

**Additional Paper for “Biological Data Mining”**Paper reference

I. I. Artamonova, G. Frishman, M.S. Gelfand and D. Frishman. **Mining sequence annotation databanks for association patterns**. Vol. 21 Suppl. 3 2005, pages iii49-iii57, doi:10.1093/bioinformatics/bti1206

Abstract

Nowadays, the size of sequence databases is increasing exponentially. There is also proportional increase in the number of protein sequences. In order to make sense of these amino acid sequences, it is necessary to annotate them (annotation: comment/explanation). However owing to their huge size, it is infeasible to manually annotate the protein sequences. One solution is automating the annotation process; but the imperfect methods used to develop these tools cause errors in annotations.

To improve the quality of automatic annotation tools, it is imperative to investigate better algorithms. This paper uses the approach of association rule mining to capture exceptions and anomalies in automated annotations. It also discusses the results which demonstrate the improvement in the annotation process.

Discussion

Association rules are used to identify a set of features that are statistically related in the underlying data. Mathematically, an association rule is represented as  $X \Rightarrow Y$ , where  $X$  and  $Y$  are the different features of the data set. The rule implies that if a database entry has feature  $X$ , then it is *likely* to have feature  $Y$  as well. For any given database, these rules are characterized by the following attributes: *support* which is defined as the number of times a given rule is satisfied over the entire database and *strength* which is defined as the probability with which this rule is satisfied over that database.

The main assumption while applying the association rule mining technique to the annotation process is that, if a given association rule is satisfied with high support and high strength, then it is highly likely that the rule reflects some biological regularity and hence could be used in the annotation process. If a rule has very high strength, then either it gives us exceptions/anomalies which are mainly due to biological reasons; but it is also likely that it captures errors in annotation process. Hence it is advisable to focus on rules with very high strength since in either case we discover something new – either a biological exception or an error.

This technique is applied to the data extracted from the Swiss-Prot database, which is a high quality database due to manual annotation and the PEDANT genome database that contains significantly higher rate of errors, since it is automatically annotated. The rules are extracted using *Apriori* algorithm, which finds frequent item sets in a given database by combining candidate item sets at every stage. An important aspect of this algorithm is that it is significantly faster than other existing algorithms for similar purpose. Details of *Apriori* algorithm are beyond the scope of this paper.

The results obtained by this technique support the initial assumption that, exceptions obtained by applying the association rules to annotations generated by automated process, generally point to errors in the annotation process. Another advantage of this technique is that it can be applied to large automatically generated annotation data, for which no efficient method of verification is available.

This paper is similar to the one presented in the class in the sense that both deal with automation processes to analyze the massive dataset of protein sequences. The similarity further extends to the fact that both papers try to identify a set of rules that would characterize the protein database. However, the difference lies in the problems tackled by both papers. Whereas the previous paper addresses the problem of improving accuracy of prediction methods used for classification of proteins, this paper discusses an approach to increase the accuracy of automated annotation process.