Large scale protein function prediction tools

Raquel C. de Melo Minardi[1], François Artiguenave[1], Goran Neshich[2]

[1]Genoscope / Institut de Génomique / CEA
2 rue Gaston Crémieux CP5706 91057 Evry cedex France
[2]Structural Computational Biology Group / CNPTIA / EMBRAPA
Av. André Torsello, 209 – Unicamp, Barão Geraldo – Campinas, SP – Brazil
{raquelcm,artigue}@genoscope.cns.fr, neshich@embrapa.cbi.cnptia.br

Abstract. *In this project we intend to reinforce a recent collaboration established between two French and Brazilian bioinformatics laboratories for the development of a structural genomics approach to functional annotation of proteins. From initial works, we explored a new methodology for protein annotation based on modeled structural data. During this work, we identified that new tools will have to be designed for large scale protein function prediction. In this proposal, we would like to analyze the results of the proposed methodology with proteins of known function as well as analyze the impact of the use of different known techniques for each of its steps (homology modeling, cavity detection, binding / active site prediction and clustering according to sub-family specificity). We will possibly develop new algorithms in order to gain quality in prediction.*

Resumo. *Neste projeto, pretendemos reforçar uma recente colaboração entre dois laboratórios franceses e um brasileiro para o desenvolvimento de uma abordagem de genômica estrutural para anotação funcional de proteínas. Em trabalhos iniciais, exploramos uma nova metodologia para anotação de proteínas baseada em modelos estruturais. Durante este trabalho, identificamos que novas ferramentas são necessárias para anotação de proteínas em larga escala. Nesta proposta, pretendemos analisar os resultados desta metodologia com proteínas de função conhecidas bem como o impacto do uso de diferentes técnicas conhecidas utilizadas em cada uma de suas etapas (modelagem por homologia, detecção de cavidades, sítio de ligação / ativo e agrupamento de acordo com a especificidade de subfamílias). Possivelmente desenvolveremos novos algoritmos com o intuito de melhorar a qualidade das predições.*

1. Objective

The aim of this collaborative project is to develop a methodology and tools for large scale protein function prediction based on structural data. The major goals are:
1. Describe proteins and their surfaces in terms of features that discriminate between target areas which are the molecular interfaces through which proteins perform their function and those that are not.
2. Organize and maintain existing data relevant to broad spectrum of academic and commercial users and generate new data for purposes of identification of protein

function and specific loci responsible for functionality: catalytic site residues and interface forming residues.
3. Offer the newly acquired knowledge about protein substrates interactions to interested partners both from academic and industrial/commercial sector.
4. Fine tune applications for testing in some of the urgent problems in food industry, agro business and pharma, such as new antibiotics, new pesticides, new drugs and new strains.
Our main objective is to raise the level of our understanding in how two molecules (a protein/enzyme from one side and an inhibitor from the other) do communicate in order to engage in close contact which results in some kind of functional modification of (in this case) protein molecule, an important element of metabolic pathways which are responsible for specific disease or economically interesting traits.

2. Motivation

Both the Brazilian and French agriculture, livestock, food and pharmaceutical industries are challenged not only by the upcoming industries in Asia and established ones in North America and Australia, but also internally regarding demands for higher productivity and healthier product output with concomitant environmental protection. It is clear that this sector may add to its competitiveness by new product discovery, by process improvements, by betterment of used strains, and by the introduction of new biotechnological processes especially those referred to us as the “green routes” for the fermentative production of rare chemicals (harvesting chemicals instead of seeds/fruits). In order to promote a progress in this area, and targeting a list of some of the most important challenges facing research and development in both drug and food industries, we may cite problems which need to be addressed, such as: a) how to identify in silico function of proteins annotated in newly sequenced genomes, b) how to recognize and identify a network of interacting proteins within metabolic pathways, c) how to predict interactive part of the protein surface through which protein will communicate with other biomolecules and also perform its function and d) how to predict the effects of drugs and food (for both humans and animal organisms) by specifically keeping costs at lowest possible levels and increasing R&D process output.
This particular collaboration aims to contribute to those strategic goals by jointly reinforcing the competitiveness of the French and Brazilian pharma and agro (food) industries by enhancing research effort through joint work where complementary expertise of French and Brazilian partners will synergistically act in order to offer some improvement for the processes such as the analysis of protein sequences and structures as well as how specific amino acids mediate the interactions of these proteins with their substrates. Noteworthy is to mention that these tools may be used by scientists to collaboratively compile their own data on, for example, the specific anomalies associated with any given genomic locus and to further analyze their new data on the basis of already existing knowledge.

3. Partners

3.1. Genoscope contribution

The Genoscope (the French National Sequencing Center), has participated in the Human Genome Project and in international consortium in the domain of plant genomes and has revealed major events in the evolution of eukaryote genomes (vertebrates, ciliates, plants). The Genoscope has recently joined the CEA (Comissariat à l'Energie Atomique) and as so, has integrated in its main priorities, energy and environment issues. Consequently, the Genoscope is now enlarging the field of analysis of sequence data in extending the analysis to the experimental identification of biological functions. Internal projects are mainly in the domains of the environment and biodiversity and will open up perspectives for the development in the biotechnology industry and sustainable development. We started applying structural bioinformatics techniques in enzyme function predictions and are developing a methodology which is based on homology modeling [Tramontano et al. 2001], cavity analysis [Dundas et al. 2006], Hidden Markov Models (HHMs), conceptual clustering [Fisher 1987] and molecular docking of metabolites. We already have some interesting results that are being experimentally tested and show the applicability of such type of methods. For specific protein families, we intend to test the bioinformatics predictions using the high trough put screening platform setup at Genoscope, which will give experimental proofs of bioinformatics prediction and will allow to annotate new protein families.

3.2. Embrapa contribution

Embrapa will offer STING platform [Neshich et al., 2006], Interface, STING_DB and STING_RDB and also work in the development of new algorithms and tools to predict binding and catalytic sites. STING_DB as a source of structure/function descriptors which are to be used as vectors in 1D representation of 3D proteins. The Sting database operates with a collection of both publicly available data (e.g., PDB [Berman et al., 2000], HSSP [Schneider and Sander, 1996; Schneider et al., 1997], Prosite [Hulo et al., 2006], and UniProt [Apweiler et al., 2004]) and its proprietary protein sequence and structure (PSS) descriptors, such as geometric parameters (e.g., cavity, curvature), physic-chemical parameters (e.g., electrostatic potential), and conservation related parameters (e.g., SH2Qs, evolutionary pressure). The data consolidated and integrated into STING_RDB (STING relational database) make this database one of the most comprehensive databases available for analysis of protein sequence, structure and function and is updated on a weekly basis.

7. Planning

The project has well defined tasks, which ultimately lead to the following milestones:
1. Within one year, the clear guidelines for construction of an algorithm and building of user interface for automatic determination of catalytic site and interface forming amino acids, based on defined value intervals for sequence, structure, function and stability

descriptors, which by themselves have already been recorded within STING_DB or are still to be added as the new ones), would have been established.
2. In second year, the web based program for identification of catalytic site and interface forming AA from the sequence and structural data will be established.
3. In third year, the automatic targeting system for determination of function and function modification from the genomic data, based on both sequence and structure (homology models could be used if experimental structure information available is not sufficient [Tramontano et al., 2001]) will be established.
4. At the end of fourth year, the fully developed automatic system for identification of protein catalytic site, structural alignment and interface area prediction will be available to users, based on data analysis and data mining of the STING platform, which offers the largest collection of physicochemical parameters describing proteins structure, stability, function and interaction with other macromolecules.

References

[Apweiler et al. 2004] Apweiler, R., Bairoch, A., Wu, C., Barker, W., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M., Natale, D., O'Donovan, C., Redaschi, N., Yeh, L. (2004). UniProt: the universal protein knowledgebase. Nucleid Acids Res., 32:(Database issue):D115D119.
[Berman et al.] Berman, H., Bhat, T., Bourne, P., Feng, Z., Gilliland, G., Weissig, H., Westbrook, J. (2000). The Protein Data Bank and the challenge of structural genomics. Nat. Strcut. Biol., Suppl.9579.
[Dundas et al. 2006] Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., and Liang, J. (2006). CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acid Res., 34:W116W118.
[Fisher 1987] Fisher, D. (1987). Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2:139–172.
[Hulo et al. 2006] Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., LangendijkGenevaux, P., Pagni, M., Sigrist, C. (2006). The PROSITE database. Nucleic Acids Res., 34(Database issue):D227D230.
[Neshich et al. 2006] Neshich, G., Mazoni, I., Oliveira, S., Yamagishi, M., KuserFalcão, P., Borro, L., Morita, D., Souza, K., Almeida, G., Rodrigues, D., Jardine, J., Togawa, R., Mancini, A., Higa, R., Cruz, S., Vieira, F., Santos, E., Melo, R., and Santoro, M. (2006). The star sting server: a multiplatform environment for protein structure analysis. Genet. Mol. Res., 5(4):717–726.
[Schneider and Sander 1996] Schenider, R., and Sander, C. (1996). The HSSP database of protein structuresequence alignments. Nucleic Acids Res., 24(1):201205.
[Schneider et al. 1997] Schneider, R., Daruvar, A., and Sander, C. (1997). The HSSP database of protein structure sequence alignments.
[Tramontano et al. 2001] Tramontano, A., Leplae, R., and Morea, V. (2001). Analysis and assessment of comparative modeling predictions in casp4. Proteins, Suppl. 5:22–38.