Public Abstract
First Name:Jilong
Middle Name:
Last Name:Li
Adviser's First Name:Jianlin
Adviser's Last Name:Cheng
Co-Adviser's First Name:
Co-Adviser's Last Name:
Graduation Term:SP 2016
Department:Computer Science
Degree:PhD
Title:COMPUTATIONAL METHODS FOR PROTEIN STRUCTURE PREDICTION AND NEXT-GENERATION SEQUENCING DATA ANALYSIS

With the wide application of next-generation sequencing technologies, the number of protein sequences is increasing exponentially. However, only a tiny portion of proteins have experimentally verified structures. The huge protein sequence-structure gap could be reduced by computational methods including template-based modeling and template-free modeling. The major challenges in template-based modeling include integrating multiple templates and building conformations for gapped regions. Chapter 2 describes a stochastic point cloud sampling method for multi-template protein model generation. The stochastic sampling and simulated annealing protocol in the method has the capability to improve the global quality and reduce atom clashes in protein models.

Two popular approaches for improving protein structure prediction include enlarging the sampling space of template-based modeling and integrating template-based modeling with template-free modeling when no good templates or only partial templates can be found for a target protein. Chapters 3 and 4 introduce a large-scale conformation sampling and evaluation system for protein structure prediction. The system increases the diversity of template sampling, sequence alignment sampling and model generation, combines template-based modeling and template-free modeling in order to generate a pool of models of good quality, and applies an array of protein model quality assessment methods to evaluate and rank the predicted models.

Next-generation sequencing of RNAs (RNA-Seq) generates hundreds of millions of short reads. Analyzing these reads is increasingly being used to foster novel discovery in biomedical research. Chapter 5 describes a bioinformatics pipeline for RNA-Seq data analysis, which converts gigabytes of raw RNA-Seq data into kilobytes of valuable biological knowledge through a five-step data mining and knowledge discovery process. The experiments show that the data mining process successfully produced valuable biological knowledge and reduced the size of the initial data over a thousand-fold.