Chapter 7.29
A Bayesian Framework for Improving Clustering Accuracy of Protein Sequences Based on Association Rules

Peng-Yeng Yin
National Chi Nan University, Taiwan

Shyong-Jian Shyu
Ming Chuan University, Taiwan

Guan-Shieng Huang
National Chi Nan University, Taiwan

Shuang-Te Liao
Ming Chuan University, Taiwan

ABSTRACT

With the advent of new sequencing technology for biological data, the number of sequenced proteins stored in public databases has become an explosion. The structural, functional, and phylogenetic analyses of proteins would benefit from exploring databases by using data mining techniques. Clustering algorithms can assign proteins into clusters such that proteins in the same cluster are more similar in homology than those in different clusters. This procedure not only simplifies the analysis task but also enhances the accuracy of the results. Most of the existing protein-clustering algorithms compute the similarity between proteins based on one-to-one pairwise sequence alignment instead of multiple sequences alignment; the latter is prohibited due to expensive computation. Hence the accuracy of the clustering result is deteriorated. Further, the traditional clustering methods are ad-hoc and the resulting clustering often converges to local optima. This chapter presents a Bayesian framework for improving clustering accuracy of protein sequences based on association rules. The experimental results manifest that the proposed framework can significantly improve the performance of traditional clustering methods.

INTRODUCTION

One of the central problems of bioinformatics is to predict structural, functional, and phylogenetic features of proteins. A protein can be viewed as a sequence of amino acids with 20 letters (which is called the primary structure). The explosive growth of protein databases has made it possible to cluster proteins with similar properties into a family in order to understand their structural, functional, and phylogenetic relationships. For example, there are 181,821 protein sequences in the Swiss-Prot database (release 47.1) and 1,748,002 sequences in its supplement TrEMBL database (release 30.1) up to May 24, 2005. According to the secondary structural content and organization, proteins were originally classified into four classes: α, β, $\alpha+\beta$, and α/β (Levitt & Chothia, 1976). Several others (including multi-domain, membrane and cell surface, and small proteins) have been added in the SCOP database (Lo Conte, Brenner, Hubbard, Chothia, & Murzin, 2002). Family is a group of proteins that share more than 50% identities when aligned, the SCOP database (release 1.67) reports 2630 families.

Pairwise comparisons between sequences provide good predictions of the biological similarity for related sequences. Alignment algorithms such as the Smith-Waterman algorithm and the Needleman-Wunsch algorithm and their variants are proved to be useful. Substitution matrices like PAMs and BLOSUMs are designed so that one can detect the similarity even between distant sequences. However, the statistical tests for distant homologous sequences are not usually significant (Hubbard, Lesk, & Tramontano, 1996). Pairwise alignment fails to represent shared similarities among three or more sequences because it leaves the problem of how to represent the similarities between the first and the third sequences after the first two sequences have been aligned. It is suggested in many literatures that multiple sequence alignment should be a better choice. While this sounds reasonable, it causes some problems we address here. The most critical issue is the time efficiency. The natural extension of the dynamic programming algorithm from the pairwise alignment to the multiple alignment requires exponential time (Carrillo & Lipman, 1988), and many problems related to finding the multiple alignment are known to be NP-hard (Wang & Jiang, 1994). The second issue is that calculating a distance matrix by pairwise-alignment algorithm is fundamental. ClustalW (Thompson, Higgins, & Gibson, 1994) is one of the most popular softwares for multiple-alignment problems. It implements the so-called progressive method, a heuristic that combines the sub-alignments into a big one under the guidance of a phylogenetic tree. In fact, the tree is built from a pre-computed distance matrix using pairwise alignment.

Many protein clustering techniques exist for sorting the proteins but the resulting clustering could be of low accuracy due to two reasons. First, these clustering techniques are conducted according to homology similarity, thus a preprocessing of sequence alignment should be applied to construct a homology proximity matrix (or similarity matrix). As we have mentioned, applying multiple sequence alignment among all proteins in a large data set is prohibited because of expensive computation. Instead, an all-against-all pairwise alignment is adopted for saving computation time but it may cause deterioration in accuracy. Second, most of the traditional clustering techniques, such as hierarchical merging, iterative partitioning, and graph-based clustering, often converge to local optima and are not established on statistical inference basis (Jain, Murty, & Flynn, 1999).

This chapter proposes a Bayesian framework for improving clustering accuracy of protein sequences based on association rules. With the initial clustering result obtained by using a traditional method based on the distance matrix, the strong association rules of protein subsequences for each cluster can be generated. These rules satisfying both minimum support and minimum confidence can serve as features to assign proteins to new