Poster I-41

Single-family analysis of binding site structural similarities

Authors:
Rafael Najmanovich (*EBI*)
Richard Morris (*EBI*)
Janet Thornton (*EBI*)

Short Abstract: We describe a method for the detection of pairwise local structural similarities between members of a given protein family using all non-hydrogen atoms in protein clefts allowing larger, over-predicted and apo-form binding sites to be analyzed. We uncover previously unknown similarities between members of the human cytosolic sulfotransferase family.

Long Abstract:

A renewed emphasis is being placed in recent years into family-wide approaches to structural genomics with the goal of reaching a better understanding of protein function vis-à-vis molecular recognition and ligand binding.

While the overall fold as well as the general function is conserved within a given protein family, variations in specific positions within the fold or larger differences in loops can be responsible for subtle changes in the regulation or function of various members of the protein family and might be responsible particularly for differences in substrate specificity and selectivity. Furthermore, local structural similarities in active/binding site regions, can provide a different perspective into the evolution of a protein than that obtained using the overall sequence or structure (Via, Ferre et al. 2000; Campbell, Gold et al. 2003). The study of such family-wide similarities is a crucial step in understanding protein function at the molecular level with practical implications in the field of drug design.

Methods for the detection of local (binding site) structural similarities have been developed in recent years primarily for the prediction of function from structure rather than the analysis of members of a given protein family. Such methods can be useful in providing hints about the function of a query protein particularly in cases where sequence-based methods as well as methods based on the overall fold of the protein are unable to give clues.

Methods for the prediction of function from structure via the detection of local structural similarities can be thought as composed of three interconnected parts: structure representation, search method and scoring procedure. The various methods developed over the years can differ in any of these components. Furthermore, some methods perform the search of the query structure against pre-defined structural templates. These templates can be automatically generated or curated, in which case their biological relevance is known (Najmanovich, Torrance et al. 2005).

Given the sets of atoms defining the clefts under comparison, the question that needs to be answered is what is the largest subset of atoms in both clefts in direct correspondence with

each other geometrically as well as chemically. This is a combinatorial optimization problem where, in principle, each possible set of atom correspondences might be a solution and the largest such set is the global solution. Graph theory offers a means to solve this problem via the detection of the maximal (largest) clique in an association graph. Further details on graph theory can be found elsewhere (Gross and Yellen 2004). In the present work, we use the standard algorithm of Bron and Kerbosh (Bron and Kerbosch 1973) for the detection of cliques.

Depending on the number of atoms being compared, the size of the association graph might make it practically unfeasible to detect the largest clique when considering all non-hydrogen binding site atoms. In order to overcome this difficulty we perform the graph matching in two stages.

In the first stage, an initial superimposition is performed via the detection of the largest clique in an association graph constructed using only C? atoms of identical residues in the two clefts. A maximum distance difference of 2.0 Å is used to create edges in the association graph, imposing an upper bound of the same magnitude in the coordinates root mean square distance (RMSD) of corresponding C? atoms.

Once the largest C? clique is obtained its transformation matrix and translation vector are used to superimpose all atoms in the two clefts using the least square method of Arun et al. (Arun, Huang et al. 1987) based on the singular value decomposition of the coordinates variance-covariance matrix.

In the second graph matching stage, all non-hydrogen atoms are used. Association graph nodes are created with the requirement that two atoms, one from each cleft, be of the same atom type as well as that their spatial distance be within 1.5 Å. This spatial distance constraint is used to decrease the size of the association graph and is the reason why the initial superimposition is performed. In the present work, we use eight atoms type classes (Sobolev, Wade et al. 1996; Sobolev, Sorokine et al. 1999) comprising the following classes: Hydrophilic, Acceptor, Donor, Hydrophobic, Aromatic, Neutral, Neutral-donor and Neutral-acceptor. Similar to the first stage, a maximum distance difference in defining association graph edges is used. This second threshold is set to 1.5 Å and again defines an upper bound of that magnitude in the RMSD between corresponding non-hydrogen cleft atoms.

We define a Local Tanimoto Score of structural similarity (LTS3D) to measure local structural similarity:

(Eq. 2)

where represents the size of the largest clique (number of similar atoms in either first or second stage) and is the sum of the total number of atoms in each cleft being compared. The same measure is used to calculate local structural similarity considering only C? atoms (LTS3D-C?) or all non-hydrogen atoms (LTS3D-all). The normalization describes the total number of C? atoms or the total number of non-hydrogen atoms in both clefts respectively.

The method for the detection of local structural similarities and the Tanimoto coefficients of

sequence and structural similarity developed here are able in conjunction to detect binding site sequence and structural similarities between members of the human cytosolic sulfotransferase family. Binding site sequence and structural comparisons uncovered similarities between members of two subfamilies of the human cytosolic sulfotransferase family not seen through overall sequence comparisons.

Arun, K. S., T. S. Huang, et al. (1987). "Least-Squares Fitting Of 2 3-D Point Sets." Ieee Transactions On Pattern Analysis And Machine Intelligence 9(5): 699-700.
Bron, C. and J. Kerbosch (1973). "Algorithm 457: finding all cliques of an undirected graph." Communications of the ACM 16(9): 575-577.
Campbell, S. J., N. D. Gold, et al. (2003). "Ligand binding: functional site location, similarity and docking." Current Opinion in Structural Biology 13(3): 389-395.
Gross, J. L. and J. Yellen (2004). Handbook of graph theory. Florida, CRC Press.
Najmanovich, R. J., J. W. Torrance, et al. (2005). "Prediction of protein function from structure: insights from methods for the detection of local structural similarities." Biotechniques 38(6): 847, 849, 851.
Sobolev, V., A. Sorokine, et al. (1999). "Automated analysis of interatomic contacts in proteins." Bioinformatics 15(4): 327-332.
Sobolev, V., R. C. Wade, et al. (1996). "Molecular docking using surface complementarity." Proteins-Structure Function and Genetics 25(1): 120-129.
Via, A., F. Ferre, et al. (2000). "Protein surface similarities: a survey of methods to describe and compare protein surfaces." Cellular and Molecular Life Sciences 57(13-14): 1970-1977.