John von Neumann Institute for Computing

Analysis and Classification of the Structural Interactome

J. Teyra, M. Paszkowski-Rogacz, G. Anders,
M. T. Pisabarro

published in

From Computational Biophysics to Systems Biology (CBSB08),
Proceedings of the NIC Workshop 2008,
Ulrich H. E. Hansmann, Jan H. Meinke, Sandipan Mohanty,
Walter Nadler, Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **40**, ISBN 978-3-9810843-6-8, pp. 397-400, 2008.

http://www.fz-juelich.de/nic-series/volume40

Analysis and Classification of the Structural Interactome

Joan Teyra, Maciej Paszkowski-Rogacz, Gerd Anders, and M. Teresa Pisabarro

BIOTEC TU Dresden, Tatzberg 47/49, 01307 Dresden, Germany
E-mail: joan.teyra@biotec.tu-dresden.de

1 Introduction

Protein interactions are essential for intra-cellular communication in biological processes. Proteins are composed of small units or domains that can physically interact together forming multi-domain protein complexes. A single protein can have several binding regions, and each region can engage distinct ligands, either simultaneously or at successive stages of signalling. Detailed information about protein interactions is critical for our understanding of the principles governing protein recognition mechanisms. The structures of many proteins have been experimentally determined in complex with different ligands bound either in the same or different binding regions. Thus, the structural interactome requires the development of tools to classify protein binding regions. A proper classification may provide a general view of the regions that a protein uses to bind others and also facilitate a detailed comparative analysis of the interacting information for specific protein binding regions at atomic level. Such classification might be of potential use for deciphering protein interaction networks, understanding protein function, rational engineering and design. We present the SCOWLP database and web-interface [1, 2], a framework to study protein interfaces and for comparative analysis of protein family binding regions (PBRs).

2 Methodology

SCOWLP was developed following several steps:

2.1 Extraction of Interfaces and Contacting Domains

An accurate definition of the interacting residues is crucial to have a proper clustering of a family PBR. Our database includes all protein-interacting components of the PDB including peptides and solvent, which until now have been excluded from systematic protein interface analysis and databases. The inclusion of water enriches the definition of protein interfaces by considering residues interacting exclusively by water, defined as wet spots [3]. In our database all interface interactions are described at atom, residue and domain level by using interacting rules based on atomic physicochemical criteria [1]. The definition of a domain was extracted from the SCOP database [4]. We consider ”interface” all domain-domain interactions ; that means those belonging to the same protein and also to different proteins. SCOWLP contains 79,803 interfaces contained in 2,561 SCOP families. We grouped the domains participating in each interface by SCOP families, obtaining for each family a list of contacting domains with the residues forming part of the binding region.

2.2 Pair-Wise Structural Alignments (PSAs)

A reliable alignment is indispensable to calculate the similarities among binding regions. For this purpose we used MAMMOTH, which has shown proven accuracy to structurally align protein families [5]. We performed all-against-all PSAs of the contacting domains for each family to be able to measure the similarity among binding regions. SCOWLP contains about 160,000 contacting domains uneven distributed by families. This represents 276 million PSAs performed in a cluster of five Pentium IV 2.6 GHz. The alignments were performed taking the C atoms into account and using a gap penalty function for opening and extension [6]. The root-mean-squared deviation (RMSD) was not considered for measuring the similarity between two interfaces, as the superimposed members of the same family share a common structure.

2.3 Similarity Index (Si)

The residues described in SCOWLP forming and interface were mapped onto the domain-pair structural alignment. We calculated a similarity index (Si) based on the number of interacting residues that overlap and the length of both interacting regions by:

$S_i(a,b) = \frac{2IR_{overlap}(a,b)}{IR_{length(a)} - IR_{gaps(a)} + IR_{length(b)} - IR_{gaps(b)}}$, where a and b represent the two domain structures aligned. The number of interacting residues that match in the PSA is represented by IRoverlap(a, b). This value is divided by the average number of the interacting residues in both domains excluding the interacting residues located in gap regions in the structural alignment (IRgaps).

2.4 Clustering Binding Regions

Based on the calculated Si, we clustered the binding regions of each SCOP family using the agglomerative hierarchical algorithm following several steps: 1) Define as a cluster each contacting domain. 2) Find the closest pair of clusters and merge them into a single cluster. 3) Re-compute the distances between the new cluster and each of the remaining clusters. 4) Repeat steps 2 and 3 until all contacting domains are clustered into a single cluster. To re-compute the distances we used the complete-linkage method, which considers the distance between two clusters to be equal to the minimum similarityof the two members.

2.5 Binding Region Definition by Si Cut-Offs

The result of the clustering can be represented in an intuitive tree or dendrogram, which shows how the individual contacting domains are successively merged at greater distances into larger and fewer clusters. The final PBRs depend on the Si cut-off that is set up. Based on our observations of a representative group of families we set up an empirical maximum similarity cut-off value of 0.4. We pre-calculated the results for Si cut-offs at 0, 0.1, 0.2, 0.3 and 0.4 to offer a range of values that allow flexibility in the final analysis of PBRs. The SCOWLP web application offers the possibility to display the classification at any of these cut-off values. Our classification clustered 160,000 contacting domains from 2,561 families in 9,334 binding regions. About 65% of the families contain more than one binding region. These values are obtained for similarity zero and may vary depending on the similarity cut-off applied.

2.6 Interface Definitions

In order to differentiate binding regions having single-interfaces from multi-interfaces, we identified in each binding region the partner for each contacting domain. Each binding region was divided in sub-clusters when there were different domain families interacting in the same binding region. This resulted in a total of 10,300 interfaces. The classification shows a 78% of the binding regions having a single-interface and the rest having mainly 2 or 3 interfaces per region. These numbers have to be carefully interpreted by taking into account the limitation of the structural information contained in the PDB (i.e. 1,715 binding regions contain a unique member in the PDB and therefore only one known interface per binding region).

3 Conclusions

SCOWLP database contains detailed interacting information of protein interfaces and the hierarchical classification of PBRs. It represents a framework to study protein interfaces and for comparative analysis of protein family binding regions. This comparison can be performed at atomic level and allows the user to study interactome conservation and variability. The new SCOWLP classification may be of great utility for reconstruction of protein complexes, understanding protein networks and ligand design. The web application is available at `http://www.scowlp.org`.

References

1. J. Teyra, A. Doms, M. Schroeder, M. T. Pisabarro, *SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces*, BMC Bioinformatics. **7**, 1004, 2006.
2. J. Teyra, M. Paszkowski-Rogacz, G. Anders, M. T. Pisabarro, *SCOWLP classification: structural comparison and analysis of protein binding regions*, BMC Bioinformatics. **9**, 9, 2008.
3. J. Teyra, M. T. Pisabarro, *Characterization of interfacial solvent in protein complexes and contribution of wet spots to the interface description*, Proteins **67**, 2577–2587, 2007.
4. `http://scop.mrc-lmb.cam.ac.uk`.
5. D. Lupyan D, A. Leo-Macias A, A. R. Ortiz, *A new progressive-iterative algorithm for multiple structure alignment*, Bioinformatics **21**, 3255–3263, 2005.
6. A. R. Ortiz, C. E. Strauss, O. C. Olmea, *MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison*, Protein Sci **11**, 2606–2621, 2002.