IMPROVED COMPUTATIONAL METHODS OF PROTEIN SEQUENCE ALIGNMENT, MODEL SELECTION AND TERTIARY STRUCTURE PREDICTION

Xin Deng

Dr. Jianlin Cheng, Dissertation Supervisor

ABSTRACT

Protein sequence and profile alignment has been used essentially in most bioinformatics tasks such as protein structure modeling, function prediction, and phylogenetic analysis. We designed a new algorithm MSACompro to incorporate predicted secondary structure, relative solvent accessibility, and residue-residue contact information into multiple protein sequence alignment. Our experiments showed that it improved multiple sequence alignment accuracy over most existing methods without using the structural information and performed comparably to the method using structural features and additional homologous sequences by slightly lower scores. We also developed HHpacom, a new profile-profile pairwise alignment by integrating secondary structure, solvent accessibility, torsion angle and inferred residue pair coupling information. The evaluation showed that the secondary structure, relative solvent accessibility and torsion angle information significantly improved the alignment accuracy in comparison with the state of the art methods HHsearch and HHsuite. The evolutionary constraint information did help in some cases, especially the alignments of the proteins which are of short lengths, typically 100 to 500 residues.

Protein Model selection is also a key step in protein tertiary structure prediction. We developed two SVM model quality assessment methods taking query-template alignment as input. The assessment results illustrated that this could help improve the model selection, protein structure prediction and many other bioinformatics problems.

Moreover, we also developed a protein tertiary structure prediction pipeline, of which many components were built in our group's MULTICOM system. The MULTICOM performed well in the CASP10 (Critical Assessment of Techniques for Protein Structure Prediction) competition.