Functional annotation of genes through statistical analysis of biomedical articles

T. Theodosiou, L. Angelis, A. Vakali
Department of Informatics,
Aristotle University of Thessaloniki, 54124, Greece
{theodos,lef,avakali}@csd.auth.gr

Abstract

One of the most elaborate and important tasks in biology is the functional annotation of genes. Biologists have developed standardized and structured vocabularies, called bio-ontologies, to assist them in describing the different functions. A critical issue in the assignment of functions to genes is the utilization of knowledge from published biomedical articles. The purpose of this paper is to present a unified and comprehensive statistical methodology for functionally annotating genes using biomedical literature. Specifically, classification models are built using the discriminant analysis method while validation, analysis and interpretation of the results is based on graphical methods and various performance metrics and techniques. The general conclusions from the study are very promising, in the sense that the proposed methodology not only performs well in the assignment of functions to genes, but also provides useful and interpretable results regarding the discriminating power of certain keywords in the texts.

1. Introduction

Functional annotation of genes has become a crucial task in biology since it is very important to enhance genes description and understanding by extending the information related with them. In this context, biologists have improvised biological ontologies, like EcoCyc [9] and the popular Gene Ontology (GO) ([4], [2]), to provide a standard, controlled and structured vocabulary for describing a gene's functions.

In order to functionally annotate a gene product and its relevant gene with particular GO codes, various sources of information may be searched such as published literature, sequence analysis results, 3D analysis of the products of the gene, etc. Since it is quite important to focus on the most emerging and updated of these sources, a popular choice is to search on the vast number of article publications in order to involve the most recent and modern advances in gene processes and functions.

Despite the fact that any computational method is prone to errors and cannot be considered as accurate and reliable as the manual annotation of genes [4] it is evident that manually discovering and analyzing all the relationships of genes to functions established inside published articles is a very difficult and time consuming task.

Earlier research efforts for the gene annotation problem have applied natural language techniques to the published biomedical literature, by using informative keywords or predefined terms ([1], [6]) or sequence similarities [7]. Furthermore, other research efforts ([10], [12]) have utilized GO codes and applied data mining and machine learning methods to published biological literature stored in the PubMed database. Specifically, in [10] 21 GO codes were used in order to compare the performance of three different classification models (maximum entropy, naive Bayes and nearest- neighborhood) concluding that the maximum entropy method outperforms. In [12], a 12 GO codes set was used in order to evaluate Support Vector Machines (SVM) showing that SVMs outperform the maximum entropy classification. Both methods require the tuning of hyperparameters, which can be computationally expensive. Moreover, the SVM algorithm is complicated, difficult to implement and computationally demanding [15], [16].

The motivation for our work is based on the idea that the huge sets of published biological articles are a challenging ground where advanced multivariate statistical methods (such as discriminant analysis, factor analysis, graphical methods etc.) can be applied in order to extract functional annotation of genes. Our work focuses on the application of a *multivariate statistical methodology* for analyzing large biological articles with special emphasis on their classification ability. The proposed statistical methodology uses for classification the linear discriminant analysis (LDA), which is a simple statistical method that works without the need for tuning, and competes with more sophisticated methods [15], [16]. In our methodology LDA serves not only to build a classification model (with respect to GO codes), but also to explore the interrelations of the variables in the data set. These relations can help us to project our data onto spaces of lower dimensionality and to explain the significant differences in the data in terms of much fewer variables than those in the original data set. The results of the proposed statistical analysis can be used for inferences regarding the discriminating power of the considered biological texts (in terms of keywords) to provide functional annotation for genes. The effectiveness of LDA is evaluated by specific performance metrics (such as recall, precision, F-measure and cross-validation) and emphasized by graphical representations (such as boxplots, Andrews' curves and scatterplots). The experimentation carried out on biomedical abstracts from the PubMed database and the classification model trained for 12 GO codes (similarly to [12]), has shown the power and benefits of using the proposed methodology in annotating genes.

The remainder of the paper is structured as follows: Section 2 describes the proposed classification approach. Section 3 presents the results of experimentation. Conclusions and future work is given in Section 4.

2. Biological Text Classification Approach

2.1. Biological Document Representation

The source of published biomedical literature is the PubMed database (http://www.pubmed.com). The articles in this study are represented by the use of the popular Vector Space Model [11] which transforms a document into a vector of weighted words, suitable for statistical analysis. The steps towards identifying the vector for each document are [3]:

1. tokenization: extraction of all words appearing in an entire set of documents;
2. elimination of non informative words (stopwords) such as ”a”, ”and”, ”the”, etc.;
3. stemming: use only the root of each word;
4. counting of the number of occurrences of each word in each document;
5. elimination of non-content-bearing high-frequency and low-frequency words;
6. construction of a weight vector. The weights are binary (0 for absence of a word and 1 for presence).

The dataset in this study is a set A of articles where each article is a binary vector of size p (number of keywords). In order to proceed to the construction of the classification model, we manually assign a relevant GO code to each vector of the training set.

2.2. The Linear Discriminant Analysis

The method used to build the classification model is the linear discriminant analysis (LDA). The general purpose of LDA is to model the relationship between a dependent categorical variable with a set of independent or exploratory variables. In our context the independent variables are all binary (the document vectors) and represent the existence of keywords whereas the categorical dependent variable is the corresponding GO code with 12 possible values - categories. The specific goals of our study are: (a) to explain the differences between the GO codes in terms of the keywords and (b) to utilize the model for future predictions.

The LDA procedure produces a number of statistical results leading to the estimation of the probability that a vector (of word weights) belongs in a particular group. The original idea of LDA was the projection of the $n \times p$ data matrix $\mathbf{X}$ (n is the number of PubMed articles and p the number of words) onto a single dimension using the *Fisher's linear discrimination function* $\mathbf{z} = \mathbf{Xa}$. The vector of coefficients $\mathbf{a}$ should be chosen in such a way that an optimal discrimination is achieved via the maximization of the ratio of the *between-group-sum of square* to the *within-group-sum of squares*.

In the case where the categories of the dependent variable (the GO categories in our context) are $k > 12$, this idea can be generalized and the method computes

$$m = \min(k-1, p) \, (1)$$

new variables, the canonical discriminant functions which can be used to exhibit the differences among the categories. LDA can be applied either to the original data matrix or to the data resulting from a standard data reduction technique, for example principal component analysis (PCA) or more generally, factor analysis. For detailed description of LDA we refer to [13].

The efficiency of LDA is measured by various performance metrics such as precision, recall and F-measure. *Precision* is the ratio of the correct classifications of a specific group to all the documents assigned to that group. *Recall* is the ratio of the correct classifications of a specific group to all the documents of the group contained in the dataset. The *F-measure* is defined as the harmonic mean between precision and recall [8]. All these measures are very easily computed from the *classification matrix* which shows the number of cases that were correctly classified and those that were misclassified [14].

3. Experimentation

The proposed method was tested under a dataset of 12 GO terms, a subset of the 21 GO terms used in [10]. The main criterion for choosing these 12 GO terms was the large amount of relevant articles returned by the PubMed retrieval system for each one of them. For practical purposes, each GO code was represented by a group number ranging from 1 to 12. The assignment of the numbers was random and is listed in Table 1. Due to the space limitations, we outline the most important results of this study.

As a training dataset we use a set of $n = 9009$ articles published until 1999. Thus, the data matrix consists of $n = 9009$ vector representations of binary values for a set of $p = 1642$ words, accompanied by one of $k = 12$ GO codes. Since in our data the number of independent variables is very large ($p = 1642$) and the number of GO codes is $k = 12$, the procedure results in only 11 variables ($m = 11$), according to Formula 1, which can be used for further analysis and interpretation.

Table 1. GO codes, Terms and corresponding group number

GO code	GO term	Group no.
GO:0006914	Autophagy	1
GO:0007049	Cell cycle	2
GO:0008283	Cell proliferation	3
GO:0007267	Cell cell signalling	4
GO:0006943	Chemimechanical coupling	5
GO:0007126	Meiosis	6
GO:0008152	Metabolism	7
GO:0007048	Oncogenesis	8
GO:0006950	Stress response	9
GO:0006810	Trasnport	10
GO:0008219	Cell death	11
GO:0007165	Signal transduction	12

Two test datasets were used:

Test dataset I (hold-out sample): subsets of articles until 1999, part of the training set. Three different types of hold-out samples (10%, 20% and 30% of the total number of abstracts) were used to validate the classification model. For each type, 10 different samples were selected randomly. Every sample was used for validating the classification model constructed by the remaining training dataset. The accuracies of the ten different models were averaged in order to have more statistically reliable estimation for the overall performance.

Test dataset II: 8225 articles published in the period 2000 – 2004 (not a part of the training dataset). The purpose of this partition was to simulate a real case where an existing corpus of documents is the basis for future predictions.

Table 2 shows that the average predictive accuracy of the LDA model for each one of the three hold-out samples (test dataset I) is very good, i.e. above 80% and is not affected much by the increase of the size of the samples. The average fitting accuracy (the correct classification of the training articles) is also high, i.e. above 90%. The results show that the LDA model is general enough, i.e. avoids over-fitting to the training dataset, for classifying correctly new data.

Table 3 shows the classification matrix for the test dataset II. The rows correspond to the actual members of each group (GO category), whereas the columns show the predicted members for each group. For example, the correct classifications for the GO code group 1 were 144 out of 161, which is the actual group size. Table 4 shows the precision, recall and F-measure of each group, which are calculated from the classification matrix. The mean precision is 77.2% and the mean recall is 74.3%. The mean F-measure has a value of 75.4%, which indicates that the proposed classification approach has quite satisfactory results.

In studying the accuracy results (Table 4), we notice that group 3 (cell proliferation) has the lowest F-measure (32.1%). This might be due to the fact that either the training set was not very informative and/or the test set included articles with inaccurate information about it. However, further investigation revealed that the training dataset was actually informative about 'cell proliferation', but the information was highly correlated to other GO categories. For example, from the definition of 'cell proliferation' [4], we can see that the correlation with 'oncogenesis' is high and so the two terms are referenced together very often. A similar, but negative, correlation exists with 'cell death'.

Another important issue that could explain the difference in the accuracy of the model for each GO code (group) is the position of the GO term in the GO structure. For example, the fact that cell proliferation is a very general GO category could decrease the accuracy of the model. The granularity of the GO codes is also mentioned in [10] as an issue for the accuracy of the classification.

Table 2. Average fitting and predictive accuracy for each hold-out percentage.

	Training dataset	Test dataset	Fitting accuracy	Predictive accuracy
Type 1	8108 (90%)	909 (10%)	92.04%	81.90%
Type 2	7207 (80%)	1802 (20%)	92.69%	81.87%
Type 3	6306 (70%)	2703 (30%)	93.63%	80.81%

Table 3. Classification matrix for the test dataset II

Actual Group	**Predicted Group**												**Actual group Size**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	
1	144	1	0	0	0	0	0	1	3	2	1	9	**161**
2	0	515	55	1	9	10	28	48	39	9	50	50	**814**
3	0	16	55	0	3	0	2	7	2	3	14	10	**112**
4	0	4	1	101	0	1	2	1	2	5	0	6	**123**
5	0	19	8	4	426	1	20	107	12	14	1	48	**563**
6	0	19	0	11	3	430	5	0	6	4	1	11	**490**
7	1	22	6	2	29	0	878	24	48	36	18	48	**1112**
8	1	36	32	0	1	0	21	397	24	6	32	52	**602**
9	2	39	10	3	6	8	66	24	806	11	21	65	**1061**
10	1	1	0	2	1	0	0	0	0	240	1	1	**247**
11	1	32	15	4	4	0	16	34	24	12	1289	59	**1490**
12	1	37	49	9	39	2	45	71	22	38	59	1078	**1450**
Predicted Group Size	**151**	**741**	**231**	**137**	**521**	**452**	**1083**	**617**	**988**	**380**	**1487**	**1437**	**8225**

Table 4. Accuracy results of the test dataset II

Group	Recall	Precision	F-measure
1	89.4%	95.4%	92.3%
2	63.3%	69.5%	66.3%
3	49.1%	23.8%	32.1%
4	82.1%	73.7%	77.7%
5	75.7%	81.8%	78.6%
6	87.8%	95.1%	91.3%
7	89.0%	81.1%	84.9%
8	66.0%	64.3%	65.1%
9	76.0%	81.6%	78.7%
10	97.2%	63.2%	76.6%
11	86.5%	86.7%	86.6%
12	74.3%	75.0%	74.6%
Mean	77.2%	74.3%	75.4%

Interesting results can be obtained from the study of the 11 canonical discriminant function scores. It is important to notice that the functions are ordered in the sense that the first is the most important for the discrimination between groups, and so on [13]. We present some figures to visualize performance metrics by using several graphical tools such as box-plots, and scatter-plots. In the box-plot of Figure 1 it is shown that the distribution of the first function of the test dataset II clearly separates categories 6 and 10. In Figure 2 it is shown that by plotting the centroids (means) of the 12 groups with respect to the first two functions, GO categories 1, 4, 6, 10 and 11 are clearly discriminated. The 11 discriminant functions combined together can classify efficiently the data to all GO categories. The centroids of the discriminant functions can also be used to produce the Andrews' curves [5] (Figure 3). Each curve corresponds to a different GO category and as the curves are quite dissimilar in shape, indicate that the 12 GO categories can be efficiently discriminated (especially groups 1, 4, 6, 10 and 11).

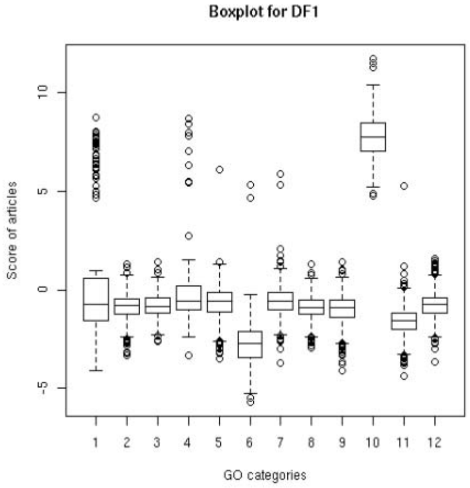

Figure 1. Graphical display of the discriminative ability of the canonical functions

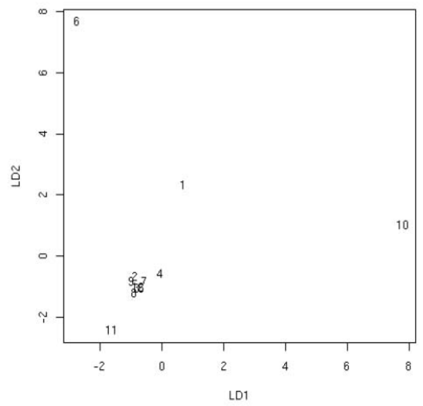

Figure 2. Graphical display of the discriminative ability of the canonical functions

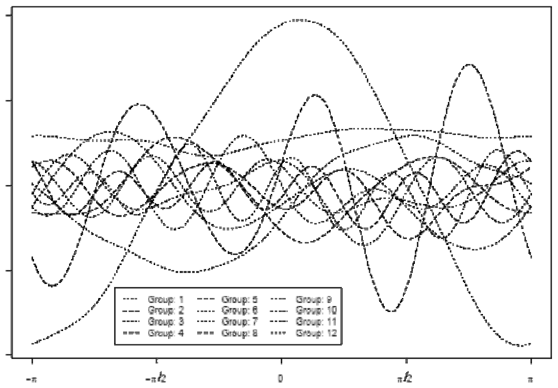

Figure 3. Andrews' curves for each of the 12 GO code groups

4. Conclusion

This work demonstrates that the proposed statistical methodology can be used to facilitate the process of automatically assigning a GO category to a gene product, by exploiting the information from published biological literature. The statistical analysis of the LDA results, with various performance metrics and graphical analysis techniques, allows the better understanding of the datasets focusing on the underlying relationships. Our future plans involve a thorough statistical analysis of the relevant datasets including the structural information of the GO codes. Special emphasis should be given to the discriminant functions that can be viewed as new variables for describing the corpus of articles in much fewer dimensions than the original data.

5. References

[1] M.A. Andrade and A. Valencia, "Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system.", *Proc Int Conf Intell Syst Mol Biol,* 1997, pp.25–32.

[2] E. Camon, M. Magrane, D. Barrell, D. Binns, W. Fleischmann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, A. Cox, and R. Apweiler., "The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro.", *Genome Res, 13(4),* Apr 2003, pp. 662–72

[3] H. Schutze and C. Manning,. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[4] Gene Ontology Consortium, "Creating the gene ontology resource: design and implementation.", *Genome Res, 11(8),* Aug 2001,pp. 1425–33

[5] D. F. Andrews, "Plots of high-dimensional data.", *Biometrics 28,* March 1972, pp. 125–136.

[6] F. Eisenhaber and P. Bork, "Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries.", *Bioinformatics 15(7-8),* 1999, pp. 528–35.

[7] S. Hennig, D. Groth and H. Lehrach, "Automated Gene Ontology annotation for anonymous sequence data.", *Nucleic Acids Res 31(13),* Jul 2003, pp. 3712–5.

[8] C. J. Van Rijsbergen, *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow, 1979.

[9] I.M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I.T. Paulsen, M. Peralta-Gil and P.D. Karp, "EcoCyc: A comprehensive database resource for Escherichia coli", *Nucleic Acids Research 33,* 2005, pp. 334-7

[10] S. Raychaudhuri, J. T Chang, P. D. Sutphin and R. B. Altman, "Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature.", *Genome Res 12(1),* Jan2002, pp. 203–14.

[11] G. Salton, "Automatic text analysis.", *Science 168,* 1970, pp. 335–343.

[12] H. Kazawa, T. Izumitani, H. Taira, "Assigning gene ontology categories (go) to yeast genes using text-based supervised learning methods.", *In Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004),* 2004.

[13] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S. Fourth Edition*. Springer, 4th edition, 2002.

[14] Hair, Anderson, Tatham, Black, *Mutlivariate Data Analysis Prentice Hall*, PTR., 5th edition, 1998.

[15] Kurt Hornik, David Meyer, Friedrich Leisch, "*Benchmarking support vector machines"*, Technical Report 78, Vienna University of Economics and Business Administration 2002.

[16] Yu-Shan Shih, Tjen-Sien Lim, Wei-Yin Loh, "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms.", *Machine Learning 40,* 2000, pp. 203–229.