Structural conservation in single domain proteins: implications for homology modeling

G.D'Alfonso, A. Tramontano and A. Lahm

IRBM P.Angeletti, Dept. Computational Biology and Chemistry, via Pontina km 30.600, Pomezia (Rome)

Large-scale sequencing projects are widening the gap between the known protein universe and the fraction for which structural information has been experimentally obtained. Through the application of homology (comparative) modeling and more general structure prediction techniques, this gap can however be narrowed, providing indirect structural information for a considerable number of proteins. Within this perspective, homology (comparative) modeling will gain in importance, as will the use of models derived by this technique.

Although a perfect scenario would see a researcher browsing through the PDB (or a related structural) database and selecting his/her most appropriate structural template, a more realistic view is to assume internet-servers with pre-generated models that can be downloaded. Ideally, the server would offer alternative models, if available, and provide the appropriate information needed for model selection. This requires a procedure, preferably automatic, that identifies alternative non-canonical template structures within a family of related proteins such as a SCOP superfamily.

Analyses such as that described here can be a first step towards such a procedure.

Here we discuss how well a sequence alignment, the most common starting point for generating a model, reflects the structural conservation between homologous proteins and we show that sequence information is able to direct construction of acceptable models as far as the structural core is concerned. Here we have discussed how well the backbone of the core region (conserved secondary structure elements) can be modelled. The distribution of the expected average rmsd error against sequence identity is similar to the trend observed in comparisons of protein structures (Chothia & Lesk , 1986). Since our approach uses only sequence information, the distribution should be regarded as an estimate of the lower limit of the expected accuracy in comparative modeling which should improve when additional data, such as secondary structure predictions, are included. An encouraging feature of protein structures is the rather well conserved local conformation of single loop regions without insertions and deletions. The situation is unfortunately different for stem regions around insertions/deletions. Even when omitting the two residues directly flanking the insertion/deletion point, the observed average error is at least twice as high compared to loop segments with no insertions/deletions, regardless of the selected stem length. Together with global shifts in the position and orientation of secondary structure elements, and, consequently, also of the intervening loop elements, this increases the overall error when all loop segments are superimposed simultaneously. Loop modeling, and, in particular, loops with insertions/deletions, therefore remains the major obstacle in model generation.

Chothia, C., and Lesk, A.M. (1986) The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823-826.