

Additional paper for “Biological Data Mining”Paper Reference:

Irena I.Artamonova, Goar Frishman, Mikhail S. Gelfand and Dmitriy Frishman, Mining sequence annotation databanks for association patterns, Bioinformatics Vol. 21 Suppl. 3 2005, pages iii49-iii57

Abstract:

Using experimentally verified information from scientific literature is a time-consuming and creative process, so manually annotation of all proteins in this way is impossible. Most of the annotations in the databanks are based on the similarity with known annotations but this method has some limitations and in some cases it has error. These errors propagate in the databases and transfer these transitive information to unrelated proteins. There are large-scale software systems do this transitive annotation using fixed recognition thresholds to assign a function and structure to the proteins. But the problem of the automatic transitive annotations is the development of intelligent systems to improve the quality of annotations. There are different approaches to this problem; one can be based on the consistency of assigned features to find the erroneous or missed ones. Another approach for intelligent filtering and improvement of annotations is a rule-based approach. Rule-based techniques detect common patterns, rules or anomalies and they can automatically predict the annotations and find the errors. The approach presented in this paper involves automatic learning of rules from a highly reliable database to improve the annotation in the same database or in the automatically generated database using data mining algorithms. It is called Association Rules to Improve Annotation, ARIA.

Discussion:

This paper presents a method to find the annotation errors using the association rule mining technique. Association rules relate a multi-entry database with a finite number of features for each entry in the database to a different feature(LHS=>RHS) The rule is that all the entries in the database have probably this new feature. The characteristics of a rule can be its coverage, number of entries in the database satisfying the LHS, its support, the number of entries satisfying both LHS and RHS, its strength, the probability that an entry satisfies RHS given it satisfies the LHS.

But the main assumption in the application of association rule mining technique to improve the annotation is that if the database annotation satisfy a rule ‘A and B imply C’ with a high support and strength then such a rule reflects some biological regularity or maybe a peculiarity of the annotation process. For the strengths very close to 1, the rule has a minor number of exceptions.

The target here is to improve the automatic annotation in the PEDANT Genome Database. The annotation produced by PEDANT has a large amount of errors because they have not been verified manually. To evaluate their approach, they analyze the association rules generated by Swiss-Prot database and then apply the method to the annotation generated by PEDANT and present both manual and computational results. They extract item sets from both Swiss-Prot database and PEDANT genome database and they used them as input data to extract rules using *Apriori* algorithm.

They demonstrate that exception from strong rules are flagged as potential annotation errors. They found out that statistical properties of association rules for automatically generated PEDANT annotation and manually Swiss-Prot annotation are similar.

Like Wieser et al. (2004) paper, this paper is also try to present a method to automatically annotate the large number of protein sequences in the databases. Wieser et al tries to detect annotation errors based on a decision tree approach using a data-mining algorithm and they want to improve the precision of automated annotation but this paper tries to detect anomalies in annotations based on association rule mining and they want to find the sources of errors in the automatic software systems to improve the quality of generated annotation.

ARIA shows that exceptions from strong association rules are the sources of annotation errors but Xanthippe system (Wieser et al., 2004) says they propagate in the entire database, not in individual protein families. ARIA uses association rules mining that is more efficient than C4.5 data mining algorithm used in Xanthippe system. andARIA is much faster than Xanthippe. As these two methods have different applications, ARIA for large automatically generated dataset with a gold standard of correctness but Xanthippe for a highly organized database but no such standard available for it, the rules generated in them are different.