

# Prediction of Gene Ontology Annotations based on Gene Functional Clustering

Marco Tagliasacchi, Roberto Sarati and Marco Masseroli

*Dipartimento di Elettronica e Informazione, Politecnico di Milano - Italy*

**Abstract**—We propose an algorithm that predicts potentially missing Gene Ontology annotations, in order to speed up the time-consuming annotation curation process. The proposed method extends a previous work based on the singular value decomposition of the gene-term annotation matrix and incorporates gene clustering, based on gene functional similarity computed by means of the Gene Ontology annotations. We tested the prediction method by performing  $K$ -fold cross-validation on the genomes of two organisms, *Saccharomyces cerevisiae* (SGD) and *Drosophila melanogaster* (FlyBase).

**Keywords**—Annotation prediction; Singular Value Decomposition; gene similarity metrics

## I. INTRODUCTION

Several controlled vocabularies are routinely used to annotate genes and proteins. The most widely used is the Gene Ontology (GO) [1]. It comprises three ontologies describing specie-independent biological process (BP), molecular function (MF) and cellular component (CC) attributes of genes and gene products. Databases providing controlled annotations contain the biological knowledge that has been gathered over the years. Despite their relevance, there are important issues that affect these databases [2], since annotations might be incomplete, incorrect or not exhaustive.

Computational tools have been used to assess the relevance of inferred annotations, or produce a ranked list of missed annotations in order to speed up the curation process [3], [4], [5]. We propose a method for predicting annotations to GO terms based solely on previously available GO annotations. Our first contribution consists in extending and enhancing the method in [5], hereafter denoted SVD method, by incorporating a gene (or gene product) clustering algorithm based on the functional similarity between gene (or gene product) pairs. The proposed method, denoted SIM method, computes a separate set of eigen-terms for each identified cluster, while the original SVD method computes a global set of eigen-terms.

## II. PREDICTION METHODS

### A. SVD Prediction Method

Let  $\mathbf{A}$  denote a modified gene-to-term matrix, where  $\mathbf{A}(i, j) = 1$ , if gene  $i$  is annotated to term  $j$  or to any descendant of  $j$ . Otherwise,  $\mathbf{A}(i, j) = 0$ . Thus, the  $i$ -th row  $\mathbf{a}_i^T$  of the matrix  $\mathbf{A}$  contains all the direct and indirect annotations of gene  $i$ . According to the work in [5], annotation prediction can be performed by computing the singular value

decomposition (SVD) of the matrix  $\mathbf{A}$ , i.e.  $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ , where  $\mathbf{U}$  is a  $m \times p$  unitary matrix (i.e.  $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ ),  $\mathbf{\Sigma}$  is a non-negative diagonal matrix of size  $p \times p$ , and  $\mathbf{V}$  is a  $n \times p$  unitary matrix, where  $p = \min(m, n)$ . For any positive integer  $k < \text{rank}(\mathbf{A})$ , it is possible to generate a matrix  $\hat{\mathbf{A}} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^T$ , where  $\hat{\mathbf{U}}$  ( $\hat{\mathbf{V}}$ ) is a  $m \times k$  ( $n \times k$ ) matrix obtained retaining the first  $k$  columns of  $\mathbf{U}$  ( $\mathbf{V}$ ) and  $\hat{\mathbf{\Sigma}}$  is a  $k \times k$  diagonal matrix with the  $k$  largest singular values along the diagonal.

The study of the matrix  $\hat{\mathbf{A}}$  reveals semantic relationships of the gene-function associations [5]. A large value of  $\hat{\mathbf{A}}(i, j)$  suggests that gene  $i$  should be annotated to term  $j$ , whereas a value close to zero suggests the opposite. Note that the  $i$ -th row of  $\hat{\mathbf{A}}$  can also be written as  $\hat{\mathbf{a}}_i^T = \mathbf{a}_i^T \hat{\mathbf{V}} \hat{\mathbf{V}}^T$ . Thus, the original annotation profile is first transformed in the eigen-term domain, while retaining only the first  $k$  eigen-terms by the multiplication with  $\hat{\mathbf{V}}$ , and then mapped back to the original domain by means of  $\hat{\mathbf{V}}^T$ .

### B. SIM Prediction Method

The SVD method implicitly adopts a global term-to-term correlation matrix  $\mathbf{T} = \mathbf{A}^T\mathbf{A}$ , which is estimated from the whole corpus of available annotations. Instead, we propose an adaptive approach, which clusters genes based on their original annotation profiles and estimates a set of distinct correlation matrices  $\mathbf{T}_c$ ,  $c = 0, \dots, C$ , where  $C$  denotes the number of clusters and  $\mathbf{T}_0 = \mathbf{T}$ . For each matrix  $\mathbf{T}_c$ , a corresponding set of  $k$  eigenvectors  $\hat{\mathbf{V}}_c$  is computed, originating  $C + 1$  predicted annotation profiles for the  $i$ -th gene:

$$\hat{\mathbf{a}}_{i,c}^T = \mathbf{a}_i^T \hat{\mathbf{V}}_c \hat{\mathbf{V}}_c^T \quad c = 0, \dots, C \quad (1)$$

The selected predicted annotation profile  $\hat{\mathbf{a}}_{i,c}^T$  for the  $i$ -th gene is the one that minimizes the variation, measured by means of the ell-2 norm, with respect to the original annotation profile of the gene:

$$c_i^* = \arg \min_{c=0, \dots, C} \|\hat{\mathbf{a}}_{i,c} - \mathbf{a}_i\|_2 \quad (2)$$

In order to estimate the correlation matrices  $\mathbf{T}_c$ , we need to cluster genes based on their functional similarity expressed by their annotations. To this end, we can exploit the SVD of the matrix  $\mathbf{A}$  as suggested in [6]. In fact, each column  $\mathbf{u}_c$  of the matrix  $\mathbf{U}$  represents a cluster, and the value  $\mathbf{U}(i, c)$  indicates the membership of gene  $i$  to the  $c$ -th cluster. Therefore, each gene might belong to more than one cluster with different degrees of membership. We notice

that the columns of  $\mathbf{U}$  are a set of eigenvectors for the matrix  $\mathbf{G} = \mathbf{A}\mathbf{A}^T$ , i.e. the similarity between gene pairs is measured by the inner product of their annotation profiles. Since  $\mathbf{A}$  is binary-valued,  $\mathbf{G}(i_1, i_2)$  is the count of common terms in the annotation profiles of genes  $i_1$  and  $i_2$ . The estimation of  $\mathbf{T}_c$  proceeds as follows. First, for each cluster, we generate a modified gene-to-term matrix  $\mathbf{A}_c = \mathbf{W}_c\mathbf{A}$ , where  $\mathbf{W}_c \in \mathbb{R}^{m \times m}$  is a diagonal matrix with the entries of  $\mathbf{u}_c$  along the main diagonal. Therefore, the  $i$ -th row is weighted by the membership score of the corresponding gene to the  $c$ -cluster. Then, we compute  $\mathbf{T}_c = \mathbf{A}_c^T \mathbf{A}_c$ .

A more accurate clustering can be obtained by incorporating the functional similarity between GO terms. We propose to perform the gene clustering by computing the eigenvectors of the modified matrix  $\tilde{\mathbf{G}} = \mathbf{A}\mathbf{S}\mathbf{A}^T$  where  $\mathbf{S} \in \mathbb{R}^{n \times n}$  denotes the term similarity matrix. Given a pair of ontology terms,  $j_1$  and  $j_2$ , the term functional similarity  $\mathbf{S}(j_1, j_2)$  can be computed using different methods (see [4] and references therein). Here we adopt the Lin’s similarity metrics

### III. SVD VS. SIM COMPARATIVE ANALYSIS

We assessed the performance of the SVD and SIM methods by performing  $K$ -fold cross-validation as in [3], [4], and discarding terms used to annotate less than  $M$  genes. As for the SIM method, we evaluated two variants. In SIM1, we set  $\mathbf{S} = \mathbf{I}$ , i.e. the clustering step does not rely on the functional similarity between terms. In SIM2, the matrix  $\mathbf{S}$  is computed by means of the Lin’s metrics. In both cases we heuristically set a fixed number of clusters  $C = 5$  for all ontologies.

Based on a threshold value  $\tau$ , if  $\mathbf{A}(i, j) = 0$  and  $\hat{\mathbf{A}}(i, j) > \tau$ , a new annotation is suggested; this case is denoted as a false positive (FP). Conversely, if  $\mathbf{A}(i, j) = 1$  and  $\hat{\mathbf{A}}(i, j) \leq \tau$ , an existing annotation is suggested to be semantically inconsistent with the available data; this case is denoted as a false negative (FN). True positives (TP) and true negatives (TN) are similarly defined.

To improve reliability, in  $\tilde{\mathbf{A}}$  we retained GO terms used to annotate at least  $M = 3$  or  $M = 10$  genes of the considered organism and excluded annotations with evidence code IEA (inferred electronic annotations).

As an aggregated indicator of the prediction performance, we computed the area above the FN rate vs. FP rate curve (AAC) in the  $[0, 0.01]$  range, for both *Saccharomyces cerevisiae* (SGD) and *Drosophila melanogaster* (FlyBase). In Table I we show the results for SGD only. Indeed, we are typically interested in the low range of FP rate, since it corresponds to top-ranked predictions of newly inferred annotations (FP) with the highest score. The normalized AAC metrics is bounded in the  $[0, 1]$  interval, where a value close to 1 implies more accurate predictions. In all cases, we computed the AAC metrics considering only the prediction of GO terms with depth  $L$  from the root of the ontology greater than either 2 or 6. Using the AAC metrics, Table I shows that the SIM method generally outperforms the

Table I  
AREA ABOVE THE CURVE OF FALSE NEGATIVE (FN) RATE VS. FALSE POSITIVE (FP) RATE OF THE GO ANNOTATIONS OF *Saccharomyces cerevisiae* (SGD)

$M$	method	$k$	BP		MF		CC	
			$L > 2$	$L > 6$	$L > 2$	$L > 6$	$L > 2$	$L > 6$
3	SVD	20	0.58	0.35	0.47	0.51	0.39	0.50
		40	0.65	0.57	0.57	0.60	0.32	0.51
	SIM1	20	0.64	0.53	0.56	0.64	<b>0.41</b>	<b>0.60</b>
		40	0.70	0.61	0.52	0.61	0.37	0.56
	SIM2	20	0.64	0.50	<b>0.59</b>	<b>0.70</b>	0.37	0.56
		40	<b>0.71</b>	<b>0.62</b>	0.55	0.62	0.36	<b>0.60</b>
10	SVD	20	0.53	0.34	0.43	0.49	0.35	0.43
		40	0.60	0.53	0.53	0.59	0.31	0.47
	SIM1	20	0.62	0.52	0.50	0.56	<b>0.43</b>	<b>0.56</b>
		40	0.65	<b>0.60</b>	0.46	0.39	0.39	<b>0.56</b>
	SIM2	20	0.63	0.52	<b>0.54</b>	<b>0.65</b>	0.37	0.53
		40	<b>0.67</b>	0.58	0.49	0.47	0.35	<b>0.56</b>

SVD method for all GO ontologies. In most cases, SIM2 outperforms SIM1 (in 53% of cases SIM2 was better than, or equal to, the SIM1 method), showing that clustering based on the functional similarity between terms might be beneficial. Nevertheless, most of the performance gain between SIM and SVD stems from the adaptive nature of SIM, regardless on how clustering is actually performed. In fact, the SVD method, which computes similarities between clusters in terms of frequency of co-annotation, is bound to be biased towards the larger clusters, since it is unnormalized. The SIM method counterbalances such a bias with its adaptive approach of clustering genes (or gene products) according to their original annotation profile. A similar analysis was conducted on the GO annotations of other organisms showing comparable results.

### REFERENCES

- [1] The Gene Ontology Consortium, “Creating the gene ontology resource: Design and implementation,” *Genome Res.*, vol. 11, pp. 1425–1433, 2001.
- [2] P. D. Karp, “What we do not know about sequence analysis and sequence databases,” *Bioinformatics*, vol. 14, pp. 753–754, 1998.
- [3] O. D. King, R. E. Foulger, S. S. Dwight, J. V. White, and F. P. Roth, “Predicting gene function from patterns of annotation,” *Genome Res.*, vol. 13, no. 5, pp. 896–904, 2003.
- [4] Y. Tao, L. Sam, J. Li, C. Friedman, and Y. A. Lussier, “Information theory applied to the sparse gene ontology annotation network to predict novel gene function,” *Bioinformatics*, vol. 23, no. 13, pp. 529–538, 2007.
- [5] P. Khatri, B. Done, A. Rao, A. Done, and S. Draghici, “A semantic analysis of the annotations of the human genome,” *Bioinformatics*, vol. 21, no. 16, pp. 3416–3421, 2005.
- [6] P. Drineas, “Clustering large graphs via the singular value decomposition: Theoretical advances in data clustering (guest editors: Nina mishra and rajeev motwani),” *Machine Learning*, vol. 56, pp. 9–33, July 2004.