Protein contact prediction with correlated mutation analysis using mixed physiochemical constraints

Horner D (1), Pesole G (2)

1) Dipartimento di Scienze Biomolecolari e Biotecnologie, Universita degli Studi di Milano
2) Dipartimento di Biochimica e Biologia Molecolare, Universita di Bari

Motivation

While models of sequence evolution used in many bioinformatic and evolutionary approaches assume that individual sites evolve independently of one-another, it has long been expected that in certain situations, for example amino acids that constitute structurally or functionally important contacts in proteins, should undergo a form of correlated, or compensatory evolution. This assumption forms the basis of many algorithms designed to predict protein structural contacts from alignments of homologous protein sequences. While such approaches have shown considerable potential, levels of accuracy typically fall short of those required to use correlated mutation analysis in structure prediction approaches. Interestingly, very few Correlated Mutation Analysis (CMA) algorithms incorporate phylogenetic information (the tree describing evolutionary relationships between the sequences under analysis). Furthermore, while several CMA algorithms incorporate models of biophysical properties of different amino acids, these typically assume that a single biophysical parameter governs all potentially correlated substitutions for a given pair of sites. We present here an updated version of our algorithm which incorporates phylogenetic information in the detection of pairs of sites undergoing potentially correlated evolution. In the current implementation, the biophysical parameters used to quantify the degree of correlation between pairs of sites are allowed to vary over the tree.Furthermore, we introduce an improved MonteCarlo data simulation procedure that allows rapid evaluation of the significance of results obtained.

Methods

Phyogenetic trees describing the relationships between homologs of proteins of known structure were estimated by standard methods. Ancestral sequences (at internal nodes on phylogenetic trees) were reconstructed under the maximum likelihood criterion. The magnitude of correlation between pairs of sites isevaluated by calculating correlation coefficients based on changes in biophysical characteristics implied by substitutions inferred on the relevant phylogenetic tree. The current implementation of this protocol allows the simultaneous calculation of correlation scores based on different biophysical parameters and the optimization of the correlation criterion in different parts of the tree. The significance of calculated correlation scores is evaluated using a MonteCarlo simulation in which a null distribution of correlation scores for every pair of sites is calculated by repeatedy independently distributing observed substitutions over the tree according to relative branch lengths.

Results

Allowing the biophysical parameter governing correlated substitutions between a pair of sitesto vary over the phylogenetic tree increases the sensitivity of CMA with our algorithm. The currently implemented measure of significance substantially decreases the number of false positive predictions with repect to a previous version which did not incorporate lengths of branches on the phylogenetic tree.

Contact email: david.horner@unimi.it