Coevolution in protein families: a functional correlation study.

Arianna Bertolino

July 10, 2011

Abstract

During the course of evolution, the set of proteins derived from a common ancestral protein (i.e. a protein family) accumulates random mutations and insertions/deletions, displaying a mean sequence similarity of 20-40%. Despite this sequence heterogeneity, the three-dimentional structure is conserved across species. This interesting phenomenon suggests that we are facing a kind of constrained evolution where the set of possible moves in the space of sequences is such that the structural properties have to be preserved.

The signature of such constrained evolution reflects in a complex residue-residue correlation pattern emerging from the multiple sequence alignment (MSA). What kind of structural information can we infer from a protein family MSA?

Answering to this question in full generality is a formidable challenge. In this study we restrict our goal to the prediction of the set of residues in physical contacts (i.e. the set of residue pairs at distance lower than $8\mathring{A}$ in the native 3-D structure).

Local correlation based analysis (e.g. mutual information) are attractive measures because they explicitly show the degree of statistical association between residues, but they have very important shortcomings that affect their predictive power.

A first problem is that correlation may result from direct coupling effects, but also from indirect effects via intermediate residues. Only the direct ones are informative about structural constraints. In order to disentangle direct and indirect correlations we use a global inference approach based on the maximum-entropy principle. The idea is to infer a global statistical model for the whole amino acid sequence of the protein domain under study. This distribution has to be coherent to the empirical data, i.e. the frequency counts. Within mean field approximation the inference problem can be solved in a single step, without using any iterative scheme.

We tested the efficiency and the prediction capacity of the method for intraprotein contacts on 131 proteins: we achieved the highest performance compared to all other methods.

Concerning interprotein interactions, a specificity analysis on two component systems, the most important signaling system in bacteria, has

been performed: as a result we have been able to predict efficiently inter-pathways cross-talk and orphans interaction pathways already known experimentally.

A further problem is that phylogentic effects between related species generate non-functional correlations in the data-set. We have tested several tree-based and sequence-based corrections. Sequence-based corrections are performing better and faster than other methods: they significantly increase the predictive power both of mutual information-based approaches and of our global inference model.