Enhanced Probabilistic Latent Semantic Analysis with Weighting Schemes to Predict Genomic Annotations

Pietro Pinoli, Davide Chicco and Marco Masseroli

Abstract— **Genomic annotations with functional controlled terms, such as the Gene Ontology (GO) ones, are paramount in modern biology. Yet, they are known to be incomplete, since the current biological knowledge is far to be definitive. In this scenario, computational methods that are able to support and quicken the curation of these annotations can be very useful. In a previous work, we discussed the benefits of using the Probabilistic Latent Semantic Analysis algorithm in order to predict novel GO annotations, compared to some Singular Value Decomposition (SVD) based approaches. In this paper, we propose a further enhancement of that method, which aims at weighting the available associations between genes and functional terms before using them as input to the predictive system. The tests that we performed on the annotations of human genes to GO functional terms showed the efficacy of our approach.**

I. INTRODUCTION

In the recent years we have been witnessing an exponential growth in biomedical and biomolecular information, with a large amount of data becoming available for investigation. Nowadays the focus is not only on generating new data, but also in storing, managing, querying, searching and analysing them efficiently and effectively. Several solutions to express and store the current biomedical knowledge have been proposed. Among them, those that describe it in a controlled and computable form are the most valuable. In particular the associations of a gene (or gene product) with one or more controlled vocabulary terms, which describe its functional properties, are paramount to perform their in silico analysis and to be able to interpret experimental results. Some consortia maintain a list of controlled annotation terms and the sets of gene and proteins associated with them; we refer to those associations as functional annotations. Often, semantic relations between terms are also provided (i.e. terms are organized in an ontological structure); the most widely used ontology for annotating genes and proteins is the Gene Ontology (GO) [1]. It consists of three sub-ontologies describing species independent Biological Processes (BP), Molecular Functions (MF) and Cellular Components (CC).

With the progression of the biomedical knowledge, new annotations are continuously added or revised, and therefore annotation databases are usually neither complete nor accurate. In this scenario, methods that are able to improve the available annotations, both in coverage and quality, are very useful. In the recent years several of such approaches have been proposed [2]; some of them take advantage of data from heterogeneous sources, implementing a sort of annotation transfer, while other methods are solely based on previously available annotations. Among the latter class of them, a noteworthy method is the one proposed by Draghici et al. [3], which is based on the Singular Value Decomposition (SVD). Draghici and colleagues also subsequently showed how it is possible to improve the performances of the SVD-based annotation prediction system in term of precision, by weighting the associations between genes and terms [4]. In 2012, we proposed an annotation prediction algorithm based on the Probabilistic Latent Semantic Analysis (pLSA) [5], named pLSAnorm [6], that overcomes the traditional methods. In this paper, we show how to further improve pLSAnorm by weighting the associations between genes and terms.

After this introduction, the paper is organized as follows: in Section II we describe the pLSA-based predictive system that we improve; Section III is about the schemes that we proposed to use to weight the associations between genes and functional terms; in Section IV we describe the validation procedure that we used to assess the quality of our method; the results are reported in Section V; finally, in Section VI we draw some conclusions.

II. PREDICTIVE SYSTEMS

A. *Probabilistic Latent Semantic Analysis*

Although pLSA is a very general method, for the sake of clarity, we introduce it within the field of functional genomic annotations. Consider a set of genes G and a set of functional terms F. Let $\mathbf{A} \in \{0,1\}^{|G|\times|F|}$ be the annotation matrix, where each row represents a gene of G and each column a term of F. An entry $\mathbf{A}(g,f)$ is equal to 1 if the gene g is annotated to the term f, or to any of its descendants (whether the controlled vocabulary is organized into an ontology); otherwise, it is set to 0.

The aim of the pLSA algorithm is to compute a matrix $\mathbf{X} \in [0,1]^{|G|\times|F|}$ where each entry $\mathbf{X}(g,f)$ provides an estimation of the probability of the event $\mathbf{A}(g,f) = 1$. The model on which pLSA is build is named *aspect model* and is depicted in Fig. 1. It is a latent variable model for general co-occurrence data. The core of the model is the set of latent variables, called *topics*, $T = \{t_1, t_2, \ldots, t_T\}$.

A multinomial distribution γ over the gene set gives the probability of choosing a gene within the corpus, $\gamma(x) = P(g_x)$. A set of multinomial distributions $\Delta = \{\delta_1, \ldots, \delta_G\}$,

Manuscript received July 30, 2013. This research is part of the Search Computing project (2008-2013) funded by the European Research Council (ERC), IDEAS Advanced Grant.

P. Pinoli, D. Chicco and M. Masseroli are with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy pinoli@elet.polimi.it, davide.chicco@elet.polimi.it masseroli@elet.polimi.it

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

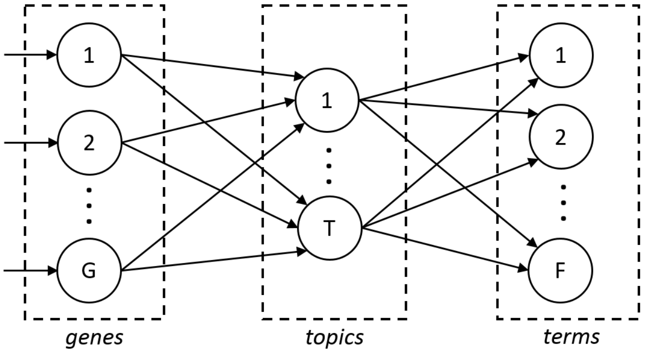

Fig. 1. Graphical representation of the aspect model

gives the conditional probability of a topic given a gene, such that $\delta_x(k) = P(t_k|g_x)$. Finally, a set of probability distributions (again multinomial) $\Phi = \{\phi_1, \ldots, \phi_T\}$ represents for each topic t and for each functional term f the probability of choosing the f given t, that is $\phi_t(f) = P(f|t)$.

Given an aspect model, the annotation probability between a gene g and a term f can be interpreted as follows:

1) Select gene g with probability $\gamma(g)$
2) Select a topic t with probability $\delta_g(t)$
3) Select the term f with probability $\phi_t(f)$

Translating the process into joint probabilities results in the expression:

$$\mathbf{X}(g,f) = P(g)P(f|g) = P(g)\sum_{t\in T} P(f|t)P(t|g) \qquad (1)$$

In pLSA all the distribution probabilities, γ and the ones in the sets Δ and Φ, are randomly initialized; then the Expectation Maximization (EM) [7] algorithm is run in order to find a set of values to be assigned to those probabilities that corresponds to a (local) maximum in the log-likelihood function:

$$\mathcal{L} = \sum_{g\in G}\sum_{f\in F} \mathbf{A}(g,f)\log P(g,f) \qquad (2)$$

B. pLSAnorm

In the pLSA variant that we proposed, named pLSAnorm [6], we introduced two changes: firstly, we consider the probability distribution γ over the genes to be uniform, such that $\forall g \in G, \gamma(g) = P(g) = 1/|G|$.

Secondly, we do not consider the $\mathbf{X}$ as the final result, but we normalize each row of the $\mathbf{X}$ matrix with respect to the row maximum. In practice, we apply the following normalization procedure:

1) $\forall g \in G, M_g = max_{f\in F}\mathbf{X}(g,f)$
2) $\hat{\mathbf{X}}(g,f) = \mathbf{X}(g,f)/M_g$

This last step is meant to reduce the bias due to the different number of terms annotated to different genes; indeed, for each row of the $\mathbf{X}$ matrix corresponding to the gene g, the property $\sum_{f\in F} P(g)P(f|g) = 1$ holds. Therefore, given the fact that the distribution probability γ is uniform, in a gene associated with a large number of functional terms, the entries of the matrix $\mathbf{X}$ corresponding to the annotated terms will have an average value lower than the ones related to a gene associated with fewer terms. This bias leads to difficulties in discerning predicted annotations; difficulties that are overcome with the normalization. Finally, notice that even if $\hat{\mathbf{X}}$ loses all the statistical properties of the matrix $\mathbf{X}$, this does not introduce any kind of problem in this particular application of the latent semantic analysis.

C. Anomaly Correction

Since GO terms are organized into ontologies, when a gene is annotated to a term describing one of its features, it must also be annotated to the ancestors of that term. This is satisfied in the $\mathbf{X}$ matrix, but not necessarily in the $\hat{\mathbf{X}}$ matrix, where the predicted annotation value between a gene g and a term f could be larger than the predicted annotation value between g and any of the f ancestors. This leads to some issues, since the system could predict an annotation between a gene and a term but not the annotation between the same gene and an ancestor of that term. In order to overcome this problem, our framework performs a further step, named *anomaly correction*. It consists in updating parent-term annotation values with the greater annotation value of their child-terms.

D. Thresholding

Once the $\hat{\mathbf{X}}$ is computed and anomalies corrected, in order to identify predicted annotations, we binarize the $\hat{\mathbf{X}}$; a threshold $\tau \in [0,1]$ is chosen and a gene g is considered annotated to a certain functional term f if $\hat{\mathbf{X}}(g,f) > \tau$. We refer to those couples $< g, f >$ such that $\mathbf{A}(g,f) = 0$ and $\hat{\mathbf{X}}(g,f) > \tau$ as *annotations predicted* (AP); whereas we refer to those couples having $\mathbf{A}(g,f) = 1$ and $\hat{\mathbf{X}}(g,f) \leq \tau$ as *annotations to be reviewed* (AR).

III. WEIGHTING THE ANNOTATIONS

Previous works in the field of Information Retrieval [8] showed how the performances of a predictive system, in term of precision and recall, can be improved by moving from a binary to a real data representation. In our case, a real data representation can be obtained by weighting the associations between genes and functional terms. We defined two criteria for weighting those associations; they are inspired to the TF-IDF (*term-frequency, inverse-document-frequency*) [8], which is a popular technique in text mining that is often used as weighting factor:

- if a feature term is included in multiple paths to the ontology root from any of the ontology terms associated with a given gene, then this feature is more strongly related to that gene than a feature represented by a term of the same ontology that is included in less paths;
- if a feature is associated only with a limited number of genes, it is a good discriminator of the genes; thus, the annotation to the term representing that feature should be considered as more important, since it brings a higher amount of information.

These two criteria can be expressed by using two statistics: *term-frequency* (*tf*) and *inverse-gene-frequency* (*igf*). *Term-frequency* measures how important an annotation term is to a certain gene. For each gene g and ontological term f, $tf(g, f)$ is the number of paths to the ontology root from any term annotated to g which include f; therefore, it provides a local weight to each annotation.

Inverse-gene-frequency measures how much important an annotation to a particular term is; it provides a global weight decreasing the importance of the most common annotation terms, usually the ones very close to the ontology root, while increasing the relevance of the rare ones. For each annotation term f we can compute:

$$igf(f) = \ln \frac{|G|}{|\ \textit{genes annotated to f}\ |} \tag{3}$$

These two statistics can be combined in order to build different weighting schemes. Draghici et al. [4] proposed some of those schemes. Accordingly, we refer to each weighting schema with a three letter code: the first letter specifies the local weight used by the schema, the second letter denotes the used global weight and the last letter indicates which normalization function is applied. For example, the ATC schema uses the *Augmented* local weight (A) as local weight, *igf* (T) as global weight and *Cosine* (C) normalization. The possible combinations of weights and normalization functions described in Table I lead to nine potential weighting schemes, but only seven of them are distinct since NTM and NTC schemes are equivalent to the MTM and MTC schemes, respectively. In fact, these schemes differ only in the local weight (N vs. M), which after the M or C normalization become equal (e.g. in MTM $w_{nloc} = w_{loc}/max_F\{w_{loc}\}$; but $w_{nloc} = max\{w_{loc}/max_F\{w_{loc}\}\} = 1$, thus $w_{nloc,MTM} = tf(g, f)/max_F\{tf(g, f)\} = w_{nloc,NTM}$). Notice also that our schemes slightly differ from the ones proposed by Draghici and colleagues in [4] because of the different formulation of the statistics *tf* and *igf*.

Each schema can be applied to the annotation matrix $\mathbf{A}$ by: (a) multiplying every gene annotation by the corresponding local weight factor and by the global weight factor of the specific annotation term, and (b) normalizing the new real valued annotations by the normalization function.

By changing the annotation matrix, the weighting schemes implicitly change the log-likelihood function, (2) of the model that the EM algorithm maximizes. The new objective function is:

$$\mathcal{L}_w = \sum_{g \in G} \sum_{f \in F} \mathbf{A}_w(g, f) \log P(g, f) \tag{4}$$

where $\mathbf{A}_w$ is the weighted annotation matrix. Thus, some annotations are more relevant than others because of the larger portion of the likelihood that they affect.

IV. VALIDATION

We tested the effects of the weighting schemes on pLSAnorm in predicting new GO functional annotations from a set of available annotations of *Homo sapiens* genes to GO terms. We compared the predictions given by the weighted version of pLSAnorm with the ones given by its unweighted variant, so as to asses the actual improvement given by the weighting schemes. Furthermore, we compared the predictions made on a set of annotations available on a certain date to the set of the same annotations available several months later, in order to check how many of those predictions were confirmed in the newer annotation set. We retrieved the annotation sets from the *Genomic and Proteomic Data Warehouse* (GPDW) [9], which integrates biological information from several sources; the GPDW is updated bimonthly and old versions are stored.

In order to measure the quality of an annotation prediction system we implemented the following procedure:

1) we extracted a set of annotations of *Homo sapiens* genes to terms of an ontology from an older version of the GPDW; we filtered out the less reliable annotations (i.e. the ones with *Inferred from Electronic Annotation* (IEA) evidence, which have been inferred by a computational tool and have not been verified by any human curator);
2) by running the predictive algorithm on those annotations, we got a list of AP; for the thresholding, we used $\tau = 0.5$ as predicted annotation heuristic threshold;
3) we compared the obtained AP with the annotations present in a newer version of the GPDW and counted how many AP were confirmed, both with evidence equal or different from IEA.

In all our tests, we extracted the input set of annotation (i.e. the $\mathbf{A}$ matrix) from the GPDW version of July 2009 and checked the generated predictions against the GPDW version created 3 years and 10 months later, on May 2013. We report the dimensions of the input datasets in the upper part of Table II.

The cardinality of the topic set is the main parameter of our method and of the pLSA approach. Yet, a general procedure to estimate the best number of pLSA topics is still unknown. The more topics are used, the larger the expressivity of the aspect model is; in contrast, a high number of topics leads to overfitting problems. In our experiments we heuristically chose to use the ten percent of the number

TABLE I

WEIGHTING SCHEMES

Code	Name	Description
		Local Weight
N	No-Transformation	$\forall f, g : w_{loc} = tf(g, f)$
M	Maximum	$\forall f, g : w_{loc} = tf(g, f)/max_f tf(g, f)$
A	Augmented	$\forall f, g : w_{loc} = 0.5 + \frac{0.5 * tf(g,f)}{max_f tf(g,f)}$
		Global Weight
T	Term Weight	$\forall f : w_{glob} = igf(f)$
		Normalization
N	None	Normalization factor is not used
M	Maximum	$w_{norm}(g, f) = w(g, f)/max_f w(g, f)$
C	Cosine	$w_{norm}(g, f) = w(g, f)/\sqrt{\sum_f w(g, f)}$

TABLE II

TEST RESULTS

Dataset	*Homo sapiens* GO Biological Process			*Homo sapiens* GO Molecular Function			*Homo sapiens* GO Cellular Component		
# genes	7,897			8,579			7,833		
# terms	3,523			1,233			583		
# annotations	241,825			69,027			107,880		
# topics	352			123			58		
Weigthing Schema	$\#AP$	$\#Conf.$	$\%Conf.$	$\#AP$	$\#Conf.$	$\%Conf.$	$\#AP$	$\#Conf.$	$\%Conf.$
None	2,076	385	18.54	1,902	208	10.90	594	87	14.64
NTN	1,072	255	**23.78**	2,232	**250**	**11.20**	1,656	**645**	**38.94**
NTM	1,079	142	13.16	1,637	197	**12.03**	1,741	**576**	**33.08**
NTC	928	191	**20.58**	1,546	174	**11.25**	718	**209**	**29.11**
MTN	1,310	185	14.12	1,763	168	9.53	879	**371**	**42.21**
ATN	6,719	**1,318**	**19.61**	4,221	**510**	**12.08**	3,383	**1,354**	**40.02**
ATC	7,252	**1,087**	14.98	4,757	**442**	9.29	2,329	**791**	**33.96**
ATM	7,159	**1,109**	15.49	4,786	**577**	**12.06**	2,551	**881**	**34.54**

Sizes of the GO annotation datasets used (table upper part) and prediction test results (table lower part). # AP: number of annotations predicted; # Conf.: number of predicted annotations found confirmed in the newer GPDW version; % Conf.: percentage of predicted annotations found confirmed in the newer GPDW version. In bold are those values which indicate an improvement with respect to the unweighted variant.

of annotation terms as number of topics.

V. RESULTS

In the lower part of Table II the prediction test results are reported; for each weighting schema and dataset we report the number of AP ($\#AP$) and the number ($\#Conf.$) and percentage ($\%Conf.$) of the AP that are actually present in the newer version of the GPDW. The percentages show how the weighting schemes can improve the precision of a predictive system based on pLSAnorm in 15 out of 21 cases (71.43%); furthermore, in 11 of those cases they have also increased the number of correct predictions.

Nonetheless, not all the schemes induce equal improvement; among them, NTN and NTC are those that seem to perform better. The schemes ATN, ATC and ATM also perform well, but by generating a higher amount of predictions above the threshold (compared to the other schemes), they implicitly produce many not confirmed predictions. Conversely, the MTN schema decreases the performance of the predictive system in 2 out of 3 datasets, even if it is the best schema in the third dataset.

Although these results already show the relevance of the proposed weighting schemes, they could represent only partially the real improvement given by the weighting schemes; in fact, some predictions could be correct even if they are not present in the newer annotation set used for the validation.

VI. CONCLUSIONS

In this paper, we have shown that the proposed weighting schemes can enhance the predicted annotations. These schemes, associated with probabilistic latent class analysis methods, are able to predict novel functional annotations useful to biologists and physicians. In addition, the independence of such schemes from the organism and the terminology considered make them a remarkable tool for improving annotation sets, both in quality and quantity. Nevertheless, our tests have also underlined difficulties in the choice of the best weighting schema.

Future works will address advantages and issues related to applying these tools also to sets of annotations from other controlled vocabularies different from the GO ones. In particular, we will explore the benefits of considering jointly more than one terminology at a time, in order to take advantage of the potential semantic correlations among them.

We also plan to further verify the effectiveness of the proposed weighting schemes by assessing the quality of the top ranked predictions by means of a literature based validation procedure.

REFERENCES

[1] The Gene Ontology Consortium, ”Creating the gene ontology resource: Design and implementation”, Genome Res., vol. 11, pp. 14251433, 2001.

[2] G. Pandey et al., ”Computational Approaches for Protein Function Prediction: a Survey”, Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, Usa, 2006.

[3] 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.

[4] B. Done et al., ”A semantic analysis of genome annotations using weighting schemes”, *in IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, Honolulu, Hawaii,USA, 2007

[5] T. Hofmann, ”Probabilistic Latent Semantic Indexing”, *in Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval* (SIGIR-99), pp. 5057, 1999.

[6] M. Masseroli et al., ”Probabilistic Latent Semantic Analysis for prediction of Gene Ontology annotations”, *in Proc. IEEE World Congress on Computational Intelligence*, Brisbane, AU, 2012, pp. 2891-2898.

[7] A.P. Dempster et al., ”Maximum likelihood from incomplete data via the EM algorithm”, *J. Royal Statist. Soc. B*, pp. 39–138, 1977.

[8] G. Salton, M.J., *Introduction to modern information retrieval*, McGraw-Hill, ISBN 0-07-054484-0, 1986.

[9] A. Canakoglu et al., Integration of Biomolecular Interaction Data in a Genomic and Proteomic Data Warehouse to Support Biomedical Knowledge Discovery, *Computational Intelligence Methods for Bioinformatics and Biostatistics - LNCS*, vol. 7548, 2012, pp. 112–126, Springer, Heidelberg, Germany.