Predicting fold from sequence: A new algorithm

John B.C. Findlay

University of Leeds

One of the landmark successes in bioinformatics has been the development of sequence-based algorithms for detecting homologues and compiling evolutionary and structural relationships. This approach relies on a degree of sequence conservation but begins to break down when the sequence similarities descend into what is commonly called the "twilight zone". Yet we know that there are very many more sequences than there are 3-dimensional structural folds. This piece of work was designed, therefore, to examine whether there is a mechanism for identifying proteins which have some form of ancestral relationship but which because of evolutionary diversity have lost the sequence identifiers but still may retain the structural fold.

The method relied on the alignment and analysis of the 3-dimensional structures of members of a family, selected on the basis of their diversity. Residue positions which participated in packing contacts were then utilised as the basis of a method for identifying members of the family from sequence databases. Several "sparsities" could be used, 10% representing those positions which participated in the highest number of contacts, the 20% sparsity included the next 10% and so on. These residue positions constituted the "SIGNATURE" for the protein. They allowed both amino acid variability as seen in the sequences included in the hit list, and positional variability reflecting the addition/deletion of segments between the key residues.

The method was surprisingly successful at identifying proteins belonging to a variety of families, even when their sequences were not apparently homologous either by inspection or using sequence-based search algorithms. It now forms the basis of a system for identifying the fold of a protein based on the closeness of match with the signatures of the major folds so far classified.

References

Blades, M.J., Ison, J.C., Ranasinghe, R. and Findlay, J.B.C. (2005). Automatic generation and evaluation of sparse protein signatures for families of protein structural domains. *Protein Science*, **14**, 13-23.