Protein classification by surface analysis

Baldacci L.(1), Capozzi F.(2), Golfarelli M.(1), Lumini A.(1), Rizzi S.(1), Turano M.(2).

(1) D.E.I.S. – University of Bologna – Italy (2) C.I.R.M.M.P. – Consorzio Interuniversitario per le Risonanze Magnetiche Nucleari di Metalloproteine Paramagnetiche – Piazza Goidanich 60, 47023 Cesena (FC) Italy

Motivation

Divining the functional role of proteins in life processes, starting from their structural features, is the challenging purpose of structural biology. In this context, SCOP and CATH databases are the starting point for scientist aiming at discovering the relationships between function and molecular structure. The classification criteria adopted are based on the topological architecture of the molecule, thus the proteins are clustered by finding local folding motifs which are repeatedly encountered in the protein data bank. It is an assumption largely accepted that the overall structure deserves the appropriate distribution of chemical properties on the surface that the protein presents to its molecular target. Thanks to its "appearance" the protein could have the correct approach with the target or not. However, two proteins with similar structures may be divergent in their sequence, thus playing different functions. The structural classification as a tool for the individuation of a common physiological role is misleading in this case and a surface classification is more and more reputed necessary. Up to now, a successful strategy has not yet developed to reach such a general goal: the most exploited approach consists in adopting a surface patch already recognized to play a key role for a certain function and use it as a probe to explore the surfaces of all the proteins with known structure. This method, based on local features, will fail in all cases characterized by highly adaptable surfaces and when portions of surface far from the active site have a dominant allosteric effect on the functional region of the protein. In this paper we propose an original approach to classification of proteins based on their surface characteristics, which has the advantage of being not based on local surface features neither on already known functional meanings.

Methods

The global features we consider are electric potential, hydrophobicity and shape, thus our starting points are surface regions presenting homogeneous surface properties. Nevertheless, using a single area for classifying is not always effective, since the properties of the whole protein could be determined by a set of not necessarily neighboring regions. Given a DB of proteins, we search for complex surface patterns that can be roughly defined as a set of regions which have a given layout (i.e. relative placement on the protein surface). Proteins are classified according to the number of common patterns. Our approach to protein classification is composed by the following steps: (1) Clustering: this phase is aimed at defining areas with homogeneous surface features, that are identified using a region growing approach that is interrupted when the properties are no longer stable. The compact representation induced by clustering can be modeled by a completely connected graph whose nodes are labeled with the average value of the region features and whose edges are labeled with the information necessary to identify the relative positions of the two regions. (2) Mining: patterns are discovered using an unsupervised mining approach that works on the compact representation of the proteins. Starting from simple patterns (single regions), more complex ones are iteratively discovered by adding (possibly non-neighboring) regions such that the resulting pattern is present in at least a threshold number of proteins of the DB. Due to the nature of the problem, the presence of a given pattern in a protein is defined as the existence of a pattern similar to the

given one. The similarity function takes into account both the surface features of the regions and their relative placement. (3) Classification: it is obtained by clustering together those proteins that present a high number of common patterns. In order to avoid local minima a simulated annealing technique is adopted.

Results

Preliminary tests have been carried out showing that a known structural classification can be obtained using surface properties. We runned our techniques on a part of the SCOP DB. Our classification matches the first level of the SCOP one at the 88.5% while the percentage strongly decrease when the second level is considered. The main source of divergence between surface and structural classification is related to the wrong assumption that the same structures must share the same surface. SCOP, as well as CATH database, group proteins with similar fold in the same superfamily independently of the overall sequence homology. This means that two proteins or domains with similar structural fold may be characterized by a completely different set of chemical properties over all the surface, although a high sequence homology is retained in the inner sphere of the molecule to ensure the same fold. For this reason, the robustness of the algorithms adopted in our approach is being tested on a benchmark dataset of molecules with opportunely designed surface mutations introduced with the homology modelling technique.

Contact email: lbaldacci@deis.unibo.it