4. PREDICTING FUNCTION FROM SEQUENCE

James C. Whisstock[1] *and Arthur M. Lesk*[2]

[1]Victorian Bioinformatics Consortium, Monash University, Melbourne, Australia; [2]Cambridge Institute of Medical Research, University of Cambridge, UK.

The sequence of a genome contains the plans of the possible life of an organism, but implementation of genetic information depends on the functions of the proteins and nucleic acids that it encodes. Many individual proteins of known sequence and structure present challenges to understanding their function. In particular, a number of genes responsible for diseases have been identified but their specific functions are unknown. Whole-genome sequencing projects are a major source of proteins of unknown function. Annotation of a genome involves assignment of functions to gene products, in most cases on the basis of amino acid sequence alone. Three-dimensional structure can aid the assignment of function, motivating the challenge of structural genomics projects to make structural information available for novel uncharacterised proteins. Structure-based identification of homologues often succeeds where sequence-alone-based methods fail, because in many cases evolution retains the folding pattern long after sequence similarity becomes undetectable. Nevertheless, prediction of protein function from sequence and structure is a difficult problem, because homologous proteins often have different functions. Many methods of function prediction rely on identifying similarity in sequence and/or structure between a protein of unknown function and one or more well-understood proteins. Alternative methods include inferring conservation patterns in members of a functionally uncharacterised family for which many sequences and structures are known. However, these inferences are tenuous. Such methods provide reasonable guesses at function, but are far from foolproof. The development of whole-organism investigations permits other approaches to function prediction when the data are available. These include the use of protein–protein interaction patterns, and correlations between occurrences of related proteins in different organisms, as indicators of functional properties. Even if it is possible to ascribe a particular function to a gene product, the protein may have multiple functions. An underlying problem is that function is in many cases an ill-defined concept. Here we discuss the state of the art in function prediction and describe some of the underlying difficulties and successes.

10.1071/SRB03Ab4