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

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

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, 2 rue du Belvédère, 91120, Palaiseau, France.

[2]INSERM UMR-S 726, Equipe de Bioinformatique Génomique & Moléculaire, (EBGM), DSIMB, Institut National de Transfusion Sanguine (INTS), Université Paris Diderot - Paris 7, 6 rue Alexandre Cabanel, 75739 Paris Cedex 15, France.

[§]Corresponding author

Email addresses:

OD: olivia.doppelt@univ-paris-diderot.fr

FM: fabrice.moriaud@medit.fr

FD: francois.delfaud@medit.fr

ADB: alexandre.debrevern@univ-paris-diderot.fr

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 characterise similar local regions on protein surfaces. Each amino acid residue's potential chemical interactions are represented by specific Surface Chemical Features (SCF). 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.

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 (Wendt, Weiss et al. 2008). Knowledge of 3D protein structures linked to small-molecules can be used for structure- and ligand-based drug design approaches. (Guido, Oliva et al. 2008; Waszkowycz 2008) 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 (Bartlett, Porter et al. 2002; Porter, Bartlett et al. 2004). 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 endeavours have permitted the resolution of numerous structures characterized as "Unknown function", and multiple functional sites are not associated with any known binding partner (Fox, Goulding et al. 2008). 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*, *e.g.*, PSI-BLAST (Altschul, Madden et al. 1997), or on the characterization of functional patterns or profiles, *e.g.* PROSITE (Bairoch 1991). 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 global fold similarity, *e.g.*, CavBase (Schmitt, Kuhn et al. 2002), SiteEngine (Shulman-Peleg, Nussinov et al. 2005), FLAP (Baroni, Cruciani et al. 2007), CPASS (Powers, Copeland et al. 2006) or eF-seek (Standley, Kinjo et al. 2008).

Some of these approaches share gross features but they also have notable distinctions. For instance, SiteEngine and CavBase both associate physico-chemical proprieties 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 (Shulman-Peleg, Nussinov et al. 2005). CavBase detects related cavities based on a clique detection algorithm (Schmitt, Kuhn et al. 2002) while CPASS comparison uses an alignment of binding site pairs through a root-mean-square-difference (RMSD) scoring function (Powers, Copeland et al. 2006). Roterman has developed an innovative methodology based on irregular hydrophobicity distribution (Brylinski, Prymula et al. 2007). A few other methods are based on the detection of conserved residues to characterise binding sites, e.g. Evolutionary trace method (Lichtarge, Bourne et al. 1996; Mihalek, Res et al. 2006; Morgan, Kristensen et al. 2006) or sequence alignment with a dedicated dataset as Catalytic Site Atlas (CSA) (Porter, Bartlett et al. 2004).

In this research area, SuMo is a powerful technology to localize similar local regions on protein surfaces *i.e.*, binding sites (Jambon, Imberty et al. 2003). 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 motif against all the binding sites of the PDB can be performed in a few minutes (Jambon, Andrieu et al. 2005). MED-SuMo is the latest evolution of SuMo software developed by MEDIT-SA (MEDIT-SA). Recent developments have improved its binding site database, and have included novel functional annotation tools as presented in a recent study (Doppelt, Moriaud et al. 2007).

Proteins are also classified according to their folds (Jefferson, Walsh et al. 2008), *e.g.*, SCOP (Structural Classification of Proteins) (Murzin, Brenner et al. 1995; Andreeva, Howorth et al. 2008), that provides a manually refined classification with detailed and comprehensive descriptions of the structural and evolutionary relationships of the known protein structure (Murzin, Brenner et al. 1995; Andreeva, Howorth et al. 2008). 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, (Picard 2002),

apoptosis and tumour repression (Whitesell and Lindquist 2005). It binds ATP (see Figures 2a and 2b) and is the target of some innovative drugs including geldanamycin which has enables 50% reduction of tumour growth (Goetz, Toft et al. 2003) and celasterol which disrupts interactions between HSP90 and Cdc37 in pancreatic cancer cells (Zhang, Hamza et al. 2008). Some recent research focussed on a new potential drug, radicicol. This molecule has a very high affinity for HSP90 (20 nM) (Roe, Prodromou et al. 1999). Figure 3 shows the association of the drug with the HSP90 at the binding site normally filled with a natural ligand (Roe, Prodromou et al. 1999). However, radicicol is not specific to HSP90 as it binds bacterial Sensor Kinase PhoQ (Guarnieri, Zhang et al. 2008), and topoisomerase VI (Corbett and Berger 2006). 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).

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 pairwise and classified using the Markov Clustering algorithm (MCL) (van Dongen 2000). Results from this method highlight common and distinct functional features between the analysed proteins.

Materials and methods

Protein structure database.

SCOP web site provides the list of proteins associated to a selected fold (Murzin, Brenner et al. 1995). The *"ATPase domain of HSP90 chaperone / DNA topoisomerase II / histidine kinase"* superfamily contains 116 PDB structures (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 (Jambon, Imberty et al. 2003; Jambon, Andrieu et al. 2005; Doppelt, Moriaud et al. 2007). 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 Surface Chemical Features (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, *i.e.*, a residue is represented by a set of representative SCFs (cf. figures 4a and 4b). Their positions and orientations are filtered as shown in figure 4c. Remaining SCFs are assembled into triplets with specific geometric characteristics e.g. 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 (*i.e.*, those sharing at least two SCFs).

To compare graphs, MED-SuMo looks for compatible triplets; composed of compatible SCF (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 (Jambon, Imberty et al. 2003). 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 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 (Jambon, Imberty et al. 2003; Jambon, Andrieu et al. 2005; Doppelt, Moriaud et al. 2007). 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 Markov Clustering algorithm (MCL (van Dongen 2000)). 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 region 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 5.d). MCL is used to interpret the matrix through 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 (Enright and Ouzounis 2001; Goldovsky, Cases et al. 2005).

Results

MED-SMA Classification

To generate the MED-SuMo database, only binding sites co-crystallized with ligands with more than ten atoms are selected. 101 of the originally selected 116 PDB structures satisfy this filter. This yields a total of 146 binding sites in the final database. Several kinds of ligands are present, purines *e.g.* Adenosine Tri-Phosphate or N-Ethyl-5'-Carboxamido Adenosine, or potential drugs, *e.g.* Radicicol or Novobiocin. Of these 146 binding sites, 78 are from HSP90, 38 from Topoisomerase / MutL, 26 are from Histidine Kinase and 4 are from α-ketoacid dehydrogenase kinase C (BCK). The database parameters are set to a ligand radius of 6.0 Å and triangle parameters of 13Å and 39 Å (respectively *edge_max* and *max_edge_sum*). To classify this dataset, MED-SMA takes around 2 minutes on a 4 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 5 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 (*i.e.*, 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 respectively 22 and 6 binding sites of the 38 proteins of the topoisomerase / MutL / DNA gyrase family. 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 (Bellon, Parsons et al. 2004). So, our results reinforce the study of Bellon and co-workers. Moreover, it characterizes with elegance the fact that these two distinct local conformations are found in different related proteins.

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 1 topoisomerase VI, 5 MutL and 4 PMS2) and 1 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 non-

specific drug radicicol which binds HSP90 and topoisomerase VI (Corbett and Berger 2006), 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.

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 (Brylinski, Prymula et al. 2007). Several innovative approaches have been proposed, *i.e.* the use of hydrophobicity distribution on protein structures based on the fuzzy oil drop model (Dessailly, Lensink et al. 2007), the destabilization of limited protein regions (Brown, Krishnamurthy et al. 2007), phylogenomic classification of protein sequences

(Ramensky, Sobol et al. 2007) or the classification of known protein catalytic sites (Mao, Wang et al. 2004). Prediction of protein functional sites is an important step to identify small-molecule interactions for drug discovery (Niefind, Putter et al. 1999) and it can be very useful to optimize drug design (Yde, Ermakova et al. 2005). 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, *e.g.*, CavBase uses pseudocenters (Nebel, Herzyk et al. 2007), and the strong hypothesis that chemical similarity and activity are linked. In this field, MED-SuMo has an interesting approach using Surface Chemical Features (SCF). Each SCF represents a pertinent chemical property and is described with specific geometric rules. The search for equivalent binding sites is preformed by detection of similar graphs (Wu, Liang et al. 2008). The specific geometric rules of each SCF enable the heuristic to be quite fast. So, MED-SuMo provides 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 (Kuhn, Weskamp et al. 2007) 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 characterised 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. 3 out of 5 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 (*e.g.* 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 (Roe, Prodromou et al. 1999; Besant, Lasker et al. 2002; Corbett and Berger 2006; Guarnieri, Zhang et al. 2008).

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 to 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 *i.e.* 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 in structural genomics research. As noted by Ferrè *et al.* functional patches associated to a large collection of protein surface cavities can be used to provide functional clues for protein with unknown structures (Ferre, Ausiello et al. 2005). This observation can be 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, its effectiveness at dealing with the entire PDB, and the parallelisation of the computational process in course, show that MED-SuMo is well-suited to large-scale applications. In fact it is currently used to resolve the big challenge of the POPS

project (http://www.pops-systematic.org/) in classifying every binding site represented in the PDB.

Software Licensing.

Commercial information regarding MED-SuMo is available at 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.

Acknowledgements

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.

Figure legends

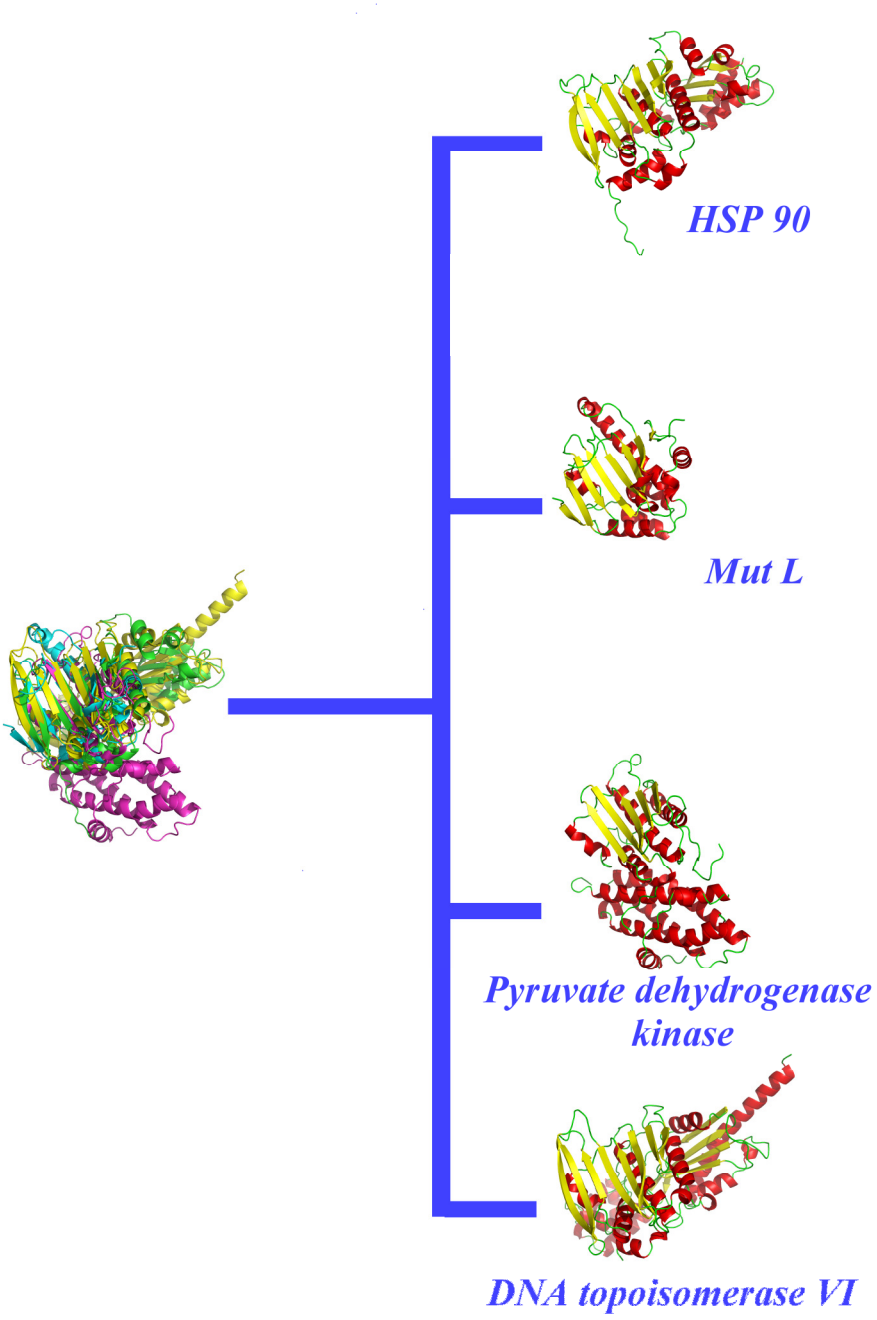

Figure 1: *Heat Shock Protein 90 (HSP90) SCOP superfamly:* ***GHKL***: HSP90, MutL proteins, pyruvate dehydrogenase kinases and DNA topoisomerases VI all share this fold.

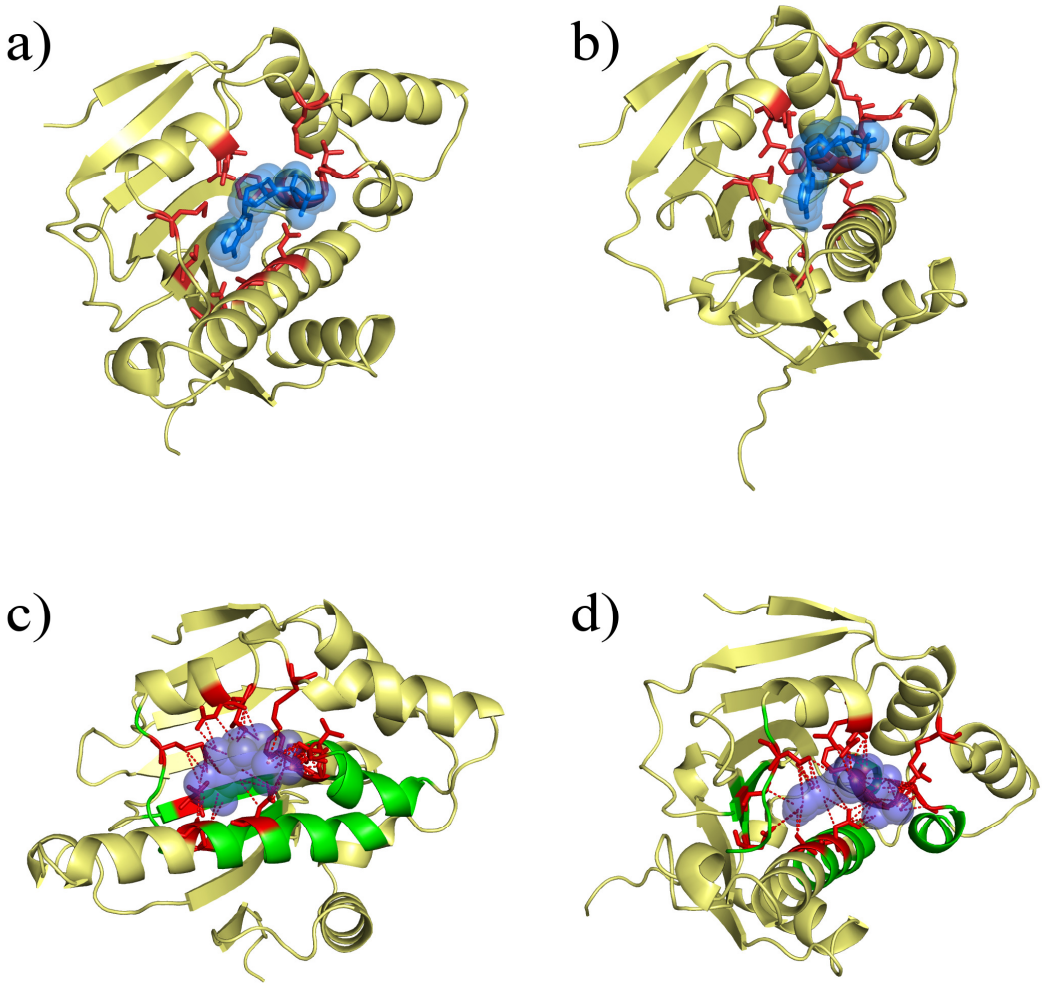

Figure 2: *An example of Heat Shock Protein 90 bound to its natural ligand***.** The protein shown is an HSP90 of *Saccharomyces cerevisiae* (PDB code 1AMW). (a-b) underlines the close contacts (in red) of the ADP (in blue). (c-d) underlines in green the common binding region of this SCOP superfamily.

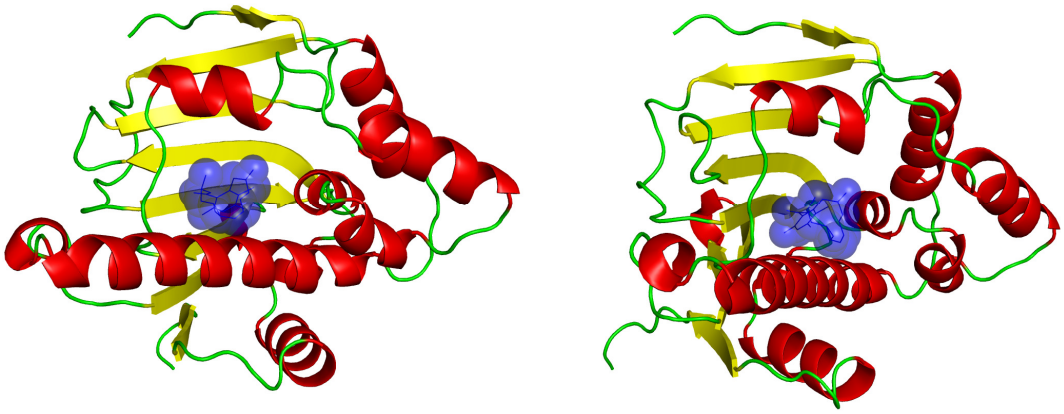

Figure 3: *An example of Heat Shock Protein 90 (HSP90) bound to radicicol*. Both views represent an HSP90 of *Saccharomyces cerevisiae* (PDB code 1BGQ) bound to the drug radicicol shown in blue (see Figure 2 to compare with the natural ligand of HSP90).

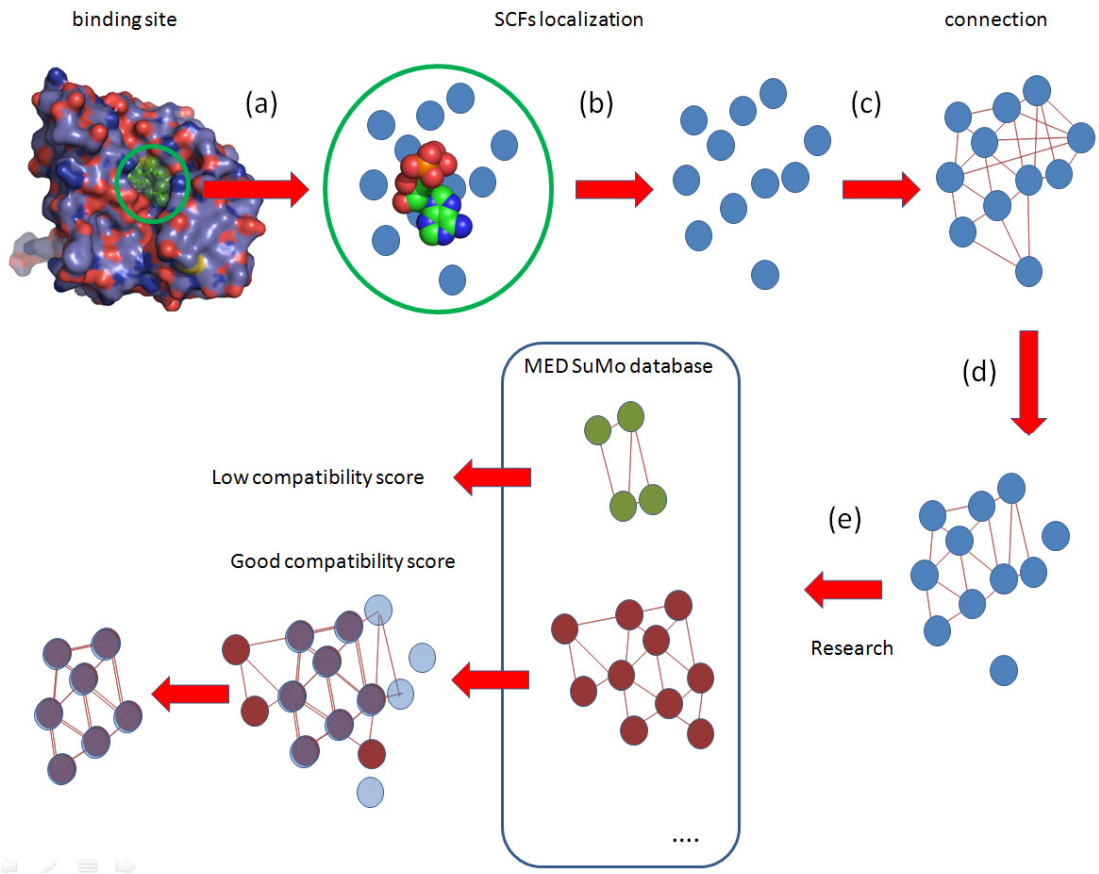

Figure 4: *MED-SuMo comparison procedure.* (a) Localisation of an interesting part of the protein surface often characterized by the presence of a co-crystallized ligand. (b) Surface Chemical Features (SCFs) are displayed on the protein structure through a lexicographic analysis of the PDB files. (c) 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, i.e., they are formed by compatible SCFs. At last, the corresponding graphs (hits) are ranked in regard to their compatibility score.

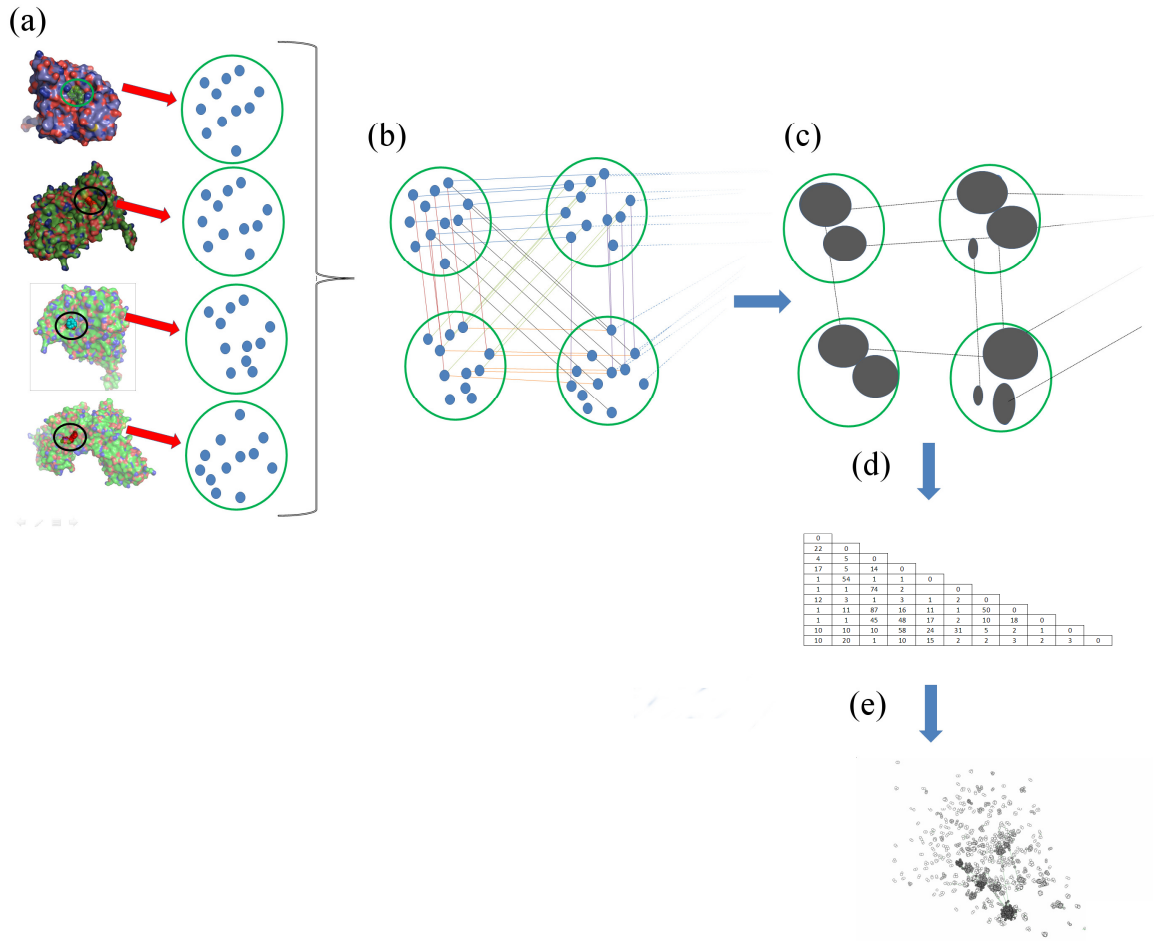

Figure 5: *Global steps of binding site classification heuristic.* MED-SuMo_Multi approach (MED-SMA) can be divided in 5 steps: (a) Database construction: all selected binding sites are stored as graph in the MED-SuMo database. (b) Pairwise comparisons: all binding sites are compared to each other to 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 clustering algorithm) to create clusters of binding sites. (e) Biolayout 2D view of the MED-SMA clusters.

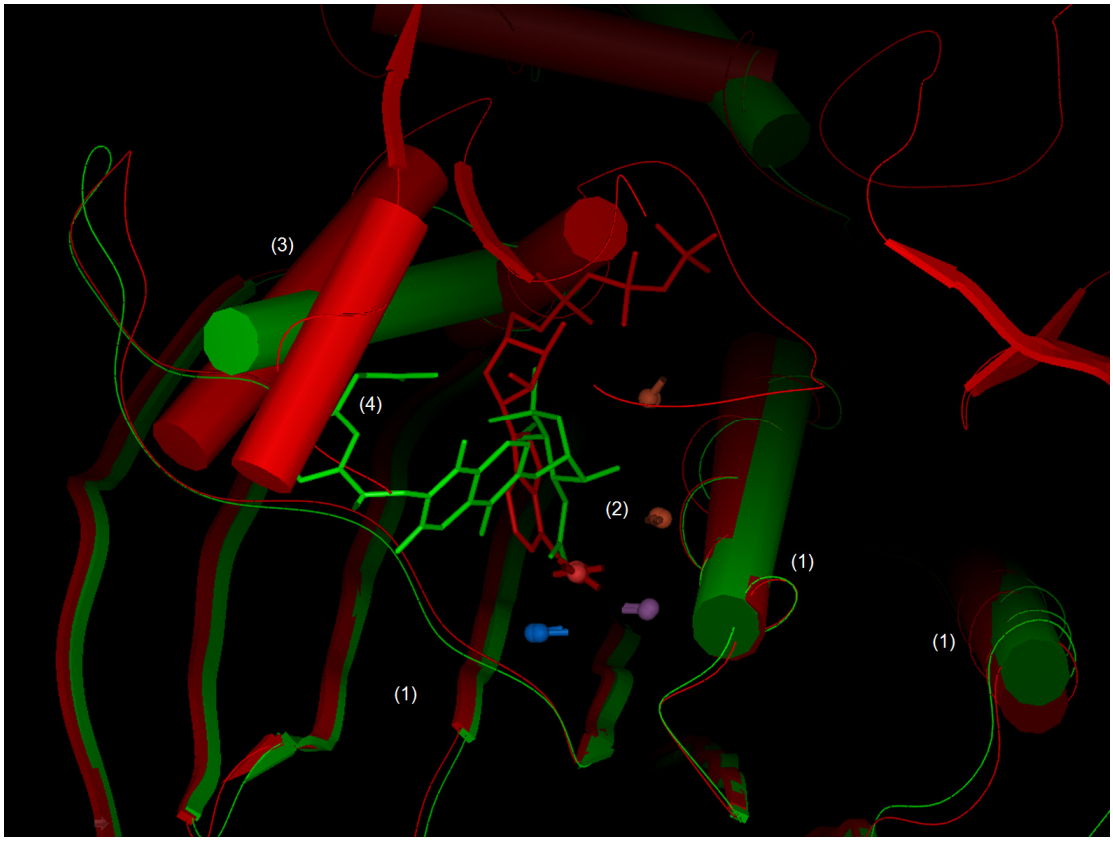

Figure 6: *Superimposition of 2 topoisomerase VI separated by MED-SMA.* PDB codes 1S16 (red) and 1S14 (green) are superimposed. They are both topoisomerase but their binding sites do not share enough similarity to be grouped in the same cluster. This figure is divided by several numbered regions: (1) 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.

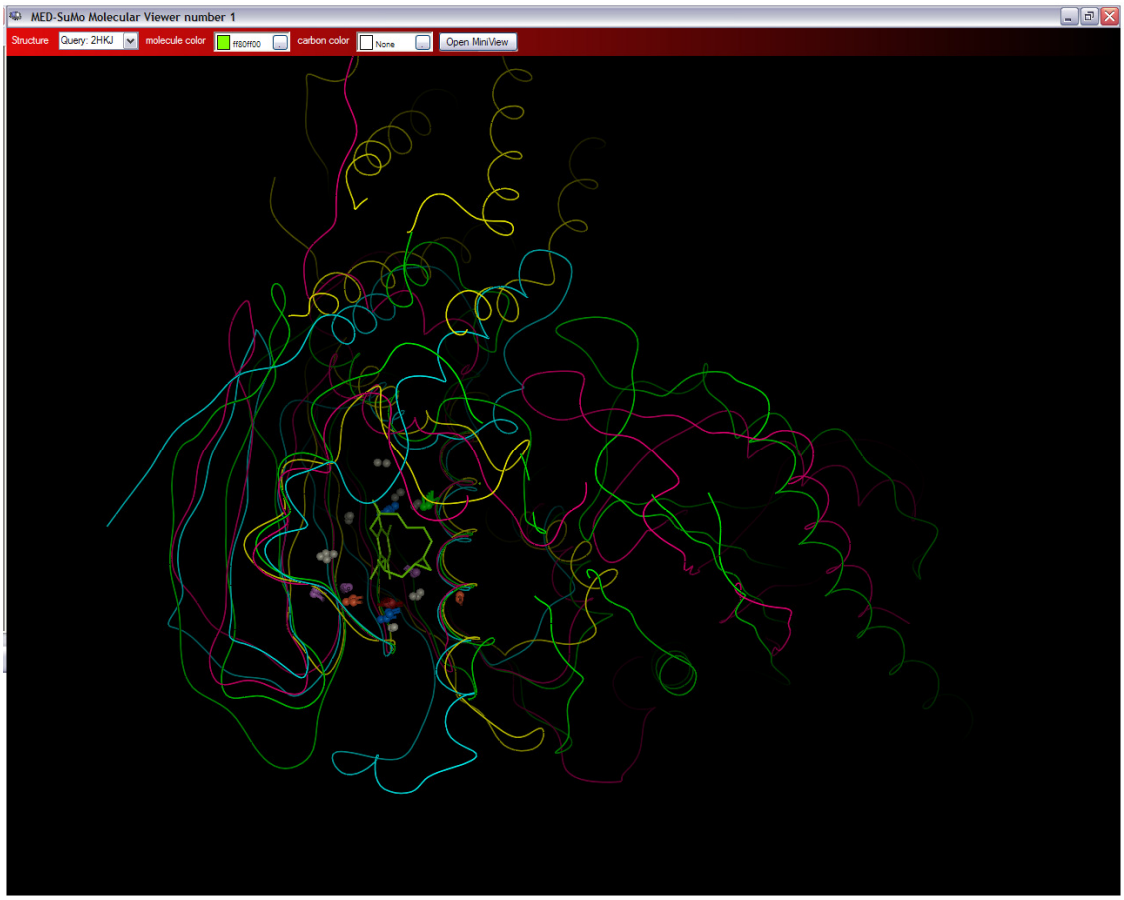

Figure 7: *Superimposition of four proteins from three distinct SCOP families but gathered in the same cluster by MED-SMA.* (PDB codes 2HKJ (green), 2CCT (cyan), 1B63 (pink) 1JM6 (yellow)). The white rectangles show similarities around the ligands and also the helices from the Bergerat fold. The rest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

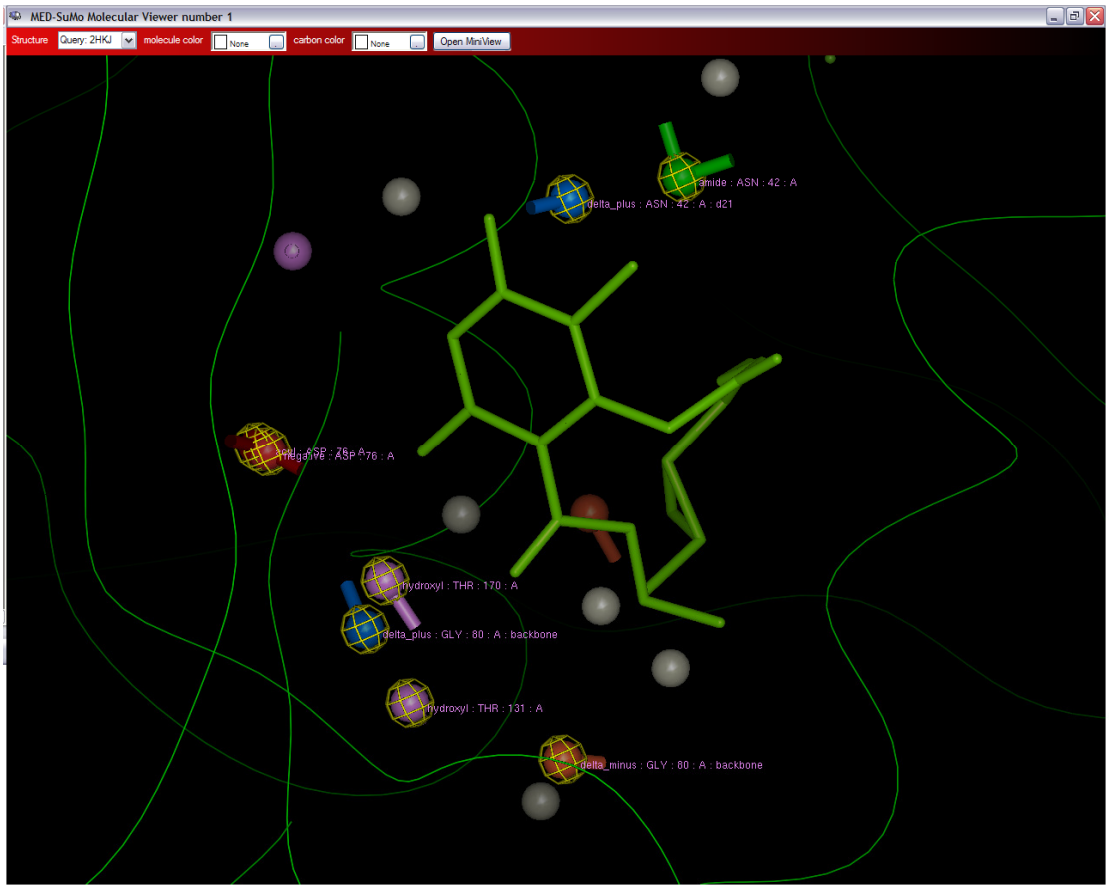

Figure 8: *A close view around the radicicol ligand.* The eight labelled SCFs (circled in yellow) are shared by all superimposed structures in figure 7. They are located all around the ligand meaning that the similarities concern the whole binding site.

Tables

		SCOP superfamily			
		HSP90	DNA gyrase MutL	Histidine Kinase	α-ketoacid dehydrogenase kinase C
MED-SMA clusters	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*

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 respectively: SCOP DNA gyrase/MutL family (for #1 and #3) and Histidine kinase. MED-SuMo clusters #2 and #4 are heterogeneous

References

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Andreeva, A., D. Howorth, et al. (2008). "Data growth and its impact on the SCOP database: new developments." Nucleic Acids Res **36**(Database issue): D419-25.

Bairoch, A. (1991). "PROSITE: a dictionary of sites and patterns in proteins." Nucleic Acids Res **19 Suppl**: 2241-5.

Baroni, M., G. Cruciani, et al. (2007). "A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application." J Chem Inf Model **47**(2): 279-94.

Bartlett, G. J., C. T. Porter, et al. (2002). "Analysis of catalytic residues in enzyme active sites." J Mol Biol **324**(1): 105-21.

Bellon, S., J. D. Parsons, et al. (2004). "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 **48**(5): 1856-64.

Besant, P. G., M. V. Lasker, et al. (2002). "Inhibition of branched-chain alpha-keto acid dehydrogenase kinase and Sln1 yeast histidine kinase by the antifungal antibiotic radicicol." Mol Pharmacol **62**(2): 289-96.

Brown, D. P., N. Krishnamurthy, et al. (2007). "Automated protein subfamily identification and classification." PLoS Comput Biol **3**(8): e160.

Brylinski, M., K. Prymula, et al. (2007). "Prediction of functional sites based on the fuzzy oil drop model." PLoS Comput Biol **3**(5): e94.

Corbett, K. D. and J. M. Berger (2006). "Structural basis for topoisomerase VI inhibition by the anti-Hsp90 drug radicicol." Nucleic Acids Res **34**(15): 4269-77.

Dessailly, B. H., M. F. Lensink, et al. (2007). "Relating destabilizing regions to known functional sites in proteins." BMC Bioinformatics **8**: 141.

Doppelt, O., F. Moriaud, et al. (2007). "Functional annotation strategy for protein structures." Bioinformation **1**(9): 357-9.

Enright, A. J. and C. A. Ouzounis (2001). "BioLayout--an automatic graph layout algorithm for similarity visualization." Bioinformatics **17**(9): 853-4.

Ferre, F., G. Ausiello, et al. (2005). "Functional annotation by identification of local surface similarities: a novel tool for structural genomics." BMC Bioinformatics **6**: 194.

Fox, B. G., C. Goulding, et al. (2008). "Structural genomics: from genes to structures with valuable materials and many questions in between." Nat Methods **5**(2): 129-32.

Goetz, M. P., D. O. Toft, et al. (2003). "The Hsp90 chaperone complex as a novel target for cancer therapy." Ann Oncol **14**(8): 1169-76.
Goldovsky, L., I. Cases, et al. (2005). "BioLayout(Java): versatile network visualisation of structural and functional relationships." Appl Bioinformatics **4**(1): 71-4.
Guarnieri, M. T., L. Zhang, et al. (2008). "The Hsp90 inhibitor radicicol interacts with the ATP-binding pocket of bacterial sensor kinase PhoQ." J Mol Biol **379**(1): 82-93.
Guido, R. V., G. Oliva, et al. (2008). "Virtual screening and its integration with modern drug design technologies." Curr Med Chem **15**(1): 37-46.
Jambon, M., O. Andrieu, et al. (2005). "The SuMo server: 3D search for protein functional sites." Bioinformatics **21**(20): 3929-30.
Jambon, M., A. Imberty, et al. (2003). "A new bioinformatic approach to detect common 3D sites in protein structures." Proteins **52**(2): 137-45.
Jefferson, E. R., T. P. Walsh, et al. (2008). "A comparison of SCOP and CATH with respect to domain-domain interactions." Proteins **70**(1): 54-62.
Kuhn, D., N. Weskamp, et al. (2007). "Functional classification of protein kinase binding sites using Cavbase." ChemMedChem **2**(10): 1432-47.
Lichtarge, O., H. R. Bourne, et al. (1996). "An evolutionary trace method defines binding surfaces common to protein families." J Mol Biol **257**(2): 342-58.
Mao, L., Y. Wang, et al. (2004). "Molecular determinants for ATP-binding in proteins: a data mining and quantum chemical analysis." J Mol Biol **336**(3): 787-807.
Mihalek, I., I. Res, et al. (2006). "Evolutionary trace report_maker: a new type of service for comparative analysis of proteins." Bioinformatics **22**(13): 1656-7.
Morgan, D. H., D. M. Kristensen, et al. (2006). "ET viewer: an application for predicting and visualizing functional sites in protein structures." Bioinformatics **22**(16): 2049-50.
Murzin, A. G., S. E. Brenner, et al. (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." J Mol Biol **247**(4): 536-40.
Nebel, J. C., P. Herzyk, et al. (2007). "Automatic generation of 3D motifs for classification of protein binding sites." BMC Bioinformatics **8**: 321.
Niefind, K., M. Putter, et al. (1999). "GTP plus water mimic ATP in the active site of protein kinase CK2." Nat Struct Biol **6**(12): 1100-3.
Picard, D. (2002). "Heat-shock protein 90, a chaperone for folding and regulation." Cell Mol Life Sci **59**(10): 1640-8.
Porter, C. T., G. J. Bartlett, et al. (2004). "The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data." Nucleic Acids Res **32**(Database issue): D129-33.
Powers, R., J. C. Copeland, et al. (2006). "Comparison of protein active site structures for functional annotation of proteins and drug design." Proteins **65**(1): 124-35.
Ramensky, V., A. Sobol, et al. (2007). "A novel approach to local similarity of protein binding sites substantially improves computational drug design results." Proteins **69**(2): 349-57.
Roe, S. M., C. Prodromou, et al. (1999). "Structural basis for inhibition of the Hsp90 molecular chaperone by the antitumor antibiotics radicicol and geldanamycin." J Med Chem **42**(2): 260-6.
Schmitt, S., D. Kuhn, et al. (2002). "A new method to detect related function among proteins independent of sequence and fold homology." J Mol Biol **323**(2): 387-406.
Shulman-Peleg, A., R. Nussinov, et al. (2005). "SiteEngines: recognition and comparison of binding sites and protein-protein interfaces." Nucleic Acids Res **33**(Web Server issue): W337-41.
Standley, D. M., A. R. Kinjo, et al. (2008). "Protein structure databases with new web services for structural biology and biomedical research." Brief Bioinform.
van Dongen, S. (2000). Graph Clustering by Flow Simulation
University of Utrecht. **PhD**.
Waszkowycz, B. (2008). "Towards improving compound selection in structure-based virtual screening." Drug Discov Today **13**(5-6): 219-26.
Wendt, K. U., M. S. Weiss, et al. (2008). "Structures and diseases." Nat Struct Mol Biol **15**(2): 117-20.
Whitesell, L. and S. L. Lindquist (2005). "HSP90 and the chaperoning of cancer." Nat Rev Cancer **5**(10): 761-72.
Wu, S., M. P. Liang, et al. (2008). "The SeqFEATURE library of 3D functional site models: comparison to existing methods and applications to protein function annotation." Genome Biol **9**(1): R8.

Yde, C. W., I. Ermakova, et al. (2005). "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 **347**(2): 399-414.
Zhang, T., A. Hamza, et al. (2008). "A novel Hsp90 inhibitor to disrupt Hsp90/Cdc37 complex against pancreatic cancer cells." Mol Cancer Ther **7**(1): 162-70.