MS10-O4 From protein sequence to function and structure with BAR+

Giuseppe Profiti[1], Rita Casadio[1], Francesco Aggazio[1], Pier Luigi Martelli[1], Piero Fariselli[1]

1. Bologna Biocomputing Group, Bologna Computational Biology Network, University of Bologna, Italy

email: giuseppe.profiti2@unibo.it

We introduced a web server that allows functional and structural annotation of protein sequences. A previous version of our method was already described and validated, the Bologna Annotation Resource PLUS(+). BAR+ is a non hierarchical clustering method relying on a comparative large-scale genome analysis. The method relies on a non hierarchical clustering procedure characterized by a stringent metric that ensures a reliable transfer of features within clusters. The set includes 13,495,736 protein sequences that derive also from 988 whole genomes. BAR+ is constructed by performing an all-against-all pairwise alignment from all protein sequences available (collected from the entire UniProt). Each protein is then taken as a node and a graph is built allowing links among nodes only when the following similarity constrains are found among two proteins: their sequence identity (SI) is ≥ 40% and the extent of the overlap after alignment (Coverage, CO) is ≥ 90%. By this clusters are simply the connected components of the graph. 70% of the whole data set of sequences fall into 913,962 clusters. Well annotated sequences are characterized by all the functional and structural annotations derived from UniProt entries. These include GO, PFAM, PDB and SCOP mapping (when available). Ligands are also listed when present in their PDB file/s. When a well annotated sequence falls into a cluster, it inherits the annotation/s that characterize the cluster. GO and PFAM features in the clusters are validated by computing a P-value. With this procedure, also distantly related homologs can inherit function and structure in a validated manner. This procedure increases the level of annotation when compared to that of Uniprot. In BAR+ when PDB templates are present within a cluster (with or without their SCOP classification), profile HMMs are computed on the basis of sequence to structure alignment and are cluster-associated (Cluster-HMM). A library of 10,858 HMMs is available for aligning even distantly related sequences to a given PDB template/s. BAR+ is available at http://bar.biocomp.unibo.it/bar2.0. A recent new improvement relies on community detection techniques that allow the identification of groups of proteins relative to a specific ligand. By this, clusters are subdivided into smaller sets of closely related sequences, enhanching the specificity of the annotation in terms of different ligand binding to the same putative template.

Keywords: Protein structure prediction; protein function prediction

MS10-O5 New protein main-chain conformational descriptors on the validation and improvement of automatic protein model building

Joana Pereira[1], Victor Lamzin[1]

1. European Molecular Biology Laboratory (EMBL), c/o DESY, Notkestrasse 85, Hamburg 22607, Germany

email: joana.pereira@embl-hamburg.de

During the process of protein automated model building, ARP/wARP [1] represents the electron density map as a set of free atoms without any chemical identity and, through the use of density and distance checks, searches for free atoms on possible Cα positions that could be forming a peptide unit. If two putative peptide units share a free atom, they are considered to be a dipeptide, the conformation of which is then evaluated against a two-parameter Ramachandran-like plot. We have found such an evaluation has proven to be very powerful with high-resolution data; however, more than two conformational degrees of freedom are required to properly account for experimental errors and to build models at 3.0 Å or lower resolution.

To address the problem, we utilised distance-geometry-based methods, which are often used in the NMR structure solution. We expanded on the premise that molecular conformation, represented by the relative three-dimensional location of atoms, can be calculated when the distances between all atoms in the molecule are known. We identified three independent parameters that describe dipeptide conformation; thereby enabling the separation of dipeptides corresponding to different secondary structural elements, from dipeptides in randomly generated conformation. By comparing the three-dimensional distribution of these parameters with the parameters calculated for random dipeptides, we were able to compute a scoring function for the evaluation of the likelihood of a dipeptide to be in a plausible conformation. Dipeptides with a score close to 1 are likely correct, while those with a score close to 0 should not be accepted.

Using this approach, we have been able to evaluate the quality of protein chain fragments built using ARP/wARP at different data resolutions. We have found that, at lower resolution, the constructed chain fragments have more dipeptide units which are likely in an incorrect conformation. We now plan to incorporate the newly developed method into the model building process and expect that this will increase the accuracy and completeness of the automatically constructed protein structures. Additionally, an overall score can also be calculated for the entire protein structure, providing the user with a general measure of the quality of the model.

[1] Langer G, Cohen SX, Lamzin VS, Perrakis A. *Nature Protocols*, 2008, **3(7)**, 1171-1179

Keywords: automated model building, peptide conformation, structure validation, software, ARP/wARP