Large scale surface comparison for the identification of functional similarities in unrelated proteins

Fabrizio Ferrè, Gabriele Ausiello, Andreas Zanzoni, Manuela Helmer-Citterich

Centre for Molecular Bioinformatics, Dept. of Biology, University of Rome Tor Vergata, Rome (Italy)

We developed a systematic large-scale approach to identifying protein surface regions sharing shape and residue similarity. We used a new fast structural comparison algorithm (LSC: Local Structure Comparison) to exhaustively analyze a set of functionally annotated protein patches (1) with a larger collection of protein cavities. From a dataset of about 10.000 protein surface patches extracted from a non redundant list of PDB proteins (p-value=10^{-7}), we collected a grand total of 65910 matches among patch pairs that were stored in the SURFACE (2) database. The functional meaning of most of the matches could be confirmed by other established methods: the presence of the same PROSITE (3) and ELM (4) motifs in the sequence, the presence of the same ligand in the PDB (5) structure, similar GO (6) terms, common SWISS-PROT (7) keywords, sequence similarity, same SCOP (8) superfamily and E.C. (9) numbers. We noticed that the fraction of matches whose functional association can be confirmed by more methods sensibly decreases with the extension of the match (Figure 1).

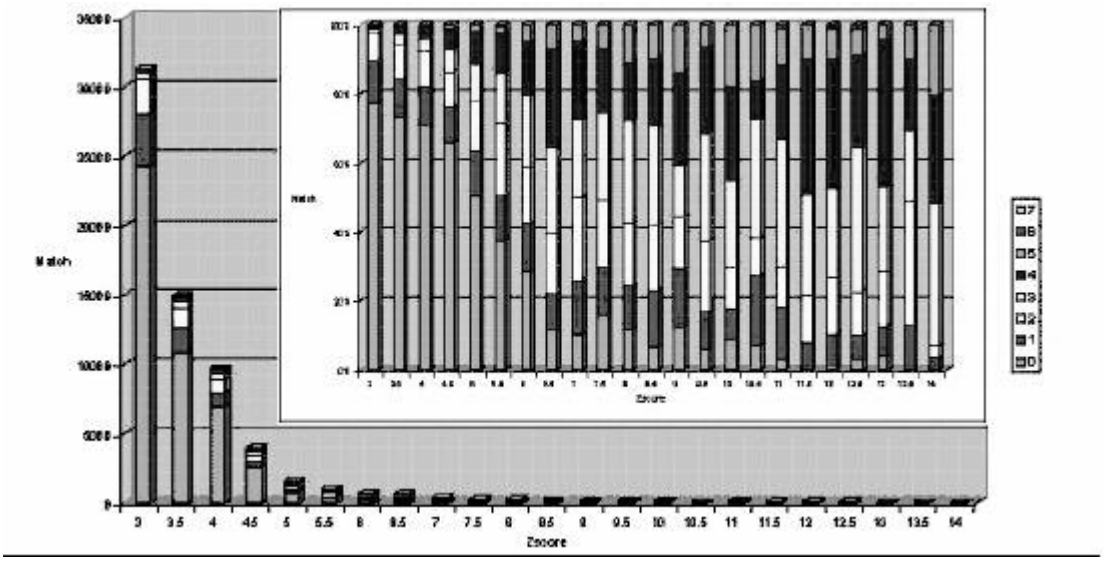

Fig 1 Validation of the matches by different methods. The graph shows the number of matches validated by 0, 1,..,7 methods in the different ranges of Z-score. The smaller graph shows the percentage of the matches confirmed by 0, 1, ..., 7 methods in the different Z-score ranges

So by considering only highly significant matches (zscore > 9), we could characterize a number of proteins solved in structural genomics projects and annotated as being of unknown function. The strategy we developed is a powerful addition to the list of available methods for the functional annotation of structures of unknown function.

References

[1] Laskowski, R.A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph*. **13**, 323-330, 307-308.

[2] Ferrè, F., Ausiello, G., Zanzoni, A., Helmer-Citterich, M. (2004). SURFACE: a database of protein surface regions for functional annotation. *Nucl.Acids Res.,* 32, 240-244.

[3] Hulo, N., Sigrist, C.J.A., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P. and Bairoch, A. (2004). Recent improvements to the PROSITE database. *Nucl. Acids. Res.,* 32, 134-137.

[4] Puntervoll, P., Linding, R., Gemuend, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D.M.A., Ausiello, G., Brannetti, B., Costantini, A., Ferrè, F., Maselli, V., Via, A., Cesareni, G., Diella, F., Superti-Furga, G., Wyrwicz, Ramu, C., McGuigan, C., Gudavalli, R., Letunic, I., Bork, P., Rychlewski, Kuester, B., Helmer-Citterich, M., Hunter, W.N., Aasland, R. and Gibson, T. (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucl. Acids Res.,* 31, 3625-3630.

[5] Bourne, P.E., Addess, K.J., Bluhm, W.F., Chen, L., Deshpande, N., Feng, Z., Fleri, W., Green, R., Merino-Ott, J.C., Townsend-Merino, W., Weissig, H., Westbrook, J. and Berman, H.M. (2004). The distribution and query systems of the RCSB Protein Data Bank. *Nucl. Acids. Res.,* 32, 223-225.

[6] The Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology. *Nature Genet.,* 25, 25-29.

[7] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.,* 31, 365-370.

[8] Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J.P., Chothia, C. and Murzin, A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acids. Res.,* 32, 226-229.

[9] Bairoch A. (2000). The ENZYME database in 2000. *Nucl. Acids Res.,* 28, 304-305.