

Assigning Gene Ontology Categories (GO) to Yeast Genes Using Text-Based Supervised Learning Methods

Tomonori Izumitani Hirotoishi Taira Hideto Kazawa
Eisaku Maeda
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{izumi, taira, kazawa, maeda}@cslab.kecl.ntt.co.jp

Abstract

We propose a method for assigning upper level Gene Ontology terms (GO categories) to genes using relevant documents. This method represents each gene as a vector using relevant documents to the gene. Then, binary classifiers are made for the GO categories using such supervised learning methods as support vector machines and maximum entropy method. We applied this method for assigning GO categories to yeast genes and achieved an average F-measure of 0.67, which is >0.3 higher than the existing method developed by Raychaudhuri et al. We also applied this method to genome-wide annotation for yeast by all GO Slim categories provided by SGD and achieved average F-measures of 0.58, 0.72, and 0.60, respectively, for the three GO parts: cellular component, molecular function, and biological process.

1. Introduction

In such genome database projects as Saccharomyces Genome Database (SGD),¹ each gene is associated with thousands of Gene Ontology (GO) terms [1]. In SGD, each gene is also annotated by GO Slim² which consists of a few dozen major categories of GO terms. In this study, we call such categories “GO categories”.

GO terms or categories are manually assigned by curators after reading many relevant scientific papers. In such a time-consuming operation automatic assignment is demanding. In this study, we focus on a problem of such automatic assignment of GO categories as GO Slim categories to genes.

Raychaudhuri et al. developed a method to assign GO categories to genes using a supervised learning method [2].

¹ <http://www.yeastgenome.org>

² <ftp://ftp.geneontology.org/pub/go/GO-slms/>

It classifies documents associated with each gene and ranks GO categories by a voting scheme.

This method does not decide the number of GO categories to be assigned to each gene. Neither does it utilize genes to which GO categories have already been assigned, although the genes give important information for the assignment. This may cause performance degradation.

To resolve these problems, we propose a method that directly assigns GO categories to genes using supervised learning methods such as a support vector machine (SVM) and a maximum entropy method (MEM). It uses genes as training data that have already been assigned by GO categories.

We applied this method to yeast genes registered in SGD and assigned two kinds of GO categories: twelve categories selected from Raychaudhuri et al.’s study and GO Slim categories.

The proposed method can provide temporary annotations for genes that have not yet been checked by curators, contributing to the efficiency of GO assignment.

2. Methods

The proposed method adopts a simple strategy for directly assigning GO categories to each gene.

First, for each gene, the system combines documents associated with the gene to produce a vector that has binary (0/1) values based on whether the combined document contains the relevant words.

Second, classifiers are made using supervised learning methods such as SVM and MEM. Each gene can have multiple GO categories; therefore, for each GO category we made a binary classifier. The assignment of GO categories is completed by integrating these classifiers’ outputs.

3. Results and Discussion

We applied the proposed method for assignment of GO categories to yeast genes registered in SGD. Two kinds of GO category sets were tested: the twelve GO categories selected from Raychaudhuri et al.'s study and GO Slim.

3.1. Assigning Raychaudhuri et al.'s twelve GO categories

Raychaudhuri et al. tested 21 GO categories from the biological process part of GO. From these 21 categories, we selected twelve significant GO categories using recent versions of GO, downloaded on 16/Feb/2004, and SGD, downloaded on 6/Nov/2003.

We assigned twelve GO categories to 3,295 genes in SGD that are associated with at least three PubMed abstracts. Performance was determined by F-measure, a harmonic mean between precision and recall, and evaluated by five-fold cross validation.

Table 1 shows the F-measures using SVM with first-order polynomial kernel (A) and MEM (B). For Raychaudhuri et al.'s method, we tried three thresholds of the number of GO categories to be assigned because it only provides the rank of GO categories. In each table, "Top 1," "Top 2," and "Top 3" denote Raychaudhuri et al.'s method, assigning the top one, two, and three GO categories, respectively.

Results show that our proposed method outperforms Raychaudhuri et al.'s method in most GO categories. SVM especially improves the F-measure in "Cell adhesion," "Cell death," and "Cell proliferation," which are not assigned effectively by Raychaudhuri et al.'s method.

3.2. Assigning GO Slim categories

Yeast GO Slim categories were selected by the SGD curators based on annotation statistics and biological significance. We used all GO Slim categories in the 2003 version of the yeast GO Slim, except for the "unknown" categories in the three GO parts.

We used SVM with a first-order polynomial kernel and separately tried some term weighting methods, TF, IDF, and TFIDF, in addition to binary vectors.

The performance was evaluated for averaged precision, recall, and F-measure through five-fold cross validations. Then the optimal weighting method was selected for each GO part.

For Three GO parts, table 2 shows the number of tested genes, the number of GO Slim categories, optimal weighting methods, and the performance of the proposed method.

Although GO Slim contains 11 - 21 more categories than Raychaudhuri et al.'s GO categories, the performances are comparable, especially in the molecular function.

(A) SVM (1st order polynomial kernel)				
GO category	Proposed method	Raychaudhuri et al.'s method		
		Top 1	Top 2	Top 3
Autophagy	0.78	0.83	0.66	0.38
Biogenesis	0.76	0.35	0.41	0.44
Cell adhesion	0.51	0.19	0.19	0.13
Cell death	0.58	0.07	0.06	0.02
Cell proliferation	0.76	0.00	0.03	0.06
Ion homeostasis	0.61	0.16	0.42	0.33
Membrane fusion	0.64	0.24	0.16	0.12
Metabolism	0.91	0.42	0.65	0.74
Signal transduction	0.76	0.41	0.30	0.21
Sporulation	0.37	0.22	0.14	0.07
Stress response	0.65	0.41	0.27	0.24
Transport	0.83	0.56	0.55	0.49
Average	0.67	0.32	0.32	0.27

(B) MEM				
GO category	Proposed method	Raychaudhuri et al.'s method		
		Top 1	Top 2	Top 3
Autophagy	0.46	0.76	0.55	0.43
Biogenesis	0.69	0.37	0.42	0.43
Cell adhesion	0.19	0.15	0.16	0.09
Cell death	0.13	0.12	0.03	0.04
Cell proliferation	0.70	0.00	0.05	0.09
Ion homeostasis	0.47	0.07	0.24	0.23
Membrane fusion	0.30	0.33	0.18	0.14
Metabolism	0.90	0.18	0.47	0.61
Signal transduction	0.52	0.40	0.25	0.19
Sporulation	0.19	0.19	0.14	0.08
Stress response	0.55	0.43	0.27	0.24
Transport	0.78	0.52	0.46	0.42
Average	0.49	0.29	0.27	0.25

Table 1. F-measure for Raychaudhuri et al.'s GO categories

GO part	(# genes)	# categories	Weighting method	Precision	Recall	F-measure
Cellular Component	(4,619)	23	Binary	0.61	0.56	0.58
Molecular Function	(3,282)	22	TFIDF	0.72	0.73	0.72
Biological Process	(3,763)	33	IDF	0.59	0.63	0.60

Table 2. Assignment of yeast GO Slim categories using SVM

4. Conclusions

We proposed a method to assign GO categories to genes using text-based supervised learning methods. This method outperforms the existing method by directly using genes, to which GO categories have already been assigned, as training data. The method also successfully assigned the GO Slim categories.

References

- [1] T. G. O. Consortium. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25:25–29, 2000.
- [2] S. Raychaudhuri, J. T. Chang, P. D. Sutphin, and R. B. Altman. Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.*, 12(1):203–214, January 2002.