

15. A novel pair-to-pair substitution matrix improves contact prediction in protein cores

Milana Frenkel-Morgenstern*,1, Eran Eyal*,2, Vladimir Sobolev2, Shmuel Pietrokovski1

*1 Department of Plant Science 2 Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, 76100, Israel, * equally contributed to this work*

We developed the P2PConPred method for predicting direct physical contacts between residues inside proteins from multiple alignments. Our method uses a novel pair-to-pair substitution matrix (P2PMAT) derived from protein multiple alignments and structures. It performs particularly well in predicting core contacts, and significantly better than other contact prediction methods.

Prediction of protein residue-residue contacts from sequence information is an important and difficult problem in structural bioinformatics. Most methods use correlated mutation analysis to detect such contacts. We developed a new approach (P2PConPred) for intra-protein contact prediction, which is oriented for detecting direct physical contacts. Our method uses a novel pair-to-pair substitution matrix (P2PMAT) derived from accurate protein multiple sequence alignments with an available representative structure. The P2PMAT matrix integrates the probabilities of contacting and non-contacting protein sites. Incorporating evolutionary (sequence) conservation of residues with information regarding correlated substitutions is natural and effective in our P2PMAT matrix. Our P2PConPred method is most sensitive for contact prediction in the protein cores. Core residues are effectively identified from sequence information alone by means of a predicted surface accessibility of proteins. Our method improves protein core contact prediction by 1.25 and 1.6 fold over the contact prediction method of Gobel et. al. [1] and that of Singer et. al. [2], respectively. Combining our approach with other approaches for calculating correlated mutations is expected to be beneficial. The basic approach we developed can also be naturally applied to other protein analysis problems, such as fold recognition and subtle sequence similarity detection.

References

1. Gobel, U., Sander, C., Schneider, R., Valencia, A., Correlated mutations and residue contacts in proteins. *Proteins*, 1994. 18: p. 309-317.
2. Singer, M., Vriend G., and Bywater R., Prediction of protein residue contacts with a PDB-derived likelihood matrix. *Prot Eng*, 2002. 15: p. 721-725.