BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 9 2011, pages 1324–1326
doi:10.1093/bioinformatics/btr120

Structural bioinformatics
Advance Access publication March 12, 2011

sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins

Jamel Meslamani, Didier Rognan and Esther Kellenberger*
Structural Chemogenomics Group, Laboratory of Therapeutic Innovation UMR7200 CNRS/University of Strasbourg, Faculté de Pharmacie, 67400 Illkirch, France

Associate Editor: Burkhard Rost

ABSTRACT

Background: The sc-PDB database is an annotated archive of druggable binding sites extracted from the Protein Data Bank. It contains all-atoms coordinates for 8166 protein–ligand complexes, chosen for their geometrical and physico-chemical properties. The sc-PDB provides a functional annotation for proteins, a chemical description for ligands and the detailed intermolecular interactions for complexes. The sc-PDB now includes a hierarchical classification of all the binding sites within a functional class.

Method: The sc-PDB entries were first clustered according to the protein name indifferent of the species. For each cluster, we identified dissimilar sites (e.g. catalytic and allosteric sites of an enzyme).

Scope and applications: The classification of sc-PDB targets by binding site diversity was intended to facilitate chemogenomics approaches to drug design. In ligand-based approaches, it avoids comparing ligands that do not share the same binding site. In structure-based approaches, it permits to quantitatively evaluate the diversity of the binding site definition (variations in size, sequence and/or structure).

Availability: The sc-PDB database is freely available at: http://bioinfo-pharma.u-strasbg.fr/scPDB.

Contact: ekellen@unistra.fr

Received on December 13, 2010; revised on February 8, 2011; accepted on February 28, 2011

1 INTRODUCTION

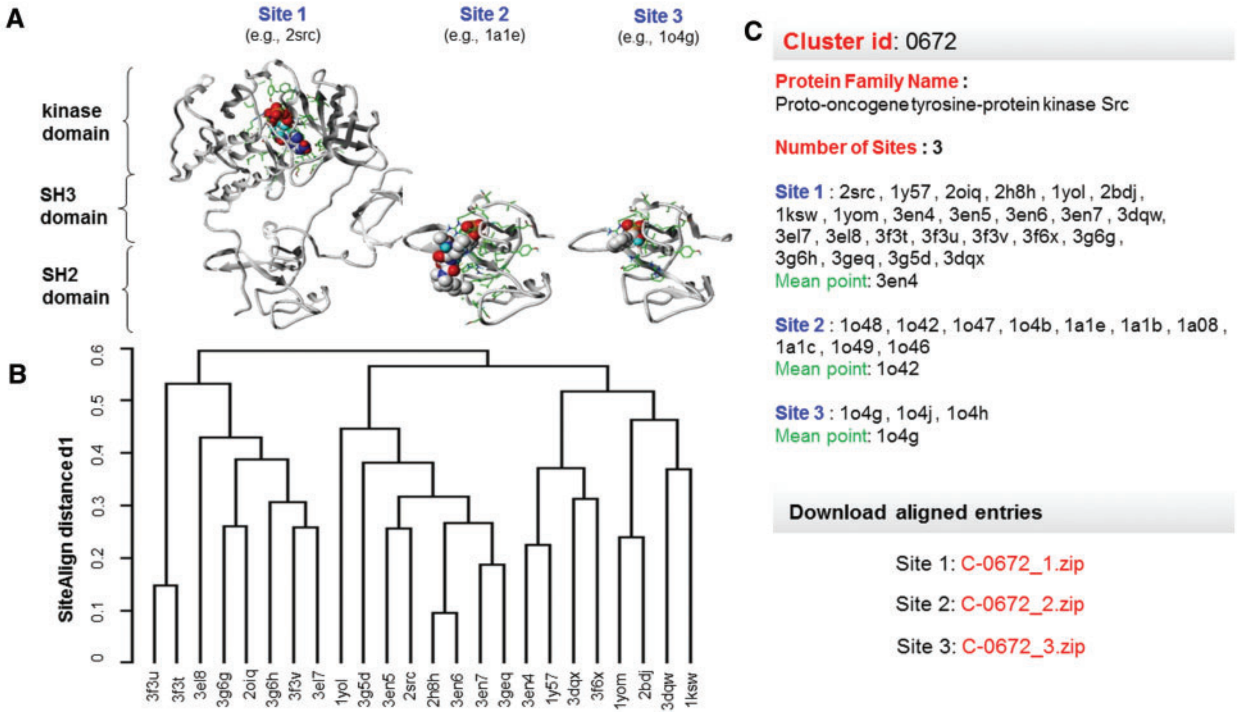
The Protein Data Bank (PDB) is the main public resource of biologically active 3D structures available to study the interactions that govern ligand binding to protein (Berman *et al*., 2007). To assist structure-based approaches in drug design, we have parsed the PDB to identify binding sites suitable for the docking of a 'drug-like' ligand and so have created a database named sc-PDB. The protein selection is based on the molecular weight, buried surface area and chemical structure of ligands as well as the volume of corresponding cavities (Kellenberger *et al*., 2006; Kellenberger *et al*., 2008). Since its creation in 2004, the database is updated annually, with regular improvements. Notably, the curation of ligand chemical structures (2005), the optimization of ligand-bound coordinates (2005) and the systematic description of the ligand binding mode (2006) give significant added values to the structural information contained in the database. The sc-PDB is annotated at a functional level and a sc-PDB *target name* is assigned to each entry. However, two sc-PDB entries with the same sc-PDB *target name* do not necessarily describe an identical binding site. For example, there are 37 copies of the tyrosine-kinase scr in the sc-PDB (Fig. 1A); in 24 entries, the binding site is the ATP binding site of the kinase domain (Site 1), while in other entries, the binding site is located in the SH2 domain and accommodates ligands of variable size (Sites 2 and 3). The present application aims at the distinction of the sc-PDB binding sites for a particular protein. A hierarchical classification was established based on the geometrical and physico-chemical diversity of binding sites. All the binding sites found similar for a given protein were structurally aligned to yield a new set of coordinates for the ligand, site and protein files.

2 CLASSIFICATION OF BINDING SITES BY LOCAL STRUCTURAL SIMILARITY

The sc-PDB data are organized into a hierarchical classification scheme. The first level of the classification is the protein itself, as defined by the sc-PDB *target name*, which combines biological information retrieved from UniprotKB (Apweiler *et al*., 2010) and PDB archives. Typically, the Uniprot recommended name was considered if the PDB file includes the appropriate cross reference (a word matching check validates the consistency between PDB and Uniprot names). In the case of polyproteins (e.g. HIV gag-pol) or multifunctional proteins, the sc-PDB *target name* was chosen according to the domain function (e.g. EC number). Further simplifications of sc-PDB names remove tags for cellular locations and maturation states. Lastly, in the absence of Uniprot reference in the PDB entry, the sc-PDB name was directly extracted from the PDB file and manually curated for a better uniformity within the database. The current 8166 sc-PDB entries represent 1168 protein families and 1470 singletons.

The second level of the classification distinguishes binding sites which significantly differs in size, shape or location in the protein. The third level reports homogeneous classes of structurally similar binding sites. The second and third levels of classification were obtained by clustering all members of a protein family (first-level classification) according to local structural similarity between sites. The all-against-all comparison of sites was performed using the 3D alignment program SiteAlign (Schalon *et al*., 2008). SiteAlign searches for the best superposition of a target site onto a query site (the largest one) by optimizing a global similarity measure which estimates the agreement between the topological and physico-chemical attributes of site-specific 1D fingerprints. The algorithm was successfully applied to identify a novel off-target for

*To whom correspondence should be addressed.



Cluster id: 0672

Protein Family Name :
Proto-oncogene tyrosine-protein kinase Src

Number of Sites : 3

Site 1 : 2src , 1y57 , 2oiq , 2h8h , 1yol , 2bdj , 1ksw , 1yom , 3en4 , 3en5 , 3en6 , 3en7 , 3dqw, 3el7 , 3el8 , 3f3t , 3f3u , 3f3v , 3f6x , 3g6g , 3g6h , 3geq , 3g5d , 3dqx
Mean point: 3en4

Site 2 : 1o48 , 1o42 , 1o47 , 1o4b , 1a1e , 1a1b , 1a08 , 1a1c , 1o49 , 1o46
Mean point: 1o42

Site 3 : 1o4g , 1o4j , 1o4h
Mean point: 1o4g

Download aligned entries

Site 1: C-0672_1.zip

Site 2: C-0672_2.zip

Site 3: C-0672_3.zip

Fig. 1. The proto-oncogene tyrosine kinase scr family. (**A**) 3D representation of the three binding sites. The protein chains are represented using ribbons, the bound ligands using balls and the binding sites residues using capped sticks. (**B**) Hierarchical classification of sc-PDB entries corresponding to Site 1. The dendrogram is available in sc-PDB web pages. (**C**) Web report of the classification.

some protein kinase inhibitors (Defranchi *et al*., 2010). Extensive validation of the program enabled the definition of an alignment score threshold value ($\mathrm{SA_{score}} > 20$) for distinguishing similar from dissimilar binding sites (Schalon *et al*., 2008). The $\mathrm{SA_{score}}$ quantifies all the binding sites discrepancies which result from changes in the sequence (e.g. between orthologous proteins, between wild-type and mutant or chimeric proteins), in the structure (different side chain positioning, motion in the backbone) and in the number of residues in the binding site (which is directly related to the size and the position of the bound ligand). In the current work, the $\mathrm{SA_{scores}}$ of the all-against-all comparison were stored in a distance matrix, which was converted into a Boolean matrix [$A(i,j)=1$ if $\mathrm{SA_{score}} \geq 20$, else $A(i,j)=0$]. The Boolean matrix can be viewed as the adjacent matrix of an undirected graph, where each node represents a site, and an edge is defined between two nodes if the corresponding sites are similar. The Boolean matrix allowed the identification of all graph components (i.e. the maximal connected subgraphs), thereby grouping the sites of a protein family into one or more clusters. Lastly, a hierarchical clustering within clusters of similar sites was generated using the complete linkage method and distance equal to $[1-\ln(\mathrm{SA_{score}})/\ln(\mathrm{SA_{scoremax}})]$.

All sc-PDB entries belonging to the same cluster were structurally aligned to the mean point entry, whose average square distance to all other sites is the smallest in the cluster. The rotation and translation matrix applied to ligand, site and protein coordinates of the target entry were obtained from the global structural alignment of the target and reference proteins using the combinatorial extension (CE) program (Shindyalov and Bourne, 1998). If the proteins contain several peptidic chains, all-against-all chain comparisons were performed, and the transformation which minimizes the distance between the binding site centres was chosen. CE was here preferred over SiteAlign because it finds better alignments for a closely related structural ensemble. In contrast to SiteAlign, CE does not take into account changes in the amino acid type or changes in the rotameric state of residues. CE superimposition of protein structures thus allows the user to directly perceive the contribution of protein flexibility to the $\mathrm{SA_{score}}$. In the example shown in Figure 1, the 24 copies of the kinase domain (here labeled as Site 1) are grouped in three main branches, which are not directly indicative of the binding site flexibility, the ligand chemotype or the ligand binding mode. The absence of correlation between $\mathrm{SA_{score}}$ and deviation in the backbone coordinates of aligned sites rules out the interpretation of the dendrogram in terms of protein flexibility (the mean RMSD computed for the alpha carbon atoms of the residues defining the smallest site is 0.5 ± 0.2 Å). Annotation of the sc-PDB files rather suggests that the dendrogram reflects variations in sequence length (the number of residues in site ranges from 34 to 54) and nature (avian and human proteins, wild-type and mutant proteins).

3 A FREELY AVAILABLE RESOURCE OF 3D-ALIGNED STRUCTURES OF LIGAND-BOUND PROTEIN BINDING SITES

The clustering achieved on the 1168 protein families of the 2010 sc-PDB release produced 783 singletons, 307 classes with two distinct sites, 60 classes with three distinct sites, 10 classes with four distinct sites, 4 classes with five distinct sites, 3 classes with six distinct sites and 1 class with eight sites. The sc-PDB classification is freely available via the database webserver (http://bioinfo-pharma.u-strasbg.fr/scPDB), by selecting the 'Clusters of Binding Site' page. The query form allows search by target name, by number of distinct sites within a protein class or by PDB ID. The result page returns all classes matching the request, provides the user with the detailed content of each class (Fig. 1B and C) and enables the download of MOL2 files of structurally aligned entries. Alternatively, all sc-PDB binding sites similar to a particular binding site entry may be retrieved and ranked by decreasing similarity to the query.

The sc-PDB classification should foster drug design applications for two main reasons: (i) it avoids comparing ligands that do not share the same binding site; (ii) it gives clues to evaluate the influence of ligand binding on binding site diversity for applications in structure-based methods (e.g. docking, site detection and comparison).

ACKNOWLEDGEMENTS

We thank Dr Nathanaël Weill, Dr Gilles Marcou and Dr Jérôme Hert for helpful discussions. We acknowledge the calculation center CC-IN2P3 (IN2P3, CNRS Villeurbanne, France) for providing us with computing resources and GENCI (project x2010075024) for allocation of computing time.

Funding: This work was supported by the Alsace Regional Council.

Conflict of Interest: none declared.

REFERENCES

Apweiler,R. *et al.* (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.

Berman,H. *et al.* (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.

Defranchi,E. *et al.* (2010) Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements. *PLoS ONE*, **5**, e12214.

Kellenberger,E. *et al*. (2008) Ranking targets in structure-based virtual screening of three-dimensional protein libraries: methods and problems. *J. Chem. Inf. Model.*, **48**, 1014–1025.

Kellenberger,E. *et al.* (2006) sc-PDB: an annotated database of druggable binding sites from the protein data bank. *J. Chem. Inf. Model.*, **46**, 717–727.

Schalon,C. *et al.* (2008) A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins*, **71**, 1755–1778.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.