英文講題：Protein Sequence Analysis and its Applications on Function Classification

中文講題：蛋白質序列分析及其在蛋白質功能分類之應用

Abstract

Automated function annotation is a major goal of post-genomic era with tremendous amount of protein sequences in the databases. Prediction of protein function is crucial for proteomic analysis, genome annotation, and drug discovery. Determination of function or structure using experimental approaches is time-consuming; thus, computational approaches become highly desirable.

We proposed two protein subcellular localization prediction methods, PSL101 and PSLDoc. PSL101 combines a structural homology approach and a support vector machine model, in which compartment-specific biological features derived from bacterial translocation pathways are incorporated. PSLDoc uses a probabilistic latent semantic analysis on gapped-dipeptides of various distances, where evolutionary information from position specific scoring matrix (PSSM) is utilized. Our methods achieve 93% in overall accuracy for Gram-negative bacteria, and compared favorably to the state-of-the-art results by 7.4% on a benchmark dataset having low homology to the training set. Experiment results demonstrate that both biological features derived from translocation pathways and feature reduction by document classification techniques can lead to a significant improvement in the prediction performance. Moreover, the proposed biological features and gapped-dipeptide signatures are interpretable and can be applied in advanced studies and experiment designs.

For RNA-binding site prediction, we propose another method, RNAProB, which incorporates a new smoothed PSSM encoding scheme in a support vector machine model. The proposed smoothed PSSM encoding considers correlation and dependency from neighboring residues for each amino acid in a protein sequence. Experiment results show that smoothed PSSM encoding significantly enhances the prediction performance, especially for sensitivity. Our method performs better than the state-of-the-art systems by 4.90%~6.83%, 7.05%~26.90%, 0.88%~5.33%, and 0.10~0.23 in terms of overall accuracy, sensitivity, specificity, and Matthew's correlation coefficient, respectively. This also supports our assumption that smoothed PSSM encoding can better resolve the ambiguity in discriminating between interacting and non-interacting residues by modeling the dependency from surrounding residues.

Because of the generality of the proposed methods, they can be extended to other research topics in the future. Moreover, the information from predicted localization and structure of proteins can be used collectively to assist biologists in both inferring protein function and finding suitable drug targets. Therefore, we believe that our work can contribute to scientific discoveries on a high-throughput basis.