Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 70 (2007) 718–725
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Predicting Gene Ontology functions based on support vector machines and statistical significance estimation

Ran Bi, Yanhong Zhou*, Feng Lu, Weiqiang Wang

Hubei Bioinformatics and Molecular Imaging Key Laboratory, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

Available online 14 October 2006

Abstract

Gene Ontology (GO) is a common language for the functional annotation of gene products. We have developed a computational tool, GOKey, to predict the GO function of proteins based on their sequence features and the support vector machine (SVM) method. Several measures, including improved handling of the problem caused by unbalanced positive and negative training data and postprocessing strategies to evaluate the posterior probability and statistical significance of SVM outputs, have been adopted to improve the prediction performance of GOKey. The GOKey has been trained to predict the 36 GO categories of the 'molecular function' of GO slims, and could be easily extended to other GO categories. The results of 5-fold cross validation with 10,603 GO-mapped proteins demonstrate that the performance of GOKey is better than that of standard SVMs. Comparisons with other computational tools for GO function prediction also show that the performance of GOKey is satisfactory. Further, GOKey has been applied to predict the GO functions for 5381 novel human proteins in the Ensembl database. The results show that 93% of the novel proteins can be assigned one or more GO terms, and some evidences supporting the predictions have been found. GOKey can be accessed at http://infosci.hust.edu.cn.

Keywords: Protein function; Gene Ontology; Support vector machines; Statistical significance

1. Introduction

The characterization of biochemical functions of gene products is an important task in the postgenomic era. With the progress in genome sequencing of more and more model organisms, however, there exists an increasing gap between the amount of available sequence data and the experimental characterization of their functions. Accurate computational function prediction, that is helpful for speeding up the functional annotation of gene products, has become an increasingly important problem in the field of bioinformatics [1–3].

Gene Ontology (GO), a set of structured vocabularies organized in a rooted directed acyclic graph (DAG), can provide consistent description of gene products in a species-independent manner [4,5], and has become a common language for the functional annotation of gene products. In recent years, many computational methods for the functional annotation of gene products have been developed based on GO, which can be classified as context based methods and sequence based methods according to the utilized information [1].

High-throughput experimental context data such as the protein–protein interaction and gene expression data provide a valuable resource, for the prediction of protein functions. The representatives of context-based methods include the Markov random field model to exploit protein–protein interaction data [6,7] and the clustering analysis to make use of gene expression data [8]. Although these methods are promising and have achieved certain success, currently, their usefulness is still limited by the quality and quantity of the data [9,10].

It is widely accepted that proteins' functions are determined, to a great extent, by their three-dimensional structures and the structures are determined by the amino acid sequences. Therefore, how to predict protein functions from sequence information is essential and attractive. The representatives of sequence-based computational tools for GO function prediction include GoFigure [11],

*Corresponding author.
E-mail address: yhzhou@hust.edu.cn (Y. Zhou).

0925-2312/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.neucom.2006.10.006

GOblet [12], OntoBlast [13], ProtFun [14] and GO^{TM} Predictor [15]. The GoFigure, GOblet and OntoBlast are based on homologous information, which predict the GO terms of uncharacterized protein sequences with the GO annotations of similar sequences found in GO-mapped protein databases [11–13]. The ProtFun is an artificial neural network-based program, which can predict specific GO terms by using sequence derived protein features such as predicted sorting signals, post-translational modifications and some physical/chemical properties [14]. The GO^{TM} Predictor, which is based on the assumption that protein functions are determined by domains, makes predictions using detected domains and a set of previously generated rules for domain-function associations [15].

Recently, a relatively new classification method, support vector machines (SVMs) [16,17], has been introduced to solve various biological problems such as microarray data analysis [18], protein fold recognition [19], protein–protein interaction prediction [20], protein secondary structure prediction [21] and protein function prediction [22,23]. Specifically, Cai et al. [22] studied how to use the SVM method and protein sequence information to predict protein function, and developed a web-based tool SVMProt for classifying proteins into such functional classes as enzymes, receptors, transporters, channels, binding proteins and so on [22,23]. However, these protein function terms, written in rich and non-formalized language, are context dependent, and the vagueness in using such terms may yield confusing annotations [3,24,25]. In this work, we have developed a computational tool, GOKey, to predict the GO functions of proteins based on their sequence features and the SVM method. Especially, on the basis of SVM predictions, several postprocessing measures have been adopted to improve the prediction performance of GOKey.

2. Materials and methods

2.1. The configuration of GOKey

The configuration of our function predictor GOKey is shown in Fig. 1. The GOKey consists of a coding unit and a calculation unit. The calculation unit, which contains a group of SVMs (each SVM corresponding to a specific GO category) and other postprocessing models for estimating the probabilities (the 'probability' modules) and statistical significance (the '*P*-value' modules) of the prediction results of SVMs, is the kernel of the system.

In principle, the GOKey can be trained to predict any GO term for which there are enough training data. Here, we adopt the 'molecular function' of the GO slims (updated on 7 December 2004), which have been applied in the annotation of several genomes [5], as an example to clarify the application of GOKey. Of the 40 nodes in 'molecular function' of GO slims, three nodes, GO:0030188 (chaperone regulation activity), GO:0045735 (nutrient reservoir activity) and GO:0030533 (triplet codon-amino

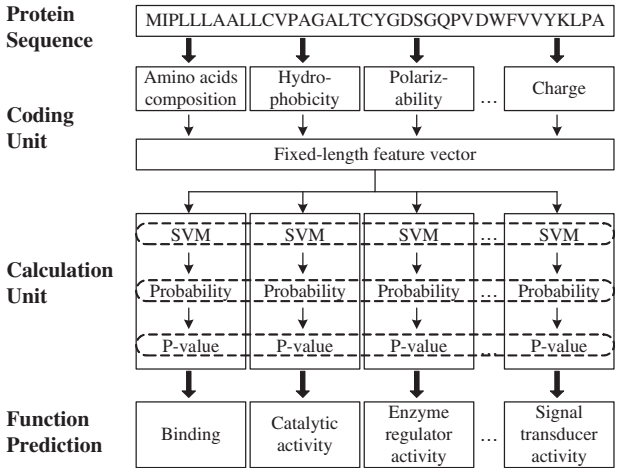

Fig. 1. The configuration of GOKey system.

acid adaptor activity) are discarded because they currently cover very few proteins (less than ten protein sequences) to build effective prediction models. In addition, GO:0005554 (molecular function unknown) is also excluded. The process for developing the GOKey that can predict the remaining 36 GO terms includes the following steps:

(1) Selecting positive and negative samples for each of the 36 GO categories. The positive samples are a set of protein sequences annotated with the GO term, whereas the negative samples are a set of sequences that do not have the corresponding GO function. The positive and negative samples are used as the benchmark data for constructing models that can effectively distinguish proteins with a specific GO function from those without.
(2) Encoding positive and negative samples using sequence features. The purpose of this step is to construct feature vectors of positive and negative samples using sequence features for building classification models.
(3) Building classification models using feature vectors. The goal of protein function prediction is to decide which GO category an uncharacterized protein belongs to. Obviously, this is a classification problem. There exist many classification methods that first constructing decision functions (classifiers) with a training dataset and then make decisions according to the values of these functions (decision values). In this study, the SVM and several postprocessing models are adopted.
(4) Validating the classification models. The performance of the classification models is evaluated with the positive and negative samples by a 5-fold cross-validation process.

2.2. Selection of positive and negative samples

To select reliable positive and negative samples for each of the 36 GO nodes, here, we use the concept of functional