

Anomaly-free Prediction of Gene Ontology Annotations using Bayesian Networks

Marco Tagliasacchi and Marco Masseroli
*Dipartimento di Elettronica e Informazione
Politecnico di Milano
Milano, Italy*

marco.tagliasacchi@polimi.it, marco.masseroli@polimi.it

Abstract

Gene and protein structural and functional annotations expressed through controlled terminologies and ontologies are paramount especially for the aim of inferring new biomedical knowledge through computational analyses. However, the available annotations are incomplete, in particular for recently studied genomes, and only a few of them are highly reliable human curated information. To support and speed up the time-consuming curation process, prioritized lists of computationally predicted annotations are hence extremely useful. In this paper we leverage 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, and we propose a novel post-processing method that uses a Bayesian network to eliminate predictions of anomalous annotations. In fact, we observed that the predicted annotation profiles might suggest that a gene shall be annotated to a term, but not to one of its ancestors, thus violating the constraint imposed by the Gene Ontology. To this end, the proposed algorithm processes the annotation profiles predicted by a SVD based method, and produces a ranked list of computationally discovered candidate annotations which is consistent with the Gene Ontology.

1. Introduction

2. Introduction

New approaches in molecular biology, particularly high-throughput microarray technologies, allow quickly and simultaneously studying thousands of genes and proteins. Such technologies are providing unprecedented amount of valuable data that foster the increasing relevance of molecular medicine in

health care research and practice. 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 orthogonal ontologies that hold a total of more than 27,200 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 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 afflict annotation databases [2]. First, the annotations are not exhaustive: only a subset of genes of sequenced organisms are 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 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 and assess the relevance of inferred annotations, or predict missed annotations with high reliability, are manifold. 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 their method provides the basis for the work presented in this paper, it will be subsequently summarized in Section 3.

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 keywords extracted from published papers are mapped to GO concepts. The reader can refer to [9] for an extensive survey on computational approaches for functional prediction of genes and gene products. Most of the methods described in [9] are applicable to newly sequenced genomes for which there are no available annotations. Conversely, in our work we focus on methods for predicting new candidate annotations for partially annotated genomes, by observing that the latter can be incomplete or contain incorrect annotations.

By providing a, possibly ranked, list of annotations to be checked by a manual curator, the aforementioned techniques can drive the discovery of previously un-

known 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 [10] and [11].

In this paper we propose a post-processing method that can be applied to the output of the method described in [5], hereafter denoted SVD method. We propose an algorithm to fix the issue related to the existence of anomalous predictions, i.e. a gene being annotated to a GO term but, at the same time, not with some of the term ancestors. We accomplish this by explicitly leveraging the semantic relationships between terms as expressed by the GO DAG. We construct a Bayesian network based on the GO topology and we exploit the output of the SVD method as prior evidence. For each gene-term pair, the output of the proposed algorithm is the a-posteriori probability that such gene is annotated to that term. Our results demonstrate that with such post-processing anomalous annotations are virtually eliminated. Although we consider here only the annotations based on GO terms, the framework can be straightforwardly extended to handle multiple ontologies as well as predictions obtained based on multiple data sources.

The rest of this paper is organized as follows. Section 3 illustrates SVD method, since it represents the starting point for our work. In Section 4 we introduce the problem of anomalous annotations. We propose a method based on Bayesian networks to remove them in Section 4.1 and we present and discuss our results in Section 4.2. Section 5 concludes the paper and provides some guidelines for future research on this topic.

3. SVD prediction method

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

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

Annotation curators are asked to always use the most specific GO term for a given functional category. As such, when a gene is annotated to a term, it is implicitly

assumed to be annotated also to the more generic terms for that category, i.e. all the ancestors in the GO DAG. As such, let $\tilde{\mathbf{A}}$ denote a modified gene-to-term matrix, where the assignment of its entries is given by:

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

The i -th row of the matrix $\tilde{\mathbf{A}}$ 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 annotation unfolding.

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

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

where \mathbf{U} is a $m \times p$ unitary matrix (i.e. $\mathbf{U}^T\mathbf{U} = \mathbf{I}$), Σ is a non-negative diagonal matrix of size $p \times p$, \mathbf{V} is a $n \times p$ unitary matrix, where $p = \min(m, n)$. Conventionally, the entries along the diagonal of Σ (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 $\tilde{\mathbf{A}}$. For any positive integer $k < r$, it is possible to generate a matrix $\tilde{\mathbf{A}}_k$, with:

$$\tilde{\mathbf{A}}_k = \mathbf{U}_k \Sigma_k \mathbf{V}_k^T \quad (4)$$

where \mathbf{U}_k (\mathbf{V}_k) is a $m \times k$ ($n \times k$) matrix obtained retaining the first k columns of \mathbf{U} (\mathbf{V}) and Σ_k is a $k \times k$ diagonal matrix with the k largest singular values along the diagonal. $\tilde{\mathbf{A}}_k$ is the optimal rank- k approximation of $\tilde{\mathbf{A}}$, i.e. the one that minimizes the norm (either the spectral norm or the Frobenius norm) $\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_k\|$ subject to the rank constraint.

In [5] it is argued that the study of the matrix $\tilde{\mathbf{A}}_k$ reveals semantic relationships of the gene-function associations. A large value of $\tilde{\mathbf{A}}_k(i, j)$ 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 $\tilde{\mathbf{A}}$ 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 $\tilde{\mathbf{A}}_k$ can be used to predict gene-to-term annotations, we point out that an alternative expression of (4) is obtained by basic linear algebra manipulations:

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

Moreover, the SVD of the matrix $\tilde{\mathbf{A}}$ is related to the eigen-decomposition of the symmetric matrices

$\mathbf{T} = \tilde{\mathbf{A}}^T\tilde{\mathbf{A}}$ and $\mathbf{G} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^T$. In fact, the columns of \mathbf{V}_k (\mathbf{U}_k) 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, $\mathbf{T}(p, q)$ is the number of times that term p and q are used to annotate the same gene in the existing annotation profile. Therefore, $\mathbf{T}(p, q)$ expresses the (unnormalized) correlation between term pairs and it can be interpreted as a similarity score of the terms p and q computed solely based on the use of these terms in available annotations. The eigenvectors of \mathbf{T} (i.e., the columns of \mathbf{V}_k) 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 $\tilde{\mathbf{A}}_k$ can be written as

$$\mathbf{a}_{k,i}^T = [\mathbf{a}_i^T \mathbf{V}_k] \mathbf{V}_k^T \quad (6)$$

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 \mathbf{V}_k , and then mapped back to the original domain by means of \mathbf{V}_k^T . This corresponds to projecting the original vector \mathbf{a}_i^T onto the k -dimensional subspace spanned by the columns of \mathbf{V}_k .

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

- If $\tilde{\mathbf{A}}(i, j) = 1$ and $\tilde{\mathbf{A}}_k(i, j) > \tau$, the annotation of gene i to term j is confirmed; this case is denoted as a true positive (TP).
- If $\tilde{\mathbf{A}}(i, j) = 0$ and $\tilde{\mathbf{A}}_k(i, j) > \tau$, a new annotation is suggested; this case is denoted as a false positive (FP).
- If $\tilde{\mathbf{A}}(i, j) = 1$ and $\tilde{\mathbf{A}}_k(i, j) \leq \tau$, an existing annotation is suggested to be semantically inconsistent with the available data; this case is denoted as a false negative (FN).
- If $\tilde{\mathbf{A}}(i, j) = 0$ and $\tilde{\mathbf{A}}_k(i, j) \leq \tau$, 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).

4. Removing anomalous predictions

As a matter of fact, the SVD method described in Section 3 predicts GO annotation profiles \mathbf{A}_k that, for some values of the threshold τ , might contain inconsistencies with respect to the GO structure. In fact, it might predict that gene i should be annotated to term j but, at the same time, that it should not be

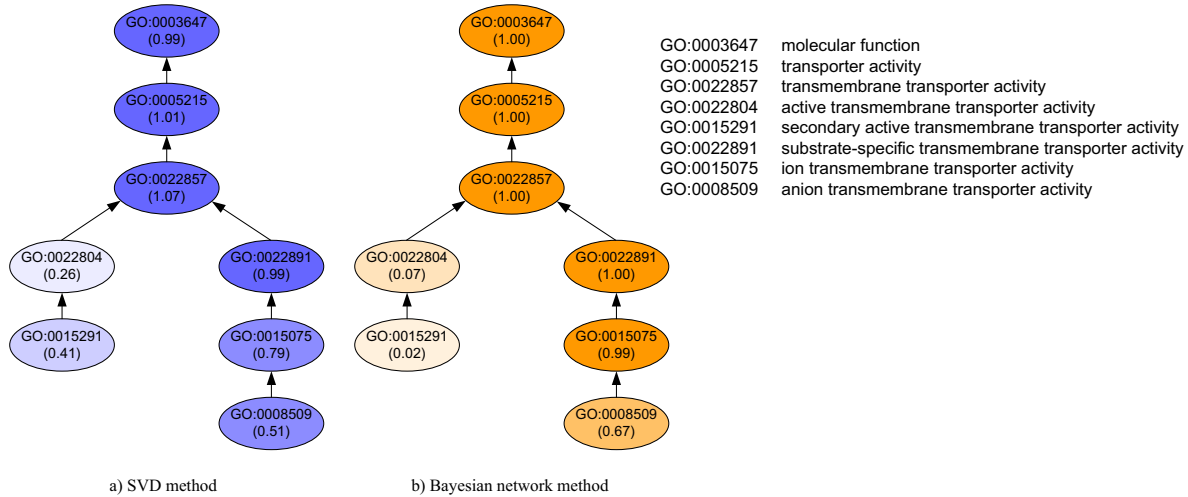


Figure 1. Example of anomaly of the SVD predicted annotation for gene S000000498 and its correction with the proposed Bayesian network method. Within each depicted Gene Ontology term node, the a-posteriori probability of the S000000498 gene to be annotated to that term is reported in brackets.

annotated to some of the ancestors of term j . Also, interpreting the real-valued output $\mathbf{A}_k(i, j)$ as a ranking score of the likelihood of gene i being annotated to term j , it might happen that $\mathbf{A}_k(i, j) > \mathbf{A}_k(i, r)$, i.e. the annotation of gene i to term j is more likely than to term r , although r is an ancestor of j . This obviously represents an anomaly that should be avoided in any prediction algorithm. The reason for this behavior can be imputed to the specific use of the GO structure in the SVD method. In fact, the GO DAG is taken into account only to perform the annotation unfolding *before* applying the SVD to the resulting matrix $\tilde{\mathbf{A}}$.

As an illustrative example, Figure 1 shows the predicted annotation profile for gene S000000498 of SGD. We observe that, for $0.26 < \tau < 0.41$, the SVD method suggests that the gene should be annotated to term GO:0015291 (*secondary active transmembrane transporter activity*) but, at the same time, it should not be annotated to its parent term GO:0022804 (*active-transmembrane transporter activity*), thus violating the GO structure constraints.

4.1. Bayesian network method

As our novel contribution, we propose an algorithm aimed at fixing the anomalous behavior illustrated above by imposing on the predicted annotations the constraints dictated by the GO structure. To this end, we designed a Bayesian network that helps enforcing consistent annotations by analyzing each gene i (a row of the matrix $\tilde{\mathbf{A}}_k$) independently. The proposed method

is similar to the approach presented in [6], where a Bayesian network is used to enforce consistency of annotation profiles inferred from microarray gene expression data by means of term-independent SVM classifiers. Unlike [6], in our work the input data consists of the annotation profiles predicted by the SVD method.

Bayesian networks are powerful computational tools suitable for representing conditional probability statements between statistically dependent random variables in the form of a graph. Each node in the graph represents a random variable (either discrete or continuous valued) and a directed edge from node p to node q indicates that q is statistically dependent from p . In our context, let t_j , with $j = 1, \dots, n$, denote a node corresponding to the j -th term in the GO. Such nodes are discrete valued and for a given gene i they can assume the values of 1, if term t_j is used to annotate the gene i , or 0, otherwise. An edge exists from node t_p to t_q if the former is a child of the latter, based either on a *is a* or *part of* relationship. In fact, if a child node t_p is annotated to a gene i , also its parent node t_q is annotated to that gene i . For each node t_j we created a continuous valued node e_j and an edge from t_j to e_j , as depicted in Figure 2. We call these e_j nodes evidence nodes.

In order to complete the specifications of the Bayesian network, we needed to determine the conditional probabilities relating the random variables. For each node t_j and gene i , we specify the probability of t_j being either 1 or 0 (i.e., used to annotate gene i or not),

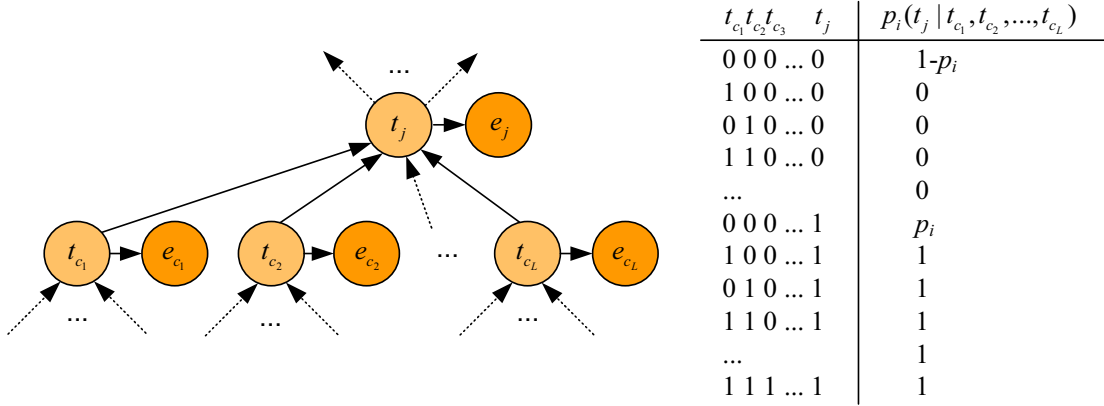


Figure 2. Structure of our designed Bayesian network. Light shading indicates binary-valued nodes, dark shading real-valued evidence nodes.

conditioned on the values assumed by its children:

$$p_i(t_j | t_{c_1}, t_{c_2}, \dots, t_{c_L}) \quad (7)$$

with L the number of children of node t_j . In general, this can be stored as a table with 2^{L+1} rows. In our context, regardless of the specific node t_j , such table has a generic structure dictated by the GO. Figure 2 shows an example for a generic node t_j . If any of the child node is equal to 1 (i.e., the corresponding GO term is used to annotate gene i), then $p_i(t_j = 1 | t_{c_1}, t_{c_2}, \dots, t_{c_L}) = 1$ and, consequently, $p_i(t_j = 0 | t_{c_1}, t_{c_2}, \dots, t_{c_L}) = 0$. In fact, the annotation to a term is unfolded into the annotation profile containing all its ancestor terms, thus including t_j . If none of the children is equal to 1 (i.e., used to annotate gene i), nothing can be said about the annotation of gene i to term t_j from a deterministic point of view. Rather, we set $p_i(t_j = 1 | t_{c_1} = 0, t_{c_2} = 0, \dots, t_{c_L} = 0) = p_{i,j}$, where the value of $p_{i,j}$ can be estimated from available data. We computed $p_{i,j}$ as a function of the number L of the child nodes of node t_j . However we found that a fixed value of $p_{i,j} = 0.005, \forall i, j$ fits reasonably the GO annotation data available for the tested organisms, regardless of the value of L .

We also needed to specify the conditional probabilities relating nodes e_j to t_j , i.e. $p_i(e_j | t_j)$. In our algorithm, each e_j node corresponds to the real valued output of the SVD method for a specific gene i . The conditional probability is modeled as a Gaussian mixture model (GMM) with two components:

$$p_i(e_j | t_j = 0) = w_{0,0}N(0, \sigma_{0,0}^2) + w_{0,1}N(0, \sigma_{0,1}^2) \quad (8)$$

$$p_i(e_j | t_j = 1) = w_{1,0}N(1, \sigma_{1,0}^2) + w_{1,1}N(1, \sigma_{1,1}^2) \quad (9)$$

where the values of the parameters are estimated using the Expectation-Maximization algorithm, independently from the gene i , i.e. $p_i(e_j | t_j) = p(e_j | t_j), \forall i$, using as training data the (e_j, t_j) pairs obtained by extracting the j -th elements from all row vectors $\mathbf{a}_{k,i}^T$ and \mathbf{a}_i^T respectively.

Once the Bayesian network is specified, it can be used for the inference of the annotations. For each gene i , the input is represented by the real-valued annotation profile produced by the SVD method, i.e. $\mathbf{a}_{k,i}^T$. This is imposed as prior evidence by setting $\mathbf{e}_i^T = [e_{i,1}, \dots, e_{i,j}, e_{i,n}]^T = \mathbf{a}_{k,i}^T$. By running the junction-tree inference algorithm [12], the proposed method produces the following output for each gene i :

- The a-posteriori marginal probability $p_i(t_j = 1)$ for each term node t_j , given the evidence and the conditional probability constraints imposed by the Bayesian network.
- The a-posteriori most probable explanation (MPE), i.e. the binary values of the t_j nodes corresponding to a mode of the joint probability density function $p_i(t_1, \dots, t_n)$, i.e. the most likely annotation profile given the evidence and compatible with the constraints imposed by the Bayesian network.

Both outputs can be useful in their respect. The values of the marginal probabilities $p_i(t_j = 1)$ can be used as a ranking score indicating the likelihood of gene i being annotated to term j , as for the SVD method, but with much less likely anomalies. In addition, thresholding can be applied in order to have an unordered list of predictions. We point out that thresholding the marginal probabilities could produce in principle anomalous annotations, although they are much less likely to

occur than in the SVD method, as it will be shown in the next section. Conversely, if one produces the predicted annotation profiles based on the generated MPE values, anomalies are ruled out by construction. In fact, an anomalous annotation would be assigned a joint probability equal to zero and thus it cannot be selected as the most probable explanation. In this case, one would not have access to an ordered list of candidate predictions though.

4.2. Results and discussion

In order to evaluate the anomaly correction performed by the proposed Bayesian network method, we considered the gene annotations of the *Saccharomyces cerevisiae* (SGD) organism, expressed through GO molecular functions terms. Figure 3(a) plots the results of the SVD method obtained, by varying the threshold τ , as counts of false positives (FP), false negatives (FN), the sum FP + FN and the count of anomalies detected. In this case we considered the full set of available annotations as input of the SVD method. We confined our analysis to GO terms that are used to annotate (directly or indirectly) at least 10 SGD genes, since the SVD method estimates the similarity between GO terms based on the count of co-occurrences. Furthermore, we excluded annotations with evidence code IEA (inferred electronic annotations) from the matrix $\tilde{\mathbf{A}}$, since they have not been checked by a manual curator. Consequently, the matrix $\tilde{\mathbf{A}}$ contained a total number of 33090 (unfolded) annotations of $m = 4333$ genes and $n = 261$ terms. Then, the reduced-rank approximation has been computed setting $k = 40$. All results represented in Figure 3(a) were obtained using the same approach as in [5], i.e. FP and FN were counted by considering the existing annotations as ground truth data besides as input of the prediction method. According to [5], the value of the threshold τ yielding the minimum FP + FN global error can be chosen to practically provide a list of predicted annotations that minimize differences with respect to the available annotations. In our considered case, such threshold value was $\tau = 0.51$, yielding 404 FP, i.e. newly predicted annotations, (1.2% of the initial annotations) and 3657 FN (11% of the initial annotations). Figure 3(b) shows analogous results for the proposed Bayesian network method. We notice that in terms of FP and FN, the proposed method produces results comparable to the SVD method, but it is more robust to the specific choice of the threshold τ , providing consistent results for a wide range of threshold values. By empirically setting the threshold to the value achieving the lowest FP + FN count, i.e.

$\tau = 0.55$, we obtained 590 FP, i.e. newly predicted annotations (1.8% of the initial annotations) and 3445 FN (10% of the initial annotations). Figure 3 also shows the count of anomalies for both methods, which was obtained as follows: for each value of the threshold τ we checked the consistency of all gene-term pairs (i, j) such that $\tilde{\mathbf{A}}_k(i, j) > \tau$. If, for any of the ancestors r of the GO term j , $\tilde{\mathbf{A}}_k(i, r) < \tau$, the counter of anomalies is increased by one. We notice that the number of anomalies for the SVD method depends heavily on the choice of the threshold τ , whereas it is consistently equal to zero for the Bayesian network method.

It is interesting to look at the anomaly rate with respect to the FP rate, depicted in Figure 4 by varying the threshold τ . The figure was obtained by dividing the count of both anomalies and FP by the total number of negative annotations (i.e., TN + FP). We are typically interested at the low range of FP rate, since it corresponds to top-ranked predictions corresponding to newly inferred annotations (FP) with the highest score. For the SVD method, we observe that at a FP rate equal to 0.01, the anomaly rate is 0.0011, i.e. 11% (2160) of the predicted annotations are inconsistent with the GO structure. This fraction drops at 7.5% (710) at a FP rate equal to 0.005 and 1.8% (35) at a FP rate equal to 0.001. If we restrict our analysis to the predicted annotations with no descendants (Figure 4(b)), i.e. the most specific GO terms predicted to be used to annotate the considered genes, the anomaly rate is equal to 6.5% (1290), 5% (490), and 1.7% (31) at, respectively, a FP rate equal to 0.01, 0.005, and 0.001.

By applying the proposed Bayesian network method as a post-processing after the SVD method, Figure 4 shows that we can drastically reduce the number of anomalous annotations with respect to the SVD method, being them identically equal to zero for all FP rates. This suggests that, for any value of the threshold τ , the annotations predicted by the Bayesian network method are more likely correct than those predicted by the SVD method, since they do not suffer from the problem of anomalies. Furthermore, the real valued marginal probabilities $p_i(t_j = 1)$ can be used to produce a ranked list of predicted terms to be used to annotate gene i , as in the SVD method. Nevertheless, with the correction imposed by the Bayesian network, the annotation of a gene i to a GO term j will never appear higher in ranking order than any of the annotations of gene i to the ancestors of term j , thus avoiding to introduce any anomalous. This can support the manual curator in the choice of the most reliable predicted annotations to be checked. In fact, such list can constitute a prioritized ranked list of more likely annotations, in particular when it is created starting

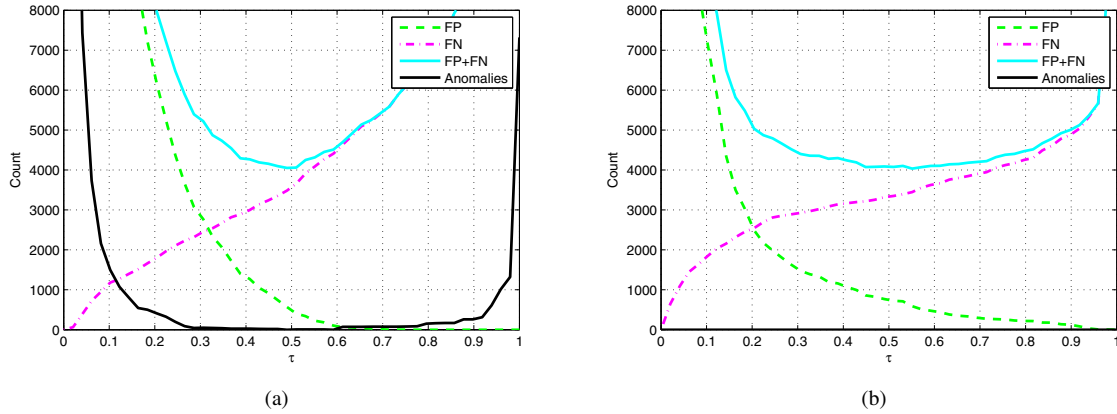


Figure 3. False positives (FP), false negatives (FN) and their sum FP + FN vs. threshold τ in predicting the GO molecular functions annotations of *Saccharomyces cerevisiae* genes. a) SVD method, b) Bayesian network method. Notice that in b) the line representing the count of anomalies is overlapped to the x-axis.

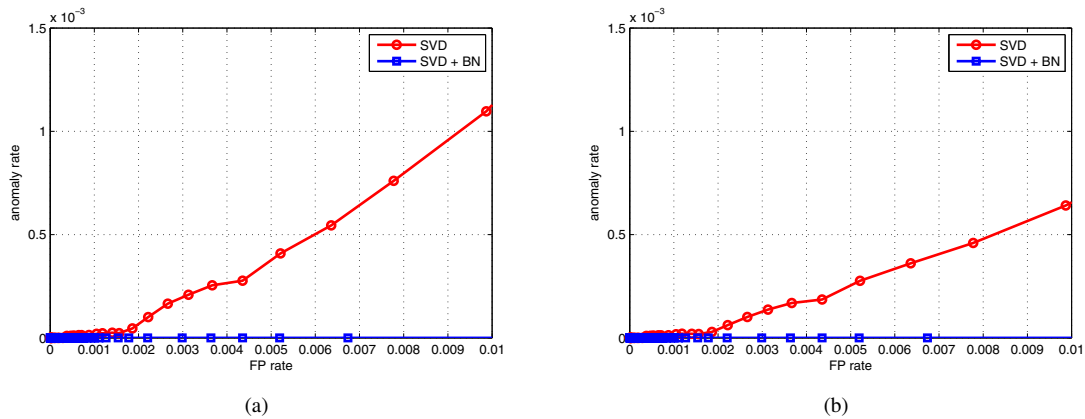


Figure 4. Anomaly rate vs. FP rate of predicted Gene Ontology molecular functions annotations for *Saccharomyces cerevisiae* genes. SVD: singular value decomposition method; SVD + BN: SVD method corrected with our Bayesian network method; a) Anomaly rate considering all predicted annotations. b) Anomaly rate considering only predicted annotations corresponding to the most specific terms for a given GO category.

only from the currently most reliable GO annotations, i.e. those having evidence code different from "inferred electronic annotation" (IEA). For practical purposes, a reasonable lower bound of the value of τ to generate this ranked list can be determined by choosing the value of the threshold τ yielding the minimum FP + FN quantity [5]. With respect to the example shown in Figure 1, we observe that with the proposed Bayesian network method the anomaly reported by the SVD method is avoided. In fact, the a-posteriori marginal probability that gene S00000498 is annotated to term GO:0015291 is smaller than the probability of the gene to be annotated to the parent term GO:0022804.

For performance comparison with the Bayesian net-

work method, we considered also a simpler method to remove anomalies from the annotation profiles predicted with the SVD method. For each gene, we processed the GO terms from the root to the leaves and we replaced the score assigned to a term by the SVD method with the minimum among the scores assigned to its ancestor and the term itself. This strategy guarantees to prevent anomalies by construction, when thresholding is applied. Nevertheless, the quantity FP + FN turns out to be consistently larger than the one obtained by the SVD method before or after the Bayesian network post-processing, thus suggesting the inefficiency of this simple approach.

5. Conclusions

In this paper we propose a novel contribution in the context of prediction of genomic ontological annotations: we describe a post-processing algorithm based on a Bayesian network that eliminates the issue of anomalous predictions, i.e. predicted annotations that are inconsistent with the ontology structure, present in the original SVD method [5]. 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.

References

- [1] T. G. O. 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 T. O. G., "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] G. Pandey, V. Kumar, and M. Steinbach, "Computational approaches for protein function prediction: A survey," Department of Computer Science and Engineering, University of Minnesota, Tech. Rep., 2006.
- [10] 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.
- [11] B. Adryan and R. Schuh, "Gene-Ontology-based clustering of gene expression data," *Bioinformatics*, vol. 20, no. 16, pp. 2851–2852, 2004.
- [12] C. Bishop, *Pattern Recognition and Machine Learning*, ser. Information Science and Statistics. Springer, 2006.