Inference of Co-Evolving Site Pairs: An Excellent Predictor of Contact residue Pairs in Protein 3D Structures

Sanzo Miyazawa[1]
sanzo.miyazawa@gmail.com

[1] 6-5-607 Miyanodai, Sakura, Chiba 285-0857, Japan

Keywords: selective constraint on amino acids, co-substitution, compensatory substitution, coevolution between sites, phylogenetic analysis, maximum likelihood estimation, partial correlation coefficients, Gaussian graphical model

Residue-residue interactions that fold a protein into a unique three-dimensional structure and make it play a specific function impose structural and functional constraints in varying degrees on each residue site. Selective constraints on residue sites are recorded in amino acid orders in homologous sequences and also in the evolutionary trace of amino acid substitutions. A challenge is to extract direct dependences between residue sites by removing phylogenetic correlations and indirect dependences through other residues within a protein or even through other molecules. Rapid growth of protein families with unknown folds requires an accurate *de novo* prediction method for protein structure. Recent attempts of disentangling direct from indirect dependences of amino acid types between residue positions in multiple sequence alignments have revealed that inferred residue-residue proximities can be sufficient information to predict a protein fold without the use of known three-dimensional structures [1,2,3]. Here, we propose an alternative method of inferring coevolving site pairs from concurrent and compensatory substitutions between sites in each branch of a phylogenetic tree [4]. First, branch lengths of the Pfam phylogenetic tree are optimized as well as other parameters by maximizing a likelihood of the tree in a mechanistic codon substitution model. Substitution probability and physico-chemical changes (volume, charge, hydrogen-bonding capability and others) accompanied by substitutions at each site in each branch of a phylogenetic tree are estimated with the likelihood of each substitution, and their direct correlations between sites are used to detect concurrent and compensatory substitutions. In order to remove phylogenetic correlations and to extract direct dependences between sites, partial correlation coefficients of the characteristic changes along branches between sites, in which linear multiple dependences on feature vectors at other sites are removed, are calculated and used to rank coevolving site pairs. Accuracy of contact prediction based on the present coevolution score is better than that [2] achieved by a maximum entropy model of protein sequences for 15 protein families taken from the Pfam release 26.0. Besides, this excellent accuracy indicates that compensatory substitutions are significant in protein evolution.

[1] Morcos, F., Pagnani, A. L. B., Bertolinod, A., Marks, D. S., Sander, C., Zecchina, R. Onuchic, J. N., Hwa, T., and Weigt, M., Direct-coupling analysis of residue coevolution captures native contacts across many protein families, *Proc. Natl. Acad. Sci. USA*, 108:E1293-E1301, 2011.

[2] Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C., Protein 3D structure computed from evolutionary sequence variation, *PLoS One*, 6:e28766, 2011.

[3] Taylor, W. R., and Sadowski, M. I., Structural Constraints on the covariance matrix derived from multiple aligned protein sequences, *PLoS One*, 6:e28265, 2011.

[4] Miyazawa, S., Prediction of contact residue pairs based on co-substitution between sites in protein structures, *PLos One*, 8:e54252, 2013.