W230–W234 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh484

Dragon TF Association Miner: a system for exploring transcription factor associations through text-mining

Hong Pan, Li Zuo, Vidhu Choudhary, Zhuo Zhang, Shoi Houi Leow, Fui Teen Chong, Yingliang Huang, Victor Wui Siong Ong, Bijayalaxmi Mohanty, Sin Lam Tan, S. P. T. Krishnan and Vladimir B. Bajic*

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613

Received February 15, 2004; Revised April 19, 2004; Accepted May 5, 2004

ABSTRACT

We present Dragon TF Association Miner (DTFAM), a system for text-mining of PubMed documents for potential functional association of transcription factors (TFs) with terms from Gene Ontology (GO) and with diseases. DTFAM has been trained and tested in the selection of relevant documents on a manually curated dataset containing >3000 PubMed abstracts relevant to transcription control. On our test data the system achieves sensitivity of 80% with specificity of 82%. DTFAM provides comprehensive tabular and graphical reports linking terms to relevant sets of documents. These documents are color-coded for easier inspection. DTFAM complements the existing biological resources by collecting, assessing, extracting and presenting associations that can reveal some of the not so easily observable connections among the entities found which could explain the functions of TFs and help decipher parts of gene transcriptional regulatory networks. DTFAM summarizes information from a large volume of documents saving time and making analysis simpler for individual users. DTFAM is freely available for academic and non-profit users at http://research.i2r.a-star.edu.sg/DRAGON/TFAM/.

INTRODUCTION

Understanding the full complexity of transcriptional control and the associated effects it can produce is a difficult and as yet unsolved problem. There are different ways in which transcription factors (TFs) influence each other or affect transcription of their target genes. Complete understanding of the control effects of individual TFs is not possible without insights into the molecular functions which they affect or the biological processes they are involved with, as well as the associations of TFs with different diseases. Such information is scattered across different resources. For an individual user, obtaining it is a very tedious task and is frequently not feasible due to the large volume of documents which have to be processed and the numerous resources which need to be consulted. However, a large portion of such useful information can be found in the abstracts of scientific documents, as deposited in the PubMed repository (1). Realizing the great potential of extracting useful information from biomedical literature by text-mining [see reviews (2–6)], many text-mining systems have been developed, such as PubGene (7), MedMiner (8), XplorMed (9), PubMatrix (10), AbXtract (11), EASE (12), VxInsight (13), SUISEKI (14), GIS (15), PreBIND (16), Genes2Diseases (17), MeKE (18), MeSHmap (19) and HAPI (20). These systems provide different types of information to the end-user—giving more insight into protein–protein interactions [see also (21)] and gene–gene relations—or extract more comprehensive relations between genes and diseases, or other important categories such as terms from Gene Ontology (GO) (22). Some of these systems are not specialized, in the sense that they allow arbitrary vocabularies to be used (10). Several systems (14,16–18,21) have built-in modules for filtering out irrelevant documents. However, none of these web-based solutions focuses on TFs and transcriptional regulation as our Dragon TF Association Miner (DTFAM) system does. DTFAM attempts to collect, assess, extract and present potential associations between TFs, GO categories and diseases (derived from the web site of the Karolinska Institute, Sweden), based on mining PubMed documents. The aim of the DTFAM system is to find clues on potential associations between different queried components, particularly those which can suggest the function of the entity found or an association of its functionality with different diseases. It analyzes documents from PubMed, selects the most relevant ones, analyzes them again and provides comprehensive tabular and graphical reports with terms linked to the relevant sets of

*To whom correspondence should be addressed. Tel: +65 6874 8800; Fax: +65 6774 8056; Email: bajicv@i2r.a-star.edu.sg

PubMed documents. Association map networks are visual representations of associations provided by the system. DTFAM complements the existing biological resources by presenting associations that can reveal some of the not so easily observable connections of the examined terms which could explain the functions of TFs and help decipher parts of gene transcriptional regulatory networks. Another crucial aspect of the DTFAM utility for biologists is that it condenses information from a large volume of documents for easy inspection and analysis, thus making analysis easier for individual users.

DTFAM has been trained and tested on a manually curated corpus of documents. The system is freely accessible for academic and non-profit users at http://research.i2r.a-star.edu.sg/DRAGON/TFAM/. We believe that Dragon TF Association Miner offers a useful set of functions to support the research of the life sciences community.

SYSTEM DESCRIPTION

The goal of the DTFAM system is to provide information about the potential association of TFs with terms from four well-controlled vocabularies in order to help biologists infer unusual functional associations. Three vocabularies are related to GO (biological process, molecular function, cellular component), while the fourth one is related to different disease states. Functional associations of TFs with any term from these four categories can be focused on any combination of these terms, such as biological process, or biological process and diseases depending on the user's selection. All GO vocabularies are general. Disease vocabulary is focused on human diseases, while the TF vocabulary contains ∼10 000 TF names (http://research.i2r.a-star.edu.sg/DRAGON/TFAM/current.php) and their synonyms collected for various species—mainly eukaryotes, but also including some prokaryotes. The process and sources used to compile this vocabulary are explained at the DTFAM web site. Some necessary data cleaning has been done for all vocabularies in order to enable more efficient text-mining.

There are several modules which operate within the system.

(i) The first module analyzes the submitted text, indexes words/terms, matches the selected features with terms based on trained models and assigns the weights of each selected term according to its document and term frequencies.

(ii) The second module analyzes the content of the processed document, and based on the previously selected features applies one of 65 previously derived models to assess whether the analyzed document contains information about TF relationships or not. If the model signals that the document contains information about TF relationships, the document is accepted for the final analysis, otherwise it is rejected. Which of the 65 models will be used depends on the user's selection of sensitivity on the main page. The higher the sensitivity, the more documents will be selected for the analysis, but this may also include a large number of irrelevant documents. The analysis of the documents is done on the whole document level and is not based on the presence of specific types of sentences.

(iii) The third module makes an inventory of all terms and expressions found in the finally selected documents, summarizes the information collected and presents this information in tabular form. This module presents information from the vocabularies in different colors for easier inspection. Also, all PubMed documents in which the terms have been found are presented to the user via links. This allows users to assess the original information and determine its relevance. For easier inspection the terms found are also color-highlighted.

(iv) The fourth module analyzes the connections (associations) between the terms and generates one or more association map networks of these terms. The association of terms is based on their co-occurrence in the same PubMed document. The nodes of the generated graphs represent the terms from the selected vocabularies. TF names are represented by ellipsoidal nodes with a yellow background. Diseases are represented by ellipsoidal nodes with a gray background. Terms from GO categories are represented by rhomboidal shapes, with biological processes having a green background, molecular functions a light blue background and cellular components a magenta background. All nodes provide links to a set of related PubMed documents, with terms color-highlighted, to allow the user to inspect and assess the relevance of proposed associations (Supplementary material). We use the Graphviz software to generate association networks in our system (23).

Data

DTFAM has been trained and tested on a corpus of manually curated data. We collected a random subset of 3000 PubMed documents related to transcription regulation. In a 3-fold blind checking, these documents were analyzed and classified by five biologists and one chemist, who assessed whether the document contains information about TF relationships or not. From such labeled data, training and test sets have been formed.

To the best of our knowledge, this is the only manually curated corpus of data used for the development of TF relationship extraction systems. It also seems to be the largest manually curated corpus of data used for development of any other (known to us) general biomedical text-mining system (such as those for protein–protein interactions).

Development of models for the selection of relevant documents

One of the key features of the DTFAM system is its ability to filter out a portion of irrelevant documents based on the expected sensitivity level of the system as specified by users. To provide this function we have developed a module which comprises 65 different models that perform this task. We have used two measures to quantify the system's ability to correctly classify positive and negative data. These measures are sensitivity and specificity. Sensitivity is defined as TP/(all positives), while specificity is defined as TN/(all negatives). Here, TP and TN denote the number of true positive and true negative predictions, respectively. TP prediction means that the system correctly selected a document as one which contains information about the relationships between TFs, while TN prediction means that the system correctly rejected a document as one that does not contain this information.

Our 3000 manually labeled documents contain both positive and negative data. Positives are those containing explicit statements about TF relationships, while negative ones are those that do not explicitly state such TF relationships although they contain TF names and different relationship expressions. For the training sets we randomly selected 30% of positive and 30% of negative data. The remaining 70% of the data in each case was used for testing.

Each distinct word in the training set is considered a potential feature. We processed documents in the training set and eliminated from documents all common words such as 'the', 'a' and 'we'. All TF names were replaced by an artificial word 'TFname' and all relationship expressions were replaced by another artificial word 'RELATIONword'. For each of the remaining words we calculated its frequency wf, i.e. the number of times the word appeared in the training data, as well as the number of documents, df, where that word was found. From all words we selected only those that had df not <100. This left us with 369 words. These words have been sorted according to their contribution to the separation of positive and negative training data as measured using linear discriminant analysis (LDA). This list we denote as LS.

The recognition models have been determined in the following manner. We eliminated outliers from the training data based on all 369 features. Then we selected the desired sensitivity level for the model. Sensitivity levels have been chosen from 0.36 up to 1.0 in steps of 0.01. A sensitivity level of 0.36 means that the system correctly recognizes 36% of positive data. For the selected sensitivity level, we determined a set of LDA models on the training set using different numbers of features in the range 2–369. The features were selected from LS, taking first the two most significant ones and finding an LDA model for them, then using the three most significant ones and determining a model for them, and so on. Then, we used a set of feed-forward artificial neural network (ANN) models. All these models have been chosen to have three layers: an input layer, a hidden layer and an output layer. The ANN models used linear neurons in the input layer and 'logsig' neurons (24) in the other two layers. The number of neurons in the input layer was equal to the number of words selected as features. The output layer had only one neuron. The hidden layer was tested with the number of neurons varying from 2 to 5. The training algorithm (25) was analogous to the one used in the Dragon Gene Start Finder system (26) and is presented in detail in (27). For the ANN models, we varied the number of neurons in the hidden layer from 2 to 5, while the number of features used varied from 2 to 369 (in the same way as for the LDA models). The final model for the selected sensitivity is chosen out of all LDA and ANN models as the best performing model on the test set. We repeated this procedure for all 65 different sensitivities. This produced 41 ANN models and 24 LDA models. The number of features used by these models ranged from 63 to 147.

Performance of the system

The performance curve ('Data and Accuracy', http://research.i2r.a-star.edu.sg/DRAGON/TFAM/data.htm) showing Sensitivity = TP/(all positives) versus Specificity = TN/(all negatives) is obtained from an assessment based on the whole abstract content, without any specific requirements that the abstract contain particular types of sentences which express such relationships.

Table 1. Results for test on 188 PubMed documents

Selected (expected) sensitivity	Measured sensitivity	Measured specificity	TP	TN
0.95	0.9815	0.3304	53	37
0.90	0.9630	0.5089	52	57
0.85	0.9259	0.5893	50	66
0.80	0.8704	0.6964	47	78
0.75	0.8704	0.7054	47	79
0.70	0.8519	0.7411	46	83
0.65	0.7963	0.8125	43	91
0.60	0.7593	0.8393	41	94
0.55	0.7407	0.8571	40	96

Additionally, we performed another test. On April 10, 2004 we collected from PubMed all documents from January and February 2004 related to transcription factor relationships. In total 188 documents were collected. We manually labeled them as positive or negative ones. Out of 188 documents, 166 contained TF names; 54 documents were positive and 112 documents were negative. We based the analysis on these 166 documents and examined the performance of DTFAM at selected sensitivities 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95. The results are summarized in Table 1. These results are generally better than what we have obtained on our test set. This is likely the consequence of the biased content of documents in this set due to its small size (only 166 documents), as opposed to the more diversified data we have used for system development.

USING THE SYSTEM

Input information

Users have to provide a set of PubMed abstracts or summary information from these abstracts for the analysis. Alternatively, they can search PubMed using the Entrez system at the NCBI web site. The set of retrieved PubMed documents should be

(i) copied in the summary format and pasted into input window A on the DTFAM main page, or
(ii) copied in the abstract format and pasted into input window B on the DTFAM main page, or
(iii) saved in the abstract format as a text file and uploaded into input window C on the DTFAM main page or
(iv) users can formulate an arbitrary query for a PubMed search following the prescribed rules for the Entrez system, and this query will be passed to the NCBI site.

After the user presses pressing the 'Submit' button, DTFAM will start analysis of the text. There is a limit of 5000 abstracts per session in order not to block the server.

There are three more pieces of information that a user needs to provide to the system:

(i) email address, to which a link to the report page will be sent (in addition to the dynamically created one),

(ii) selection of vocabularies intended to be used in the analysis, and
(iii) sensitivity at which the system is expected to operate.

Sensitivity determines how strict the internal document assessment system will be in the selection of documents which potentially contain TF relationships. The higher the sensitivity (closer to 1.0) the less restrictive the system will be, but the more irrelevant documents will be included in the analysis.

What users should or should not expect

This system assesses document relevance using one of its modules. This narrows down the set of documents which will eventually be analyzed. This action eliminates a great part of the irrelevant information from the reports, but not all.

Since this system analyzes co-occurrence of terms within the document and since the documents analyzed are abstracts of scientific reports, which present summaries of the most important findings, there obviously exists a loose relationship between the co-occurring terms. However, the actual nature of these relationships is not analyzed by the system. It is then left to the user to accept or reject the association proposed by the system.

Another issue is the completeness of information. Since the analyzed documents are abstracts, it is not likely that the system can collect all relevant information on the association of terms. Thus, the resultant association maps will represent only a subset of all possible relationships. In addition, if the sensitivity selected is <1.0, the system may eliminate some of the documents which contain relevant information.

The analysis of a large number of documents requires some processing time. Thus, users should not expect to get the results immediately. It is normal to wait several minutes for the results of the analysis of a large set of documents. Most of the time is spent on the generation of complex association map networks. Sometimes the networks produced are so large that they cannot be opened and viewed in the Internet browser. The more specific selection of documents is then suggested, as well as the selection of a smaller number of vocabularies for use in the analysis.

It should be noted that some TFs are unfortunately named using a 'common' word, such as 'So', 'Cactus', 'lung' and 'For'. These common names could sometimes be wrongly detected as TF names. Similarly, some TF names, such as '3.4', are very inappropriate for automated analysis. However, due to the exploratory nature of the analysis that DTFAM provides, we decided to keep most such names in the vocabulary since they may represent real TFs, and we leave it to the user to determine their relevance from the associated PubMed document.

Why are several association map networks usually presented?

There are two reasons for generating several association map networks from a single set of documents.

(i) The terms found in one network need not co-occur with the terms from the other networks.
(ii) Frequently, even with a very specific selection of documents, the resulting networks are very complex. In order to allow the user to view these networks in the browser we have implemented an automatic partitioning procedure to divide a large network into several smaller networks. Users can minimize this partitioning of networks by selecting fewer vocabularies to be included in the analysis and making the selection of documents even more specific.

How to use the system most efficiently

Users are advised to make their queries to PubMed sufficiently specific that the most relevant documents are collected. Although the system will successfully analyze up to 5000 documents in one session, we strongly suggest that the number of submitted documents is not >1000, and preferably should be <500. Moreover, it is advisable to use a level of sensitivity between 0.8 and 0.97, as this will filter out 85 to 50% of irrelevant documents. This will also speed up the analysis process and reduce the time required to obtain the results.

The DTFAM tool allows searches of any set of documents in the text format of PubMed abstracts. The initial selection of abstracts can focus on a specific aspect of transcriptional regulation, disease, biological process, molecular function, cellular component, combination of these or any other relevant terms indexed in PubMed. Users can include terms from GO vocabularies or disease vocabulary for the analysis. TF names will be included by default.

Example

As an example, let us assume that we want to find potential TFs involved in the toll-like receptor-mediated activation of a signaling pathway resulting in antimicrobial innate immune response (28), as well as functional relationships of the TFs found with either GO categories or diseases.

To conduct this exploration a user can select a sensitivity of 0.97, upload a file with abstracts collected from PubMed with the query 'toll antimicrobial' and select all four selectable vocabularies on the main DTFAM page. The system will produce a report of the form 'MainReportPage' (Supplementary Material). In this particular case we noticed that there were 102 documents found, of which 32 contained TF names. Out of these 32 documents the system selected 20 for final analysis. The results of this analysis are summarized in two tabular reports (Supplementary Material), and in an association map network generated by the system. An interesting observation after analyzing the network is that the system detected IkappaB and NF-kappaB as TFs relevant for this signaling pathway. The role of these two TFs in this pathway is documented in (28). Moreover, most of the entities found and presented in the network relate to immune response and found GO categories. This demonstrates that DTFAM is capable of extracting relevant biological knowledge. However, a user should not blindly accept the results of the analysis and should check the relevance of detected associations by consulting the references used by the system. We have made this task easier for the user by providing links to the documents used, and we also color-highlight the terms used in the analysis.

DIFFERENCES WITH RESPECT TO OTHER SYSTEMS

There are several defining characteristics of our DTFAM system:

(i) It is focused on exploring potential association of TFs with other important functional categories such as GO terms and diseases.

(ii) It provides both tabular and graphical reports with links to the relevant set of documents with color-highlighted terms to make the user's inspection easier.
(iii) Its module for filtering irrelevant documents has been trained on a manually curated corpus of data.
(iv) It uses five manually curated vocabularies (one for TF names and synonyms, three for GO categories, one for diseases).

Other systems referenced in this article currently do not have an option to focus on TFs and transcriptional regulation. Moreover, none of these systems has been trained on such a large corpus of manually cleaned data, and particularly not on data related to relationships between TFs. Moreover, the other systems do not allow the user to select the stringency with which irrelevant documents are filtered. We believe that the combination of all the features listed above provides a great utility for the life sciences community.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

REFERENCES

1. Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Suzek,T.O., Tatusova,T.A. and Wagner,L. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32**, D35–D40.
2. Dickman,S. (2003) Tough mining: the challenges of searching the scientific literature. *PLoS Biol.*, **1**, E48, Epub 2003 Nov 17.
3. de Bruijn,B. and Martin,J. (2002) Getting to the (c)ore of knowledge: mining biomedical literature. *Int. J. Med. Inf.*, **67**, 7–18.
4. Grivell,L. (2002) Mining the bibliome: searching for a needle in a haystack? New computing tools are needed to effectively scan the growing amount of scientific literature for useful information. *EMBO Rep.*, **3**, 200–203.
5. Andrade,M.A. and Bork,P. (2000) Automated extraction of information in molecular biology. *FEBS Lett.*, **476**, 12–17.
6. Schulze-Kremer,S. (2002) Ontologies for molecular biology and bioinformatics. *In Silico Biol.*, **2**, 179–193.
7. Jenssen,T.K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.*, **28**, 21–28.
8. Tanabe,L., Scherf,U., Smith,L.H., Lee,J.K., Hunter,L. and Weinstein,J.N. (1999) MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, **27**, 1210–1214, 1216–1217.
9. Perez-Iratxeta,C., Perez,A.J., Bork,P. and Andrade,M.A. (2003) Update on XplorMed: a web server for exploring scientific literature. *Nucleic Acids Res.*, **31**, 3866–3868.
10. Becker,K.G., Hosack,D.A., Dennis,G.,Jr, Lempicki,R.A., Bright,T.J., Cheadle,C. and Engel,J. (2003) PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics*, **4**, 61.
11. Asher,B. (2000) Decision analytics software solutions for proteomics analysis. *J. Mol. Graph Model.*, **18**, 79–82.
12. Hosack,D.A., Dennis,G., Sherman,B.T., Lane,H.C. and Lempicki,R.A. (2003) Identifying biological themes within lists of genes with EASE. *Genome Bio.*, **4**, R70.
13. Kim,S.K., Lund,J., Kiraly,M., Duke,K., Jiang,M., Stuart,J.M., Eizinger,A., Wylie,B.N. and Davidson,G.S. (2001) A gene expression map for *Caenorhabditis elegans*. *Science*, **293**, 2087–2092.
14. Blaschke,C. and Valencia,A. (2001) The potential use of SUISEKI as a protein interaction discovery tool. *Genome Inform. Ser. Workshop Genome Inform.*, **12**, 123–134.
15. Chiang,J.H., Yu,H.C. and Hsu,H.J. (2004) GIS: a biomedical text-mining system for gene information discovery. *Bioinformatics*, **20**, 120–121.
16. Donaldson,I., Martin,J., de Bruijn,B., Wolting,C., Lay,V., Tuekam,B., Zhang,S., Baskin,B., Bader,G.D., Michalickova,K., Pawson,T. and Hogue,C.W. (2003) PreBIND and Textomy—mining the biomedical literature for protein–protein interactions using a support vector machine. *BMC Bioinformatics*, **4**, 11.
17. Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2002) Association of genes to genetically inherited diseases using data mining. *Nature Genet.*, **31**, 316–319.
18. Chiang,J.H. and Yu,H.C. (2003) MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics*, **19**, 1417–1422.
19. Srinivasan,P. (2001) MeSHmap: a text mining tool for MEDLINE. *Proc. AMIA Symp.* 2001, 642–646.
20. Masys,D.R., Welsh,J.B., Fink,J.L., Gribskov,M., Klacansky,I. and Corbeil,J. (2001) Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, **7**, 319–326.
21. Ono,T., Hishigaki,H., Tanigami,A. and Takagi,T. (2001) Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics*, **17**, 155–161.
22. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
23. Gansner,E.R. and North,S.C. (2000) An open graph visualization system and its applications to software engineering. *Software Pract. Exper.*, **30**, 1203–1233.
24. Bishop,C.M. (1995) *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, UK.
25. Sha,D. and Bajic,V.B. (2002) On-line hybrid learning algorithm for MLP in identification problems. *Comp. Electr. Eng, An Int. J.*, **28**, 587–598.
26. Bajic,V.B. and Seah,S.H. (2003) Dragon gene start finder identifies approximate locations of the 5′ ends of genes. *Nucleic Acids Res.*, **31**, 3560–3563.
27. Bajic,V.B. and Seah,S.H. (2003) Dragon Gene Start Finder: an advanced system for finding approximate locations of the start of gene transcriptional units. *Genome Res.*, **13**, 1923–1929.
28. Telepnev,M., Golovliov,I., Grundstrom,T., Tarnvik,A. and Sjostedt,A. (2003) *Francisella tularensis* inhibits toll-like receptor-mediated activation of intracellular signaling and secretion of TNF-alpha and IL-1 from murine macrophages. *Cell Microbiol.*, **5**, 41–51.