

Semantically Improved Genome-Wide Prediction of Gene Ontology Annotations

Marco Masseroli, Marco Tagliasacchi, Davide Chicco

Dipartimento di Elettronica e Informazione
Politecnico di Milano

Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Email: masseroli@elet.polimi.it, tagliasacchi@elet.polimi.it, davide.chicco@elet.polimi.it

Abstract—Genomic annotations describing the structural and functional features of genes and gene products by means of controlled terminologies and ontologies are extremely valuable, in particular for computational analyses aimed at inferring new biomedical knowledge, which usually rely on available annotations. Yet, they are incomplete, especially for more recently studied genomes, and only some of the available annotations represent highly reliable human curated information. In order to help and speed up the time-consuming curation process and improve the available annotations, computational methods that are able to provide a prioritized list of predicted annotations are hence extremely useful. Starting from a previous work on the automatic prediction of Gene Ontology annotations based on the singular value decomposition (SVD) of the gene-to-term annotation matrix, in this work we propose a novel prediction algorithm that incorporates gene clustering based on gene functional similarity computed by means of the Gene Ontology annotations. We tested the prediction methods performing k-fold cross-validation on the genomes of two organisms, *Saccharomyces cerevisiae* (SGD) and *Drosophila melanogaster* (FlyBase). Results demonstrate the effectiveness of our approach.

Index Terms—Annotation prediction; Singular Value Decomposition; gene similarity metrics

I. INTRODUCTION

In molecular biology, new approaches are providing unprecedented amount of valuable data that foster the increasing relevance of molecular medicine in health care research and practice. In particular, high-throughput microarray technologies allow quickly and simultaneously studying thousands of genes and gene products. At the same time, advancements in information technologies and biomedical informatics are providing tools and techniques to manage the amount of biomedical data produced, as well as many methods for their analysis. In addition, biomedical domain experts are increasingly annotating biomolecular entities, mainly genes and their protein products, with controlled terminologies and ontologies describing their structural, functional and phenotypic biological features. Currently, several controlled vocabularies are routinely used to annotate genes and proteins. Some of them have a flat structure, i.e. no explicit relationships between the terms composing the vocabulary exist. Others are part of ontologies, where semantic relationships are defined between pairs of terms. The most widely used ontology for annotating biomolecular entities is the Gene Ontology

(GO) [1]. It comprises three ontologies that hold a total of nearly 26,000 controlled terms describing specie-independent biological process (BP), molecular function (MF) and cellular component (CC) attributes of genes and gene products. Each GO ontology is designed to capture orthogonal aspects of genes and gene products, and it is structured as a directed acyclic graph (DAG) of terms hierarchically related mainly through "is a" or "part of" relationships. An edge exists from a child term *a* to its parent term *b* if *a* "is a" specific instance of *b* or it is "part of" *b*. Furthermore, in each GO DAG it exists a unique root, which is defined as the DAG node without parents, and each term can have multiple parents.

Annotation databases contain the biological knowledge that has been gathered over the years, and provide such valuable data as public repositories. Despite their relevance, there are important issues that affect annotation databases [2]. In particular, first, the annotations are not exhaustive: only a subset of genes and gene products of sequenced organisms is known and, among those, only a small fraction has been annotated so far. Furthermore, annotation profiles might be incomplete, because the biological knowledge about the functions associated with a gene or a gene product might be yet to be discovered, or the evidence already available in the literature has not been entered into the database yet. Second, available annotations might be incorrect, e.g. those inferred from electronic annotations without the involvement of a human curator.

In this context, the contributions of computational tools able to analyze data stored in annotation databases are manifold. For example, it is possible to assess the relevance of inferred annotations, or produce a ranked list of missed annotations in order to speed up the curation process. Furthermore, since most of the bioinformatics analyses currently performed on genomic and proteomic data rely on the available annotations of genes and gene products, an improvement of such annotations both in quantity, coverage and quality is paramount to obtain better results in these analyses.

A few years ago, King et al. [3] proposed the use of decision trees and Bayesian networks for predicting annotations by learning patterns from available annotation profiles. Recently, Tao et al. [4] proposed to use a k-nearest neighbor (k-NN) classifier, whereby a gene inherits the annotations that

are common among its nearest neighbor genes, determined according to the functional distance between genes, based on the semantic similarity of GO terms used to annotate them. More simply, by using basic linear algebra tools, Khatri et al. [5] proposed a prediction algorithm based on the singular value decomposition (SVD) of the gene-to-term annotation matrix, which is implicitly based on the count of co-occurrences between pairs of terms in the available annotation database. Since this last method provides the basis for the work presented in this paper, it will be subsequently summarized in Section II-A.

Missing annotations can also be inferred by taking advantage of multiple data sources. In [6] expression levels obtained in microarray experiments are used to train a Support Vector Machine (SVM) classifier for each annotation term, and consistency among predicted annotation terms is enforced by means of a Bayesian network mapped onto the GO structure. Textual information is leveraged in [7] and [8], where the literature is mined and the keywords extracted from published papers are mapped to GO concepts.

By providing a list, possibly ranked, of annotations to be checked by a manual curator, the aforementioned techniques can drive the discovery of previously unknown annotations, as well as the detection of inconsistencies in the existing annotations. Furthermore, the updated annotation profiles can help boosting the performance of data analysis methods that rely upon them. These include, for example, querying for genes based on their similarity with a target annotation profile, or clustering genes based on their annotation profile [9] and [10].

In this paper we propose a method for predicting annotations of GO terms based solely on previously available GO annotations. Although there are some sophisticated techniques that predict gene functions leveraging other genomic data, techniques based solely on the available annotations have been demonstrated to be useful. 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 (Semantically IMproved) method, computes a separate set of eigen-terms for each identified cluster, while the original SVD method computes a global set of eigen-terms. We compare the SVD and the SIM methods by means of k-fold cross validation on the predicted gene annotations of two organisms, *Saccharomyces cerevisiae* (SGD) and *Drosophila melanogaster* (FlyBase), and we demonstrate improvements in the prediction of their GO annotations.

The rest of this paper is organized as follows. Section II-A illustrates the SVD method, since it represents the starting point for our work. Section II-B describes our proposed SIM method. We present and discuss our results in Section III-B. Section IV concludes the paper and provides some guidelines for future research on this topic.

II. PREDICTION METHODS

A. SVD Prediction Method

Let $\mathbf{A}_d \in \{0, 1\}^{m \times n}$ define the matrix representing all direct annotations of a specific GO ontology for a given organism. The m rows of \mathbf{A}_d correspond to genes (or gene products), while the n columns correspond to GO terms. The entries of \mathbf{A}_d assume values from the binary alphabet $\{0, 1\}$ according to the following rule:

$$\mathbf{A}_d(i, j) = \begin{cases} 1, & \text{if gene } i \text{ is annotated to term } j \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Annotation curators are asked to always use the most specific GO term for a given functional category. Thus, when a gene (or gene product) is annotated to a term, it is implicitly assumed to be indirectly annotated also to the more generic terms for that category, i.e. all the term ancestors in the GO DAG. As such, let \mathbf{A} denote a modified gene-to-term matrix, where the assignment of its entries is given by:

$$\mathbf{A}(i, j) = \begin{cases} 1, & \text{if gene } i \text{ is annotated to term } j \\ & \text{or to any descendant of } j \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The i -th row of the matrix \mathbf{A} , \mathbf{a}_i^T , contains all the direct and indirect annotations of gene i . Conversely, the j -th column encodes the list of genes that have been annotated (directly or indirectly) to term j . This process is sometimes defined as annotation unfolding.

According to the work in [5], annotation prediction can be performed by computing the SVD of the matrix \mathbf{A} , which is given by:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (3)$$

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 $p = \min(m, n)$. Conventionally, the entries along the diagonal of $\mathbf{\Sigma}$ (namely singular values) are sorted in non-increasing order. The number $r \leq p$ of non-zero singular values is equal to the rank of the matrix \mathbf{A} . For any positive integer $k < r$, it is possible to generate a matrix $\hat{\mathbf{A}}$, with:

$$\hat{\mathbf{A}} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^T \quad (4)$$

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. $\hat{\mathbf{A}}$ is the optimal rank- k approximation of \mathbf{A} , i.e. the one that minimizes the norm (either the spectral norm or the Frobenius norm) $\|\mathbf{A} - \hat{\mathbf{A}}\|$ subject to the rank constraint.

In [5] it is argued that the study of the matrix $\hat{\mathbf{A}}$ reveals semantic relationships of the gene-function associations. 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. As a matter of fact, the SVD of the matrix \mathbf{A} is equivalent to the method of latent semantic indexing (LSI) in information retrieval, where the input is a matrix that contains the occurrences of words in indexed documents.

In order to better understand why $\hat{\mathbf{A}}$ can be used to predict gene-to-term annotations, we point out that an alternative expression of (4) can be obtained by basic linear algebra manipulations:

$$\hat{\mathbf{A}} = \mathbf{A}\hat{\mathbf{V}}\hat{\mathbf{V}}^T \quad (5)$$

Moreover, the SVD of the matrix \mathbf{A} is related to the eigen-decomposition of the symmetric matrices $\mathbf{T} = \mathbf{A}^T\mathbf{A}$ and $\mathbf{G} = \mathbf{A}\mathbf{A}^T$. In fact, the columns of $\hat{\mathbf{V}}$ ($\hat{\mathbf{U}}$) are a set of k eigenvectors corresponding to the k largest eigenvalues of the matrix \mathbf{T} (\mathbf{G}). The matrix \mathbf{T} has a simple interpretation in our context. In fact,

$$\mathbf{T}(j_1, j_2) = \sum_{i=1}^m \mathbf{A}(i, j_1) \cdot \mathbf{A}(i, j_2) \quad (6)$$

i.e. $\mathbf{T}(j_1, j_2)$ is the number of times that terms j_1 and j_2 are used to annotate the same gene in the existing annotation profile. Therefore, $\mathbf{T}(j_1, j_2)$ expresses the (unnormalized) correlation between term pairs and it can be interpreted as a similarity score of the terms j_1 and j_2 computed solely based on the use of these terms in available annotations. The eigenvectors of \mathbf{T} (i.e. the columns of $\hat{\mathbf{V}}$) can be considered as a reduced set of eigen-terms. Intuitively, if two terms co-occur frequently, they are likely to be mapped to the same eigen-term. Based on (5), the i -th row of $\hat{\mathbf{A}}$ can be written as

$$\hat{\mathbf{a}}_i^T = \mathbf{a}_i^T \hat{\mathbf{V}}\hat{\mathbf{V}}^T \quad (7)$$

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 $\hat{\mathbf{V}}^T$. This corresponds to projecting the original vector \mathbf{a}_i^T onto the k -dimensional subspace spanned by the columns of $\hat{\mathbf{V}}$.

The entries of the matrix $\hat{\mathbf{A}}$ are real valued. In [5] it is defined a threshold τ such that, if $\hat{\mathbf{A}}(i, j) > \tau$, then gene i is predicted to be annotated to term j . Depending on the original values assumed by the matrix \mathbf{A} , the following cases might occur:

- If $\mathbf{A}(i, j) = 1$ and $\hat{\mathbf{A}}(i, j) > \tau$, the annotation of gene i to term j is confirmed; this case is denoted as a true positive (TP), with respect to the original $\mathbf{A}(i, j)$.
- 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), with respect to the original $\mathbf{A}(i, j)$.
- 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), with respect to the original $\mathbf{A}(i, j)$.
- If $\mathbf{A}(i, j) = 0$ and $\hat{\mathbf{A}}(i, j) \leq \tau$, the annotation is not present in the original annotation database and it is not suggested by the analysis; this case is denoted as a true negative (TN), with respect to the original $\mathbf{A}(i, j)$.

B. SIM Prediction Method

If an estimate of the term-to-term correlation matrix better than the one implicitly adopted by the SVD method is somehow available, it can be used to derive a different set

of eigenvectors potentially producing a better prediction by means of (7). The SVD method adopts a global 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:

$$\hat{\mathbf{a}}_{i,c}^T = \mathbf{a}_i^T \hat{\mathbf{V}}_c \hat{\mathbf{V}}_c^T \quad c = 0, \dots, C \quad (8)$$

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:

$$c_i^* = \arg \min_{c=0, \dots, C} \|\hat{\mathbf{a}}_{i,c} - \mathbf{a}_i\|_2 \quad (9)$$

We point out that when $c_i^* = 0$, for the i -th gene this corresponds to the original SVD prediction method based on the global correlation matrix \mathbf{T} . 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 singular value decomposition of the matrix \mathbf{A} defined in equation (3) as suggested in [11]. 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 $\mathbf{G} = \mathbf{A}\mathbf{A}^T$, where:

$$\mathbf{G}(i_1, i_2) = \mathbf{a}_{i_1}^T \mathbf{a}_{i_2} = \sum_{j=1}^n \mathbf{A}(i_1, j) \cdot \mathbf{A}(i_2, j) \quad (10)$$

i.e. the similarity between genes i_1 and i_2 is measured by the inner product of the 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:

$$\mathbf{A}_c = \mathbf{W}_c \mathbf{A} \quad (11)$$

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 of \mathbf{A} is weighted by the membership score of the corresponding gene to the c -cluster. Then, we compute:

$$\mathbf{T}_c = \mathbf{A}_c^T \mathbf{A}_c \quad (12)$$

A more accurate clustering can be obtained by incorporating the functional similarity between GO terms. As an illustrative example, consider two genes with the following annotation profiles: $\mathbf{a}_{i_1}^T = [1010]$, $\mathbf{a}_{i_2}^T = [1100]$. Their similarity score as computed by (10) is equal to 1, regardless of the fact that the second and third terms might represent functionally similar GO concepts. Therefore, 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:

$$\tilde{\mathbf{G}}(i_1, i_2) = \mathbf{a}_{i_1}^T \mathbf{S} \mathbf{a}_{i_2} = \sum_{j_1=1}^n \sum_{j_2=1}^n \mathbf{A}(i_1, j_1) \cdot \mathbf{S}(j_1, j_2) \cdot \mathbf{A}(i_2, j_2) \quad (13)$$

and $\mathbf{S} \in \mathbb{R}^{n \times n}$ denotes the term similarity matrix. For the previous example, the similarity score would be equal to $\mathbf{S}(1,1) + \mathbf{S}(3,1) + \mathbf{S}(1,2) + \mathbf{S}(3,2)$, i.e. the sum of the similarity scores of all the possible combinations of terms used to annotate the two genes.

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. Edge counting methods measure the term similarity based on their distance expressed as the number of edges between the ontology nodes associated with these terms. The shorter this distance, the higher the similarity [12]. Variations may define weights for the links (edges) according to their position in the ontology [13]. Compared to edge counting methods, information-theory methods [4] have been shown to be significantly less sensitive to edge density variability, which is generally present among different branches of a bio-ontology. They consist of determining the amount of information that two terms share in common. For each term j , $p(j)$ is the probability of finding j or a descendant of j in the available annotations of the considered m genes:

$$p(j) = \frac{1}{m} \sum_{i=1}^m \mathbf{A}(i, j) \quad (14)$$

Hence, the information content of a term is equal to $-\log(p(j))$. These types of methods exploit the assumption that the more information two terms share in common, the more similar they are. Several similarity (or distance) metrics have been proposed based on this approach [14], [15]. Here we adopt the Lin's similarity metrics [15], which is defined as:

$$\mathbf{S}(j_1, j_2) = \frac{2 \log p(\text{LCA}(j_1, j_2))}{\log p(j_1) + \log p(j_2)} \quad (15)$$

where $\text{LCA}(j_1, j_2)$ is the least common ancestor of both terms j_1 and j_2 , i.e. among the ancestor terms in common to both j_1 and j_2 the one that has the least probability.

Our proposed SIM algorithm is summarized in Figure 1. In the following section we compare the SVD and SIM prediction methods.

III. SVD VS. SIM COMPARATIVE ANALYSIS

A. Evaluation Method

We assessed the performance of the SVD and SIM methods while varying the value of the threshold τ by performing K -fold cross-validation for each value of the threshold τ as suggested in [3] and [4]. First, to improve performance by discarding useless terms, we extracted a subset of the matrix \mathbf{A} columns corresponding to the terms used (directly or indirectly) to annotate at least M genes. Then, we discarded those rows corresponding to genes with no annotations left. To the \tilde{m} rows of the reduced $\tilde{m} \times \tilde{n}$ matrix $\tilde{\mathbf{A}}$, we applied

a random permutation, in order to eliminate any form of correlation between genes mapped to adjacent rows. Then, we divided the permuted matrix into $K = 10$ non-overlapping partitions, each of size $\lfloor \tilde{m}/K \rfloor \times \tilde{n}$. In each step of the K -fold cross-validation process, one partition is considered as the test set, while the remaining partitions constitute the training set. For each gene-term pair (i, j) in the test set such that $\tilde{\mathbf{A}}(i, j) = 1$, we proceeded as follows:

- Discard the annotation in position (i, j) by setting to zero the corresponding entry in the test set matrix.
- Discard all the annotations corresponding to the descendants or ancestors of term j , by setting to zero the corresponding entries in the test set matrix.
- For each annotation left of gene i to a term $l \neq j$, restore part of the discarded annotations by setting to one the entries corresponding to the ancestors of l .
- Let $\hat{\mathbf{a}}_i^T$ denote the modified annotation profile obtained after applying the previous three steps; if $\hat{\mathbf{a}}_i^T$ is non-zero (i.e. at least one annotation has been left), perform the prediction according to:

$$\hat{\mathbf{a}}_{SVD,i}^T = \hat{\mathbf{a}}_i^T \hat{\mathbf{V}} \hat{\mathbf{V}}^T \quad (16)$$

$$\hat{\mathbf{a}}_{SIM,i}^T = \hat{\mathbf{a}}_i^T \hat{\mathbf{V}}_c \hat{\mathbf{V}}_c^T \quad (17)$$

- Retain the j -th entry of the previous row vectors as the prediction scores for the annotation of gene i to term j .

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 as described in equation (15) in Section II-B. In both cases we heuristically set a fixed number of clusters $C = 5$ for all ontologies.

B. Evaluation Results

Figure 2 shows the results obtained, by varying the threshold τ , with the SVD and SIM methods applied to the prediction of GO annotations of *Saccharomyces cerevisiae* (SGD) genes, starting from annotation data available in November 2009. Let FN, FP, TN and TP indicate the number of false negative,

- 1) Compute the term functional similarity matrix \mathbf{S}
- 2) Compute the gene-similarity matrix $\tilde{\mathbf{G}} = \mathbf{A}\mathbf{S}\mathbf{A}^T$
- 3) Compute a set of eigenvectors \mathbf{u}_c , $c = 1, \dots, C$ corresponding to the C largest eigenvalues of $\tilde{\mathbf{G}}$
- 4) $\hat{\mathbf{a}}_{i,0}^T = \mathbf{a}_i^T \hat{\mathbf{V}} \hat{\mathbf{V}}^T$
- 5) for $c = 1 : C$
 - a) $\mathbf{A}_c = \mathbf{W}_c \mathbf{A}$, $\mathbf{W}_c = \text{diag}(\mathbf{u}_c)$
 - b) $\mathbf{T}_c = \mathbf{A}_c^T \mathbf{A}_c$
 - c) Compute a set of eigenvectors $\hat{\mathbf{V}}_c$ corresponding to the k largest eigenvalues of \mathbf{T}_c
 - d) $\hat{\mathbf{a}}_{i,c}^T = \mathbf{a}_i^T \hat{\mathbf{V}}_c \hat{\mathbf{V}}_c^T$
- 6) $c_i^* = \arg \min_{c=0, \dots, C} \|\hat{\mathbf{a}}_{i,c} - \mathbf{a}_i\|_2$
- 7) $\hat{\mathbf{a}}_i^T = \mathbf{a}_i^T \hat{\mathbf{V}}_{c_i^*} \hat{\mathbf{V}}_{c_i^*}^T$

Fig. 1. Overview of the SIM algorithm.

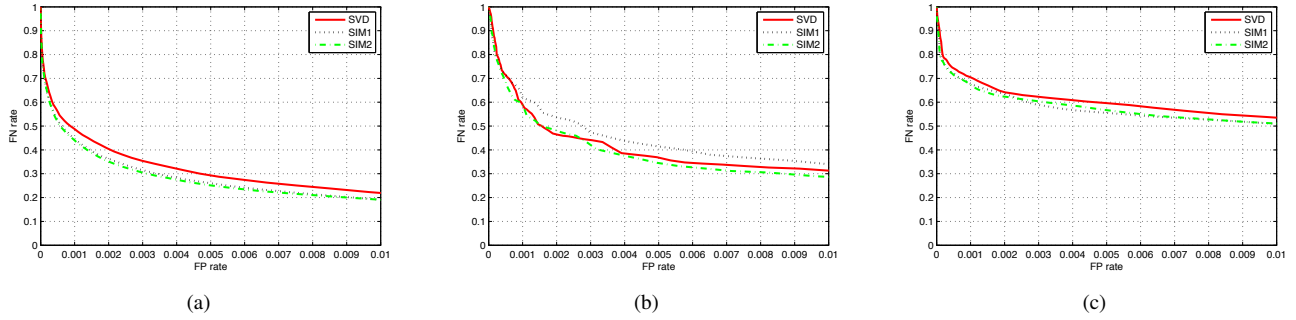


Fig. 2. False negative (FN) rate vs. false positive (FP) rate obtained by varying the threshold τ in predicting the GO annotations of *Saccharomyces cerevisiae* (SGD) genes. a) BP: Biological Process, b) MF: Molecular Function, c) CC: Cellular Component GO ontologies.

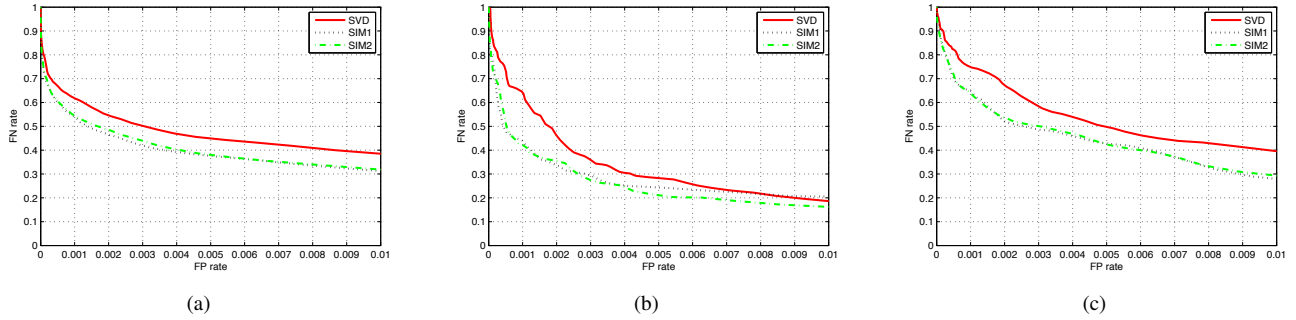


Fig. 3. False negative (FN) rate vs. false positive (FP) rate obtained by varying the threshold τ in predicting the GO annotations of *Drosophila melanogaster* (FlyBase) genes. a) BP: Biological Process, b) MF: Molecular Function, c) CC: Cellular Component GO ontologies.

false positive, true negative and true positive annotations, respectively, as defined at the end of Section II-A. The curves depict the trade-off between the false negative rate (FN / (TP + FN)) and false positive rate (FP / (FP + TN)) of annotations, when the prediction is performed using the first $k = 40$ eigenvectors of the available annotation matrix $\tilde{\mathbf{A}}$ (with $k = 40$ heuristically chosen after preliminary results). In the matrix $\tilde{\mathbf{A}}$ we heuristically retained GO terms used to annotate at least $M = 3$ genes of the considered organism and excluded annotations with evidence code IEA (inferred electronic annotations). Similar results are shown in Figure 3 for the prediction of GO annotations of *Drosophila melanogaster* (FlyBase) genes. Table I shows the number of genes and terms retained for each GO ontology, i.e. the size of the matrix $\tilde{\mathbf{A}}$.

As an aggregated indicator of the prediction performance, we computed the area above the FN rate vs. FP rate curve

TABLE I
NUMBER OF ANNOTATED GENES (\tilde{m}) AND GENE ONTOLOGY TERMS (\tilde{n}) CONSIDERED FOR PREDICTION WHEN ONLY TERMS ANNOTATING AT LEAST M GENES ARE RETAINED. BP: BIOLOGICAL PROCESS, MF: MOLECULAR FUNCTION, CC: CELLULAR COMPONENT GO ONTOLOGIES.

	BP		MF		CC		
	M	\tilde{m}	\tilde{n}	\tilde{m}	\tilde{n}	\tilde{m}	\tilde{n}
SGD	3	5351	1278	4329	482	5498	313
	10		807		261		235
FlyBase	3	6731	1683	6907	594	4740	302
	10		1084		330		211

(AAC) in the $[0, 0.01]$ range, for both *Saccharomyces cerevisiae* and *Drosophila melanogaster*. In fact, we are typically interested at the low range of FP rate, since it corresponds to top-ranked predictions of newly inferred annotations (FP) with the highest score. The AAC metrics is bounded in the $[0, 1]$ interval, where a value close to 1 implies more accurate predictions. We also repeated the experiment by varying the number of retained GO terms used to annotate at least M genes, heuristically considering $M = 3$ and $M = 10$, and varying the number k of retained eigenvectors, heuristically considering $k = 20$ and $k = 40$. In all cases, we computed the AAC metrics considering only the prediction of GO terms with depth from the root of the ontology greater than either 2 or 6. By considering the latter case, we can demonstrate the capability of predicting more specific terms. We notice that the maximum (average) depth of the ontology terms used by the prediction algorithms for the two considered organisms (SGD and FlyBase) in the three GO ontologies is 11 (5.58) for SGD and 14 (5.32) for FlyBase Biological Process, 12 (4.30) for SGD and 12 (4.43) for FlyBase Molecular Function, 13 (7.47) for SGD and 12 (6.92) for FlyBase Cellular Component ontologies, respectively. For both considered organisms, most of the GO terms are distributed between a depth of 3 and 9 from the ontology root. Using the AAC metrics, Table II shows that the SIM method generally outperforms the SVD method for all GO ontologies and tested organisms: only in 10.42% of the cases the SVD method was better than, or equal to, the SIM method. In most cases, SIM2

TABLE II

AREA ABOVE THE CURVE OF FALSE NEGATIVE (FN) RATE VS. FALSE POSITIVE (FP) RATE OF THE GO ANNOTATIONS AT GO LEVEL GREATER THAN 2 (L2) AND 6 (L6) PREDICTED WITH DIFFERENT METHODS WHEN ONLY TERMS ANNOTATING AT LEAST M GENES ARE CONSIDERED FOR PREDICTION. A) SGD, B) FLYBASE; BP: BIOLOGICAL PROCESS, MF: MOLECULAR FUNCTION, CC: CELLULAR COMPONENT GO ONTOLOGIES; K: NUMBER OF EIGENVECTORS OF THE CONSIDERED ANNOTATION MATRIX RETAINED FOR PREDICTION.

(a)									(b)								
M	method	k	BP		MF		CC		M	method	k	BP		MF		CC	
			L2	L6	L2	L6	L2	L6				L2	L6	L2	L6	L2	L6
3	SVD	20	0.58	0.35	0.47	0.51	0.39	0.50	3	SVD	20	0.56	0.50	0.57	0.65	0.37	0.46
		40	0.65	0.57	0.57	0.60	0.32	0.51			40	0.60	0.56	0.63	0.73	0.37	0.54
	SIM1	20	0.64	0.53	0.56	0.64	0.41	0.60		SIM1	20	0.60	0.54	0.59	0.70	0.39	0.56
		40	0.70	0.61	0.52	0.61	0.37	0.56			40	0.66	0.64	0.63	0.74	0.35	0.55
	SIM2	20	0.64	0.50	0.59	0.70	0.37	0.56		SIM2	20	0.60	0.51	0.61	0.73	0.40	0.54
		40	0.71	0.62	0.55	0.62	0.36	0.60			40	0.67	0.66	0.64	0.74	0.34	0.55
10	SVD	20	0.53	0.34	0.43	0.49	0.35	0.43	10	SVD	20	0.49	0.45	0.51	0.57	0.31	0.40
		40	0.60	0.53	0.53	0.59	0.31	0.47			40	0.58	0.56	0.59	0.70	0.33	0.47
	SIM1	20	0.62	0.52	0.50	0.56	0.43	0.56		SIM1	20	0.58	0.56	0.54	0.65	0.38	0.45
		40	0.65	0.60	0.46	0.39	0.39	0.56			40	0.65	0.66	0.47	0.60	0.32	0.50
	SIM2	20	0.63	0.52	0.54	0.65	0.37	0.53		SIM2	20	0.56	0.53	0.57	0.65	0.38	0.46
		40	0.67	0.58	0.49	0.47	0.35	0.56			40	0.65	0.62	0.59	0.70	0.38	0.52

outperforms SIM1 (in 53.12% 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, including *Homo sapiens*, showing comparable results.

IV. CONCLUSIONS

In this paper we propose to extend and enhance an annotation prediction method, based on the computation of the SVD of the gene-term matrix, by taking into account clustering of genes (or gene products) based on their annotation profiles. Experimental results on the GO annotations of the genes of SGD, FlyBase and other organisms confirm the effectiveness of our proposed approach. Future work will address advantages and issues related to the annotation prediction by considering all GO ontologies jointly instead of independently, in order to take advantage of potential correlations existing between them. Furthermore, since our approach is not bounded to the GO but can be applied to any ontological annotations, increasingly available multiple annotations of genes and gene products from different ontologies could be jointly considered to further improving prediction reliability. So far, results have been validated using an objective metrics computed by means of cross-validation. We plan to further verify the effectiveness of the proposed methods by assessing the quality of the top-ranked predictions by means of an expert.

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] Z. Barutcuoglu, R. E. Schapire, and O. G. Troyanskaya, "Hierarchical multi-label prediction of gene function," *Bioinformatics*, vol. 22, no. 7, pp. 830–836, 2006.
- [7] 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.*, vol. 12, no. 1, pp. 203–214, 2002.
- [8] A. Perez, C. Perez-Iratxeta, P. Bork, G. Thode, and M. A. Andrade, "Gene annotation from scientific literature using mappings between keyword systems," *Bioinformatics*, vol. 20, no. 13, pp. 2084–2091, 2004.
- [9] R. Kustra and A. Zagdanski, "Incorporating gene ontology in clustering gene expression data," in *Proc. of 19th IEEE Symposium on Computer-Based Medical Systems*, 2006, pp. 555–563.
- [10] B. Adryan and R. Schuh, "Gene-Ontology-based clustering of gene expression data," *Bioinformatics*, vol. 20, no. 16, pp. 2851–2852, 2004.
- [11] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, "Clustering Large Graphs via the Singular Value Decomposition: Theoretical Advances in Data Clustering," *Machine Learning*, vol. 56, pp. 9–33, July 2004.
- [12] P. Resnik, "Using information content to evaluate semantic similarity," in *Proc. of the 19th International Joint Conference on Artificial Intelligence*, Montreal, Canada, 1995, pp. 448–453.
- [13] J. Zhong, H. Zhu, Y. Li, and Y. Yu, "Conceptual graph matching for semantic search," in *Proc. of Conceptual Structures: Integration and Interfaces*, London, UK, 2002, pp. 92–106.
- [14] P. Resnik and M. Diab, "Measuring verb similarity," in *Proc. of 22nd Annual Meeting of the Cognitive Science Society*, Philadelphia, USA, 2000, pp. 399–404.
- [15] D. Lin, "An information-theoretic definition of similarity," in *Proc. of 15th International Conference on Machine Learning*, San Francisco, USA, 2000, pp. 296–304.