False Annotations of Proteins: Automatic Detection via Keyword-Based Clustering

Noam Kaplan* and Michal Linial

Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem 91904, Israel
* To whom correspondence should be addressed

Received line

ABSTRACT
Computational protein annotation methods occasionally introduce errors. False-positive (FP) errors are annotations that are mistakenly associated with a protein. Such false annotations introduce errors that may spread into databases through similarity with other proteins. We present a protein-clustering method that enables automatic separation of FP from true-positive hits. The method is based on the combination of each protein's annotations. Using a test set of all PROSITE signatures that are marked as FPs, we show that the method successfully separates FPs in 70% of the cases. Automatic detection of FPs may greatly facilitate the manual validation process and increase annotation sensitivity.
Contact: kaplann@cc.huji.ac.il; michall@cc.huji.ac.il

INTRODUCTION

Computational protein annotation methods are widely used. A wide variety of annotation methods exists, many of which rely on some kind of scoring. Typically, when testing whether a protein should be given a certain annotation, a score threshold is set, and proteins that score higher than the threshold are given the annotation. Obviously, some annotation mistakes may occur. Such mistakes can be divided into false positives (FPs) and false negatives (FNs). FPs (or false hits) are annotations that were mistakenly assigned to a protein (type I error). FNs (or misses) are annotations that should have been assigned to a protein but were not (type II error). Adjustment of score thresholds allows tradeoff between these two types of mistakes. FPs annotations are considered to be of graver consequence than FNs. This is partly due to the fact that introduction of a false positive annotation into a protein database may cause other proteins to become incorrectly annotated on the basis of sequence similarity (Linial 2003; Gilks et al. 2002). A systematic evaluation of the source of false annotations that already contaminated current databases was reported (Iliopoulos et al. 2003). Several automatic systems such as PEDANT (Frishman et al. 2003) and GeneQuiz (Andrade et al. 1999) were introduced with the goal of matching the performance of human experts. Still, over interpretation, FN errors, typographic mistakes and the domain-based transitivity pitfall limit the use of such fully automatic systems for inferring protein function.

Due to the importance of minimizing the amount of false annotations and maintaining highly reliable protein databases, three methods are generally used to avoid false annotations. The first method is manual validation of the annotation of each protein, which creates a serious bottleneck in the addition of new proteins and annotations to the database. The second method is using high score thresholds, thus lowering the rate of FPs but also increasing the rate of FNs. The third method is requirement for hits from different detection methods, eliminating advantages that are unique to some methods. Thus it would be beneficial to develop means by which FP annotations could be detected automatically.

Here we present such a method that uses clustering of protein functional groups to separate true positives (TP) from FPs automatically. Our method is based on the following notions: (a) protein annotations represent biological properties; (b) protein functional groups share specific combinations of biological properties, essentially constituting "property clusters"; (c) if two proteins have very different combinations of annotations, they are unlikely to share a single functional annotation (a high chance that one of them was given this annotation incorrectly). These notions are not obvious, but were shown to correctly indicate false annotations in some individual cases tested manually using the graphical annotation-analysis tool of PANDORA (Kaplan et al. 2003). Still they were not tested on a wide scale, and were not applied as automatic methods.

Using these ideas, the method attempts to separate a group of proteins into "property clusters", by introducing a measure that quantifies the similarity between the annotation combinations of two proteins. According to our basic notions, these clusters are likely to be in accordance with false and true hits.

We tested our method on the PROSITE protein signature database (Sigrist et al. 2002). The database consists of 1189 protein signatures (essentially annotations) that were assigned to a protein database. PROSITE annotation of proteins is manually validated, stating for each protein hit whether the annotation is a TP or a FP. Out of this set of 1189 signatures, we

chose a subset of all signatures that have both true and false hits, and this served as our test set. Altogether 327 such signatures were tested. For each of the signatures, the method examined the set of proteins that were assigned the signature. We called the separation successful only if at any step of the clustering process all the TPs were clustered together without any FPs. We have not counted cases of partial success.

METHOD

We created a database that includes all proteins from Swissprot 40.28 (114,033 proteins). The database also included annotation of these proteins by GO (Camon et al. 2003), Swissprot (Boeckmann et al. 2003) and InterPro (Apweilier et al. 2000). We tested 327 sets of PROSITE protein annotations (signatures). Given a set P of all proteins that were given a certain PROSITE annotation, and that there are both FPs and TPs in P, we would like to separate the set into subsets, so that one of the subsets will include all TPs and no FPs (leaving one or more subsets containing the FPs).

Annotation-based clustering is used to detect these subsets. We define an annotation as a binary property assigned to a protein (each protein may or may not have a "hit"). The clustering works in the following way: between each two proteins we define a similarity score that tries to quantify how much do the two proteins have in common from a biological perspective. The score between two proteins *a* and *b* is defined as:

$$score(a,b) = -\sum_{C} \log(p(i))$$

Where *C* is the set of annotations that both a and b share, *i* is the current annotation, and *p(i)* is the frequency of *i* in the database. This score uses the following logic: if two proteins share an annotation, they are biologically similar in some manner. The more annotations these proteins share, the more reason we have to believe that they are similar biologically. However, two proteins sharing an annotation like "Enzyme" (that appears 45991 times in our database) should receive a worse similarity score than two proteins that share a much uncommon annotation like "Heat Shock Protein" (that appears 832 times). This is taken into account by using –log(*p(i)*).

We define the similarity score between two clusters as the arithmetic average of scores of all inter-cluster protein pairs.

Starting with clusters of 1 protein each, at each clustering step the two clusters that have the highest similarity score are merged. At each step the contents of the clusters are evaluated, and if all TP proteins appear in one cluster without any FPs, we say that the clustering process successfully separated the TPs from the FPs.

Annotation source and the number of annotation for each (in parenthesis) are: SwissProt version 40.28 (865), InterPro version 5.2 (5551), GO as of July 2002 (5229), PROSITE version 17.5 (1189).

RESULTS AND DISCUSSION

Out of 327 sets of proteins that share a PROSITE signature, the method showed successful separation (as defined in Methods) in 228 sets, i.e. 69.7% of the cases. The average size of the protein sets was 156.1 and the median 76. The average and median true positive rates (TP/(TP+FP)) of the sets were 0.88 and 0.93 respectively. These general statistics about the test set indicate that the sets were large enough and had a high enough amount of TPs so that the chance of random success would be minimal.

Introduction of FP annotations into protein databases can be harmful. It has been shown that once a mistaken annotation is introduced to a database, it often transfers to other proteins that are sequentially similar causing a propagation of false annotation (Linial 2003). Due to the importance of keeping high-quality databases, either the proteins are manually checked one by one or the annotation detection sensitivity is reduced in order to minimize FPs. The error rate and the limited sensitivity of assigning structural annotations using PSI-BLAST (Muller et al. 1999) or SAM-T98 (Karplus et al. 1998) and methodologies based on HMM and SVM had been reported (Karchin et al. 2002). Naturally the process of manual validation of the annotation of protein databases is extremely time-consuming. Automatic detection of false annotations greatly facilitates the task of manual validation of annotation, and allows using lower thresholds when trying to detect protein signatures, therefore allowing higher method sensitivity.

Based on the notion that protein functional groups share specific combinations of annotations, we have introduced a method that by separating a set of proteins into their "property clusters" shows successful separation of incorrectly annotated proteins from correctly annotated proteins in 69.7% of the tested cases.

Although the separation success rate is not perfect, it should be noted that to the best of our knowledge there are currently no automatic methods that may be used to eliminate false annotations without compromising the sensitivity of detection. Furthermore, the method is not base on protein sequence. This said, some points regarding the presented method should be addressed:

(i) The clustering process is based on annotations. Therefore, it may be difficult to apply this method to proteins that are poorly annotated. Still, these cases should be relatively rare: Nearly 99% of the proteins contain at least one annotation, more than 95% of the proteins are associated with 2 annotations or more. The average number of annotations per protein is 10.9 and the median is 10 (including ENZYME and SCOP annotations). Note that both the amount and richness of annotation is constantly increasing. Furthermore, ability to detect false annotations automatically may allow an increase in the sensitivity of current methods, thereby allowing more extensive annotation of proteins. Introducing into the similarity score quantitative protein

properties that are easily determined and show some correlation with function (i.e., the protein length, its pI, number of transmembrane domains) is expected to refine and improve the sensitivity of our method.

(ii) The method is aimed to test only functional annotations. Notion c (see introduction) states that "if two proteins have very different combinations of annotations, they are unlikely to share a single functional annotation". This is not true for annotations that are only weakly related to function, as these annotations do not normally constitute "property clusters" (e.g. annotations like "mutation", "disease", "complete genome").

(iii) We call a clustering process successful if it managed at any step to separate the false annotations (as defined in Methods). However, when applying the method this step must be somehow determined, either by the process itself or in advance. Some preliminary results (not shown) indicate that there is high correlation between the "correct" separation step and intrinsic properties of the clustering process. Because the clustering is a process, we can use a measure of “time” to describe the progression of this process. We define the time at a given step of the process to be the score of the last two clusters that were merged. Using this definition of time, we define a cluster’s lifetime as the difference in time between a cluster’s creation and the cluster’s merging with another cluster. This measure of lifetime enables us to detect when a cluster reaches stability (reflected by a long lifetime), and was found to be in high correlation with the correct separation step.

(iv) At the step in which the clustering process shows successful separation there may be several clusters, one of which contains the TPs. However it is impossible to know which of the clusters is the one that contains the TPs. This problem can be solved easily if we know in advance some of the proteins that are TPs (for example when adding new proteins to a database that has been already validated). Also, in most cases the largest cluster is the one that contains the TPs (naturally this would depend on the ratio of TPs, which was shown to be high in our test set).

Separation of a protein set into functional groups by automatic means may be used not only for detection of FP annotations. Computational and experimental methods such as large-scale proteomics often result in large lists of proteins that require expert manual inspection in order to separate those into functional groups. Annotation-based clustering can facilitate the interpretation of such lists.

ACKNOWLEDGEMENTS

We thank Menachem Fromer for his support and useful suggestions. We thank the ProtoNet team for their constant support. This work is partially supported by The Sudarsky Center for Computational Biology and the Israeli Ministry of Defense.

REFERENCES

Andrade, M.A., Brown, N.P., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C., Franchini, A., Tamames, J., Valencia, A., Ouzounis, C. et al. (1999) Automated genome sequence analysis and annotation. Bioinformatics, 15, 391-412.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D. et al. (2000) InterPro--an integrated documentation resource for protein families, domains and functional sites. Bioinformatics, 16, 1145-1150.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res., 31, 365-370.

Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A. et al. (2003) The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Res., 13, 662-672.

Frishman, D., Mokrejs, M., Kosykh, D., Kastenmuller, G., Kolesov, G., Zubrzycki, I., Gruber, C., Geier, B., Kaps, A., Albermann, K. et al. (2003) The PEDANT genome database. Nucleic Acids Res, 31, 207-211.

Gilks, W.R., Audit, B., De Angelis, D., Tsoka, S. and Ouzounis, C.A. (2002) Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics, 18, 1641-1649.

Iliopoulos, I., Tsoka, S., Andrade, M.A., Enright, A.J., Carroll, M., Poullet, P., Promponas, V., Liakopoulos, T., Palaios, G., Pasquier, C. et al. (2003) Evaluation of annotation strategies using an entire genome sequence. Bioinformatics, 19, 717-726.

Kaplan, N., Vaaknin, A. and Linial, M. (2003) PANDORA: keyword-based analysis of protein sets by integration of annotation sources. Nucleic Acids Res, 31, 5617-5626.

Karchin, R., Karplus, K. and Haussler, D. (2002) Classifying G-protein coupled receptors with support vector machines. Bioinformatics, 18, 147-159.

Karplus, K., Barrett, C. and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics, 14, 846-856.

Linial, M. (2003) How incorrect annotations evolve--the case of short ORFs. Trends Biotechnol, 21, 298-300.

Muller, A., MacCallum, R.M. and Sternberg, M.J. (1999) Benchmarking PSI-BLAST in genome annotation. J Mol Biol, 293, 1257-1271.

Sigrist, C.J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. and Bucher, P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform, 3, 265-274.