Poster I-2

Protein function prediction based on 'reverse engineering' of the protein fold-recognition approach:examples from CASP6

Authors:
Michal Andrzej Kurowski (*Genesilico, IIMCB*)

Short Abstract: The most reliable approach to predict protein function is via its structure. Fold recognition - class predictions by themselves are often not good enough to enable function assignment, though. We developed an approach combining evolutionary protein families analysis with structural prediction. T0197 and T0209 CASP6 cases are described in here.

Long Abstract:
The most reliable approach to predict protein function is via its structure, e.g. by identifying a homolog in the Protein Data Bank, building a model and analyzing the functionally important residues. The accuracy of functional predictions can be greatly improved if purely structural considerations are combined with evolutionary analyzes, based on data obtained from both structural and sequence alignments. Such combined evolutionary analyzes are particularly useful for proteins with multiple domains, especially ones for which no closely related templates available, and reliable homology models cannot be obtained, and only approximate predictions can be made using fold recognition methods. However, sound biological functional assessment based on results of such work in many cases is unsatisfactory.

In this work we address such shortcomings by means of detailed evolutionary analysis based on data obtained from both structural and sequence alignments. Our method relies on the assumption that domains are the units of protein structure and function and attempts to evaluate the likelihood of compatibility between the sequence of the target protein with functionally annotated domains. The approach used in here is analogous to the the one described in [1]. Detailed study is beyond the scope of this document and is further described in a PhD thesis of the author.

The starting point for the functional prediction is the collection of results obtained from the GeneSilico metaserver [2]. First, potential domains are assigned to the target sequence based on its comparison with the libraries of alignments (e.g. the PFAM database for sequences, ASTRAL for structures). For the sequence fragments with no domain matches, potential domain boundaries are identified based on prediction of disordered regions (putative inter-domain linkers) and secondary structure. As a result, a series of sequence fragments corresponding to potential domains is generated and analyzed as hypotheses that can be either accepted or rejected. For each sequence fragment, a series of analysis carried out. The fold-recognition (FR) results (target-template alignments reported by FR methods) are obtained, and for each template the sequence profile is generated. The FR results are analyzed for consistency by structural comparisons, and the match between the target sequence and the template profile is assessed by maximum likelihood methods. In the cases,

where a potential domain can be reliably assigned to the template, the functional assignment is taken from the annotation of the closest homolog, based on the analysis of the evolutionary tree. This analysis is repeated for all putative domains and as a result, a collection of functional predictions is reported with associated maximum likelihood scores. Hypothetical trees generated are subsequently processed by tree comparison methods (CONSEL package is used). The final functional inference has to be made by the user.

Our 'reverse-engineering' approach we will be showcased based on two examples of difficult structural and/or functional assignments as it was performed during CASP6 contest.

T0197 (1XKC) is a protein from Pyrococcus furiosus. Its function has not been described. It belongs to the nucleotide cyclases CYTH family which is extensively characterized [3] by the Aravaind group as a adenylyl cyclases group CyaB unrelated to most nucleotide cyclases. None of it's members were characterized structurally until the CASP6 target was published. Similarly, no structural homologs were reviled at the time. Our approach shows significant similarity to mRNA triphosphatases which perform capping function in yeasts. This triphosphatase family does have a structural representative (1d8h) with a strikingly similar topology of a barrel built from antiparrallel sheets and beta-alpha-beta motif at the N-terminus.Both enzymes require divalent cation for their activity. Although CYTH and mRNA triphosphatases families have completely different phyletic distribution they can be reliably shown as derived from a common ancestor. T0197 adenylyl cyclase function and 1D8H triphosphatase activity mechanisms can be further fruitfully studied structurally.

T0209 (1XQB) is a hypothetical protein from Haemophilus Influenzae with unknown function. Globally, it shows no similarity to any known structure and predicting its structure and function proved to be very difficult. According to the domain search , the region spanning the N-terminus and the center of T0209 confines a UPF0066 domain. The C-terminal region does not show any detectable sequence similarity to known proteins (and it's fold has been described as new when the structure was solved). However, the evolutionary pattern of the UPF0066 region can be matched with the profile of the Acyl-CoA thioesterase family with a significant score. Although structural similarity of the UPF0066 member to the thioesterases type II Hot Dog topology [4] is only partial and no consensus was apparent among FR results, several models with similar structures could be found. Members of the UPF0066 domain are not annotated functionally in biological databases. Moreover, thioesterase type II-like proteins are known to exhibit many diverse functions. Therefore the detailed functional assignment of T0209 based on the evolutionary analysis is very difficult. However, based on our family assignments, T0209 can be implicated in the metabolism of fatty acids.

We conclude that structural predictions in even most difficult cases can provide significant insight into evolutionary relationships between protein families and guide functional predictions. The structure prediction methods often fail to produce models of high quality, but even imperfect predictions reported by FR servers that provide a skeleton-like structure can prove useful in the process of family assignment and resulting functional prediction. The 'reverse-engineering' of the FR analysis is often able to enhance the functional predictions resulting only from the 'final' structural models . The downside is that because of evolutionary patterns in the nature, the process of sequence-structure mapping cannot be easily automated. Thus, new methods must be developed that go beyond the currently existing 'consensus' approaches.

[1] M. Kurowski, A. Bhagwat, G. Papaj, and J. Bujnicki. Phylogenomic identi-fication of five new human homologs of the DNA repair enzyme AlkB. BMC Genomics, 4(1):48, 2003.

[2] M. Kurowski and J. Bujnicki. GeneSilico protein structure prediction meta-server. Nucleic Acids Res, 31(13):3305 7, 2003.

[3] L. Iyer and L. Aravind. The catalytic domains of thiamine triphosphatase and CyaB-like adenylyl cyclase define a novel superfamily of domains that bind organic phosphates. BMC Genomics, 3(1):33, 2002.

[4] S. Dillon and A. Bateman. The Hotdog fold: wrapping up a superfamily of thioesterases and dehydratases. BMC Bioinformatics, 5:109, 2004.