Poster K-11

Annotation errors and their correction by association rule mining

Authors:
Artamonova I.I. (*Institute for Bioinformatics, GSF - National Research Center for Environment and Health*)
Frishman G. (*Institute for Bioinformatics, GSF - National Research Center for Environment and Health*)
Gelfand M.S. (*Institute of Information Transmission Problems RAS*)
Frishman D. (*Department of Genome Oriented Bioinformatics, Technische Universität Munchen*)

Short Abstract: The Association Rules mining was applied to identification of possible errors in protein annotations. The errors were identified as exceptions from strong rules found by the algorithm. Suspicious features were marked for manual curation. Tests demonstrated that most exceptions indeed were caused by annotation errors.

Long Abstract:

The protein sequence data grows exponentially making it impossible to annotate manually all known proteins. Automatic annotation is technically efficient, but at the same time notoriously error-prone. The development of intelligent systems aimed at improving the quality of automatically generated annotation is one of the most difficult challenges in bioinformatics. One approach to error correction is to check the consistency of assigned features and then to mark suspicious features for manual inspection or to add those features that could have been missed.

We address this problem by applying the Association Rule mining technique to large sequence annotation databases. This technique amounts to the discovery of association relationships or correlations among a set of items and originates from the analysis of the data on market baskets. Usually, an association rule is formulated in the form Left-Hand-Side (LHS) implies Right-Hand-Side (RHS) and is interpreted as “database entries that satisfy LHS are likely to satisfy RHS”.

The main measures of the rule quality are coverage, support and strength. The coverage of a rule is the number of database entries that satisfy the LHS of the rule. The support is the number of entries satisfying both the LHS and the RHS. The strength is the fraction of entries that satisfy LHS and RHS among the entries satisfying the LHS.

Application of the association rule mining to improving the annotation is based on the following assumption: if a rule “A&B imply C” has high support and strength, then it reflects some biological regularity or maybe a peculiarity of the annotation process. If the strength is very close, but not equal, to 1, then the rule has a minor number of exceptions. While in some cases such exceptions may reflect biological reality, it is plausible that a significant fraction of them are actual errors in annotation. Hence our strategy was to find rules of high strength filter them, identify proteins that are exceptions from such rules, mark the features from the left side and add the right side feature of the rule to the annotation of such exceptional proteins.

Our main goal is to improve the automatic annotation in the PEDANT Genome Database

containing pre-computed information resulting from bioinformatics analyses of publicly available genomes. Its main mission is to provide robust and up-to-date annotation of the vast majority of protein sequences which have not been subjected to manual curation by experts in high-quality databases such as Swiss-Prot. Due to well-known pitfalls of large-scale automatic sequence annotation, PEDANT assignments contain a large amount of potentially erroneous information. Therefore we decided to first test our method on the Swiss-Prot database. For our analysis we selected the most formalized non-overlapping fields of a standard Swiss-Prot entry, such as the protein length, the highest-level taxon of the protein origin, InterPro domains, keywords and features from the feature table. Fragments and hypothetical proteins were not considered.

We calculated association rules for Swiss-Prot attributes and analyzed their statistics. A prominent feature of the rule strength distribution is the presence of two distinct peaks corresponding to very weak and very strong rules. Most weak rules originate from random combinations of frequent features. The other extreme of this distribution is constituted by very strong rules with strength values in the range between 0.97 and 1.0. According to our main assumption, the exceptions from the latter rules might be annotation errors.

The protein annotations in Swiss-Prot are evolving in time as various improvements result in constant modification of the annotations in semiannual releases. We have analyzed the annotation corrections for proteins that are exceptions from our strong rules. In the majority of cases the modified annotations satisfy our rules. That is, if a re-annotated protein is an exception from a strong association rule, then the modification involves deletion of a feature from the LHS or addition of the RHS feature. The fraction of re-annotated proteins among exceptions from the strong rules is significantly higher than the average fraction of re-annotated entries in the whole database. In addition, a limited random sample of exceptions to strong rules was checked manually. A significant fraction of annotations were found to contain errors that had lead to these exceptions. The total rate of actual annotation errors among the exceptions from the strong rules was estimated as 42%.

Additional filtration of rule exceptions based on the similarity of a targeted protein entry with other entries confirming the rule allowed us to reduce the number of considered protein annotations several-fold simultaneously increasing the fraction of correctly identified errors to 60%. Thus our method identifies several thousands of feature combinations in about 5% of all protein entries of Swiss-Prot, the majority of which are indeed errors of annotation.

Next, we have applied our technique to the PEDANT database. This resulted in more than 73000 rules, compared to ~13000 rules for Swiss-Prot. Based on the manual verification of a random sample of exceptions from strong rules, we estimated the fraction of correctly identified errors as 68%. The breakdown of association rules generated from PEDANT in terms of their constituent items indicates that most of the errors that can be corrected are related to gene functional roles. While the Swiss-Prot errors were usually caused by under-annotation due to its conservative approach, the automatically generated PEDANT annotation suffered from over-annotation.

Thus, we have developed a general methodology for improving the quality of biological data based on the notion that exceptions from strong association rules derived from annotation items often point to errors.

This work is conducted in the framework of the BioSapiens project funded by the European Commission FP6 Programme, under the thematic area "Life Sciences, Genomics and Biotechnology for Health", contract number LHSG-CT-2003-503265.