Using phylogenetic information in the detection of correlated amino acid substitutions

Horner D.(1) Pirovano W.(1) Pesole G.(1)

(1) Department of Molecular Biosciences and Biotechnology, University of Milan, Italy

Motivation

Much effort has been devoted to the detection, from multiple sequence alignments of homologous proteins, of pairs or groups of amino acid positions that evolve in a non-independent – or compensatory – manner. It is expected that such clusters of positions might either tend either to be proximal in the mature, folded protein, or to be involved in similar aspects of protein function. Several such methods have been shown to be reasonably effective in the detection of intra-protein contacts. However, all of the most successful published algorithms rely on pairwise comparisons between aligned sequences. We wished to investigate whether evolutionary information (the topology and branchlengths of the phylogenetic tree describing relationships between the sequences under study) can allow an improvement in the prediction of intra protein contacts from correlated substitutions.

Methods

We have developed and implemented a series of algorithms that use phylogenetic trees and reconstructed ancestral sequences in the identification of pairs of sites undergoing compensatory substitutions. Our method uses a substitution matrix-based correlation coefficient similar to that developed by Olmea, Pazos and Valencia (ref) and previously shown to be the most effective published algorithm (ref). In contrast to the method of Olmea, which uses all pairwise comparisons between aligned sequences, our method examines homologous positions only along branches on the tree connecting ancestral and descendent sequences. We thus examine only substitutions inferred to have occurred during the evolutionary history of the sequences under examination and can incorporate phylogenetic information such as the length of branches on the tree into the estimation of correlated sites. Furthermore, our approach – unlike pairwise comparison methods - explicitly considers the phylogenetic relationships between sequences and minimizes apparent correlation derived from the phylogenetic structure underlying the data.

Results

For thirty large test datasets, our method returns a performance comparable with those of the most effective previously published algorithms in the prediction of contact residues. Mean distances between pairs predicted by our algorithm are lower than those between pairs predicted by other methods. We show that consideration of phylogenetic structure of multiple sequence alignment can be a powerful tool in the detection of correlated substitutions.

Contact email: david.horner@unimi.it