# Protein surface analysis: from the identification of 3D surface motifs to the construction of sequence patterns

A.Via, G.Cesareni, M.Helmer Citterich

Centro di Bioinformatica Molecolare, Dipartimento di Biologia, Università Tor Vergata, Roma

Genomics is providing a growing number of proteins of unknown function. In many cases sequence alignment and/or structural comparison methods allow the association of a function to a newly determined sequence or structure. However, in some cases, the newly determined sequence or structure is too distantly related to any protein of known function, therefore these methods are not always effective. In some of these cases the function of a protein can be inferred from the occurrence in its sequence or structure of a specific cluster of residues. One of the most assessed bioinformatic tools to assign a function to a protein is the PROSITE database of sequence motifs associated to a function.
A sizable fraction (~54%) of these motifs ("leaky" patterns) selects false positives and/or does not select all the proteins known to have the function associated to a motif.

The function of a protein is, in general, determined by the chemical properties of its surface. Residues that are conserved on the surface of distantly related proteins sharing a function are likely to be responsible or at least involved in that function, independently from their position in the protein sequences.
We consider distantly related proteins sharing a functional property and we perform a structural superposition of the regions involved in the function. By means of the 3D Motif procedure [1], we identify a cluster of residues conserved on the surface of all the aligned structures. We call this cluster "3D Surface Motif".
The first part of this work consists of the construction of a database of 3D Surface Motifs to be used to search into databases of protein structures. Structural genomic projects are likely to expand the number of structures of proteins of unknown function that are entered in the PDB. We also think that it is possible that not all the interesting sites on the surface of proteins of known function have already been successfully identified.
The second part of this work uses the 3D information extracted from the 3D Surface Motifs database to improve the selectivity of the PROSITE leaky patterns and to generate new sequence patterns associated to function. We consider residues, belonging to a 3D surface motif, that are co-linear in the sequences of the corresponding superposed structures. A program analyses the HSSP or PFAM alignments of the considered proteins in a region including the residues belonging to the 3D surface motif and identifies all the conserved aa in the columns of the alignments. These conserved residues form a potential sequence pattern to be tested in a sequence database search (e.g. swissprot). After the test, if necessary, the potential sequence pattern can be made more “flexible” and then tested again. The last step can be repeated until the motif is optimized. The scheme of the procedure is described in the Figure.

[1] De Rinaldis M. et al., (1998) J.Mol.Biol., 284, 1211-1221.