Genetic Algorithm Based Data Mining for Bioinformatics

Abstract

This paper is based on our research project in the application of data discovery and analysis techniques for bioinformatics. This research has included the following four related parts:

First, improving data presentation quality based on data mining method. Chapter 4 ***Method of Online Analytical Processing on Nucleotide Sequences Database*** gives a detailed discussion on this part. This chapter proposed a new method of Online Analytical Processing on EMBL Nucleotide Sequences Database. This scheme is used to automatically restore flat file data into relational database, which in turn is converted into OLAP's data marts. The data marts greatly improved the quality and speed of analysis. We believe that this method is a powerful and flexible tool and can be seen as successful application of data mining in molecule biology.

Second, improving the traditional data mining algorithms. Chapter 5 ***K-Means Clustering Based on Genetic Algorithm*** presents such an improvement. We compared our method with the traditional K-Means method and the clustering method based on simple genetic algorithm. The comparison has proven that our method achieves a better result than the other two. The drawback of this method is a comparably lower speed in clustering.

Third, application of heuristic data mining approaches for bioinformatics. Chapter 6 ***Clustering of Amino Acid Sequences Based on Genetic Algorithm*** describes our heuristic approach to the clustering of amino acid sequences using Genetic Algorithm. The method evolves a population of medoids in a quasi-evolutionary manner, and gradually improves the fitness of the population by measuring the fitness through a function for evaluating clustering quality. This method combines Genetic Algorithm, K-Medoids method, Dynamic Programming and other new theories in Biology. Experiments have proven that our method has better performance than K-Medoids clustering method in returning more satisfying result.

Fourth, improving the existing data mining algorithms in bioinformatics. Chapter 7 ***Protein Conformation Prediction Based on Parallel Genetic Algorithm*** illustrates our solution on the topic. For the prediction of protein structure, we proposed the solution based on parallel genetic algorithm. This solution has modified the existing algorithm by omitting the process of simulated annealing, while compensating this modification by parallel programming to maintain the results quality. Experiments have demonstrated that the new solution has higher processing speed and same results quality. This solution also underlines the necessity of introducing the idea of parallel programming into the study of bioinformatics.

Keywords:

Bioinformatics Data Mining Genetic Algorithm Nucleotide Sequences Database Protein Database Online Analytical Processing Clustering of Amino Acid Sequences Protein Conformation Prediction Parallel Genetic Algorithm