Human
CL_4	1YET_1_39	GDM	HSP90_Human
CL_4	1YSZ_1_131	NEC	HSP90_Dog
CL_4	1YT0_1_80	ADP	HSP90_Dog
CL_4	1ZW9_1_137	H64	HSP90_Yeast
CL_4	1ZWH_1_58	RDE	HSP90_Yeast
CL_4	2BRC_1_20	CT5	HSP90_Yeast
CL_4	2BRE_1_19	KJ2	HSP90_Yeast
CL_4	2BRE_2_18	KJ2	HSP90_Yeast
CL_4	2BSM_1_77	BSM	HSP90_Human
CL_4	2BT0_1_81	CT5	HSP90_Human
CL_4	2BT0_2_79	CT5	HSP90_Human
CL_4	2BYH_1_69	2D7	HSP90_Human
CL_4	2BYI_1_134	2DD	HSP90_Human
CL_4	2BZ5_1_70	AB4	HSP90_Human
CL_4	2BZ5_2_65	AB4	HSP90_Human
CL_4	2CCS_1_98	4BH	HSP90_Human
CL_4	2CCT_1_30	2E1	HSP90_HumanC
CL_4	2CCU_1_97	2D9	HSP90_Human
CL_4	2CDD_1_96	CT5	HSP90_Human
CL_4	2CDD_2_94	CT5	HSP90_Human
CL_4	2EXL_1_41	GMY	HSP90_Dog
CL_4	2EXL_2_40	GMY	HSP90_Dog
CL_4	2FWY_1_12	H64	HSP90_Human
CL_4	2FWZ_1_85	H71	HSP90_Human
CL_4	2FXS_1_78	RDA	HSP90_Yeast
CL_4	2FYP_1_60	RDE	HSP90_Dog
CL_4	2FYP_2_59	RDE	HSP90_Dog
CL_4	2GFD_1_72	RDA	HSP90_Dog
CL_4	2GFD_2_67	RDA	HSP90_Dog
CL_4	2GQP_1_130	PA7	HSP90_Dog
CL_4	2GQP_2_128	PA7	HSP90_Dog
CL_4	2H55_1_122	DZ8	HSP90_Human
CL_4	2H8M_1_139	NEI	HSP90_Dog
CL_4	2H8M_2_136	NEI	HSP90_Dog
CL_4	2HCH_1_142	N5A	HSP90_Dog
CL_4	2HCH_2_140	N5A	HSP90_Dog
CL_4	2HG1_1_36	N5O	HSP90_Dog
CL_4	2HG1_2_34	N5O	HSP90_Dog
CL_4	2HKJ_1_38	RDC	TOPOVI
CL_4	2IWS_1_48	NP4	HSP90_Yeast
CL_4	2IWU_1_45	NP5	HSP90_Yeast
CL_4	2IWX_1_00	M1S	HSP90_Yeast
CL_4	2UWD_1_126	2GG	HSP90_Human
CL_5	1I58_1_129	ACP	Histidine_Kinase_CheA
CL_5	1I58_2_127	ADP	Histidine_Kinase_CheA
CL_5	1I59_1_57	ANP	Histidine_Kinase_CheA

(*Continued*)

Supplementary table 1 (*Continued*)

MED-SMA cluster ID	PDB_LIG_ID	Ligand name	SCOP family
CL_5	1I59_2_54	ADP	Histidine_Kinase_CheA
CL_5	1I5A_1_51	ACP	Histidine_Kinase_CheA
CL_5	1I5A_2_49	ACP	Histidine_Kinase_CheA
CL_5	1I5B_1_119	ANP	Histidine_Kinase_CheA
CL_5	1I5B_2_115	ANP	Histidine_Kinase_CheA
CL_5	1I5C_1_50	ADP	Histidine_Kinase_CheA
CL_5	1I5C_2_47	ADP	Histidine_Kinase_CheA
CL_5	2CH4_2_53	ANP	Histidine_Kinase_CheA