Bioinformatics: From Protein Sequences to Structure/Function Relationships

Gilbert Deléage
Institute for Biology and Chemistry of Proteins
Pôle BioInformatique Lyonnais
7, Passage du Vercors
69367 Lyon cedex 07
France
http://pbil.ibcp.fr

This talk will present the main developments made by our group in “structural bioinformatics” for several years. The talk covers several topics:

1) Methodologies to predict the secondary structure of protein and a discussion about prediction accuracy
2) Integration into ANTHEPROT/MPSA software and NPS@ webware.
3) Biological applications of secondary structure prediction related to virology
4) Presentation of a strategy to perform comparative homology modeling at low identity level (PROCSS)
5) Presentation of geno3D: a server to perform automatic molecular modeling on the Web

1) The methods to predict the secondary structure of proteins can be divided into three main classes:
   - Statistical methods (Chou and Fasman and GOR’s methods)
   - Similarity methods (Levin, SIMPA, SOPM methods)
   - Family based methods (neural methods as PHD and empirically optimized methods as SOPMA)

An historical review providing the advantages and the pitfalls of these methods will be given. Current methods have a success rate (as measured by Q3 value) usually above 72% but the accuracy is highly dependent upon several parameters that have hindered the objective comparison of methods. For example, the reference databank for comparing prediction and observation is a crucial point. Indeed, the criteria to include or to exclude a protein from that databank can be somehow biased. We do know that PDB is highly redundant (more than 20000 structures but only 2000 chains exhibiting not more than 25% identity after binary alignment). This level of identity has regularly changed during the course of the “ideal method”. At the very beginning of the story (80’s) all protein were included. In the 90’s, only protein chains sharing less than 50% identity were taken into account. More recently, authors logically admit to take a 25% threshold as the upper limit to include protein into the reference databank since above 30% identity the best approach consists in the homology modelling of the 3D structure rather than simply predict its secondary structure.

2) Integration into software/webware suite of programs
All methodologies must be usable by the final user (i.e. biologists). However, very few methods are compatible at the level of their I/O format, the sequence analysis requires a logical flow between methods the choice of which remains largely manual. For these reasons and to keep the GUI as convivial as possible, our group is developing since 90’s software such as ANTHEPROT and MPSA. These software put together in a single program most of the methods useful to perform protein sequence analysis. Moreover, all the graphical windows can be manipulated in independent ways allowing an efficient and
customized analysis. Finally these programs work in a client/server mode allowing to make remote queries onto central servers (maintaining up-to-date databases). In this model, the user may perform time consuming jobs on a server (which could be a GRID of machine in a near future), download result into the local software and interactively make some choices (programs parameters or databanks extraction) and finally launched again jobs with new set of parameters or sequences. The following figure illustrates the architecture and the interaction of methods which are connected together as Unix commands are via the “pipe”.

This service performs 2530 analyses per day (25% from France, 25% from the rest of Europe, 25% of USA and 25% for the rest of the world).

3) Biological applications
The main utility of predicted secondary structures arises from the fact that structures are more conserved than sequences. During the talk several examples will illustrate the use of secondary structure prediction to help the biologist to better understand the protein at the structure functions relationship level.

4) Comparative molecular modeling
A server to perform comparative modeling of protein has been developed. If the identity level detected by PSI-BLAST is below 30%, the system makes use of the predicted secondary structure to identify a putative template by comparing the secondary structure compatibility. The possibility to use several templates is also offered, for instance to use different parts of the sequences (modular modeling). At the result level, a fit between all models is offered as well as a set of geometrical, energetic and statistical analyses.

The results are sent by email as a link onto the whole archive containing all the web pages generated by geno3D. A typical example of the geno3D output is given in the figure below. In order to assess the validity of the modeling process at a large scale, models have been generated for all PDB proteins (found by BLAST) and sharing between 10 and 30% identity. The super computer facilities of the IN2P3 center (382 bi-processor farm) has permitted to achieve the generation of 5390 models in about 8000 hours of CPU time (cluster of 382 CPUs). Results show that at least 65% of the models were correct (ie superimposable with a RMSD less than 3 Å). The distribution of the
Quantitative analysis of models generated for proteins sharing only 10 to 35% sequence identity with the template. The model analysis has been performed by using the Dali program. The number of protein models is plotted as a function of the Dali Z-score (A) and RMSD (B) between experimental structure and the model.

Bibliography