Vinayagam Arunachalam
Dr.sc.hum.

Strategies for automatic genome annotation

Geboren am 10.05.1798 in Cheyyar, Indien
Diplom der fachrichtung Mikrobiologie an der Bharathidasan University, India

Promotionsfach: DKFZ, Bioinformatik
Doktorvater: Prof Dr. Sándor Suhai

Abstract

The thesis described a strategy developed for automatic Gene Ontology term annotation to uncharacterized cDNA or protein sequences. Function assignment is achieved with the usual homology based approach. However the method is adapted to create a complete automated system as well as to overcome the usual pit-falls faced with homology approaches. The method efficiently combined the ongoing efforts of Gene Ontology and the availability of GO-mapped sequences databases with a profound machine learning system to predict GO terms. As an application, the *Xenopus laevis* contig sequences were annotated with GO terms. The *Xenopus laevis* annotation showed remarkably enhanced annotation coverage with our approach compared to other systems. Finally the system is made available as an online tool to provide high quality annotation to the biological researchers.

Introduction

Ongoing genome sequencing and recent developments in cDNA sequencing projects have led to an exponential rise in the amount of sequence information. This has increased the need for acquiring knowledge from sequences as to their biological function. Annotating a single sequence is the gateway to interpreting its biological relevance. However, the

usefulness of these annotations is highly correlated with their quality. *Accurate* annotation has traditionally been maintained manually with the experience of individual experts and the experimental characterization. However, the increasing gap between the amount of sequence data available and the time needed for their experimental characterization makes it necessary for computational function prediction to complement manual curation. Commonly, computational functional assignment is based on homologues identified from database searches. Such an automated annotation process provides comparable results due to a uniform analysis of all query sequences across the same databases and the possibility of repeating the annotation to updated sequence data. However, crucial aspects for consideration in automated annotation are the problems associated with the databases as well as the problems associated with the inference. Though a number of excellent annotation systems have been developed to tackle these problems, little has been done to quantify the annotation accuracy by defined benchmarks *and* establish a method to provide an appropriate confidence value for each annotation.

Design and development of the system

We have developed an automated system to predict molecular function GO-terms for an unknown sequence and to define a confidence value for each prediction. Usually, for a function assignment the uncharacterized sequence is searched against known sequence databases and the function is transferred from the best hit. We introduced major modifications compared to this traditional approach to make the system entirely automated, and to overcome the usual pit-falls associated with homology approaches. The basic problem in automating the annotation process is the expression of the annotation in a rich, non-formalized language. We tackled this by using the Gene Ontology (GO), which provides controlled vocabularies to annotate gene products. The ontological terms are suited for automated handling since they are comprehensive, machine readable and processable. Apart from this, GO-terms are organism independent and as such suited well to annotate any new sequence.

In order to obtain GO annotations for unknown sequences, we searched them against GO-mapped protein databases. We selected protein databases because the functional transitivity at the protein level is more reliable. Furthermore, within these databases, each GO association is attached with the evidence of the annotation. Erroneous annotation in sequence databases is a serious problem in both cases of manual and automatic annotation, and usually it is impossible to track down such annotation errors. However with GO-mapped databases

and with their evidence of association, we could weigh the quality of different annotations. As a result of this, we reduced the possibility of spreading erroneous annotations.

Generally, homology searches result in a mixture of both appropriate (correct) and inappropriate (false) functions. Separating the correct function terms from the false ones is a crucial step. Especially in our approach we used a relaxed E-value cut-off for BLAST search. Though it yielded abundant negative samples, we were able to capture a significant portion of positive GO-terms at this cut-off. So our focus was to extract all possible information from BLAST and then to apply an intelligent system to predict the function. We used Support Vector Machines (SVMs) as the machine learning method to find the discriminating rules and to classify whether the extracted GO-terms were appropriate to the cDNA sequence or not. SVMs operate by finding a hyperplane in the space of possible inputs. This hyperplane splits the positive from the negative samples. The split is chosen to have the largest distance from the hyperplane to the nearest of the positive and negative samples.

In the training phase, these GO-terms were compared to the GO annotation of the query sequences and labeled correspondingly. We selected GO annotated cDNA sequences for training the SVM classifier. The nucleotide sequences were searched against GO-mapped protein databases and GO annotations were extracted from the significant hits. Then, each obtained GO-term was utilized as a sample for the feature table. The sample GO-terms were labeled as either correct ("+1") or false ("-1") by comparing to the original annotation. Next, the samples were attached with their features or attributes, calculated from the BLAST results. We used a broad variety of elaborated features (attributes) including sequence similarity measures, GO-term frequency, GO-term relationships between homologues, annotation quality of the homologues, and the level of annotation within the GO hierarchy. With this data the classifier was trained to distinguish between the attribute patterns that contributed to class +1 (correct prediction of a GO-term) and -1 (false prediction). To predict the function of unknown sequences, the same procedure as used for the training sequences was applied to unknown sequences for obtaining their GO-terms and corresponding attribute values. According to these attribute values the classifier assigned a class for every GO-term of the BLAST hits.

To enhance the reliability of the prediction, we used multiple SVMs for classification and applied a committee approach to combine the results with a voting scheme. The confidence values for the predicted GO-terms were assigned based on the number of votes i.e. the number of SVMs predicting a particular GO-term as correct. The performance of the system was benchmarked with 36,771 GO-annotated cDNA sequences derived from 13

organisms. The validation results show that our system was robust and the prediction performance was organism-independent. Furthermore, prediction quality of our system was comparable to the manual annotation. This shows that our approach benefited from the broad variety of potential attributes used for the functional transitivity and the vast amount of data used for training and validating. Especially the committee scheme exploited in our system provided a means to assign confidence values in a straightforward manner.

Application and availability

As an application to our system, we annotated *Xenopus laevis* sequences, since there was a demand from *Xenopus laevis* functional genomics community for quality GO annotation. We annotated the *Xenopus laevis* contig sequences from TIGR. Using our approach, we annotated 50.5% of all contig sequences presently available, and associated a confidence value for each prediction. This yielded roughly three times more annotation as compared that which is currently available. This shows a remarkably enhanced annotation coverage compared to the existing annotation. Finally, the system was converted into an online tool named GOPET (Gene Ontology term Prediction and Evaluation Tool; http://genius.embnet.dkfz-heidelberg.de/menu/biounit/open-husar). The availability of GOPET as an online tool is significantly helpful to the annotation field, since a large number of cDNA and protein sequences remains unannotated by GO.