Intelligent Extraction Versus Advanced Query: Recognize Transcription Factors from Databases

Zhuo Zhang[1], Merlin Veronika[1], See-Kiong Ng[1], and Vladimir B Bajic[2]

[1] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
[2] South African National Bioinformatics Institute, Bellville 7535, South Africa
{zzhang, skng}@i2r.a-star.edu.sg, vlad@sanbi.ac.za

Abstract. Many entries in major biological databases have incomplete functional annotation and thus, frequently, it is difficult to identify entries for a specific functional category. We combined information of protein functional domains and gene ontology descriptions for highly accurate identification of transcription factor (TF) entries in Swiss-Prot and Entrez Gene databases. Our method utilizes support vector machines and it efficiently separates TF entries from non-TF entries. The 10-fold cross validation of predictions produced on average a positive predictive value of 97.5% and sensitivity of 93.4%. Using this method we have scanned the whole Swiss-Prot and Entrez Gene databases and extracted 13826 unique TF entries. Based on a separate manual test of 500 randomly chosen extracted TF entries, we found that the non-TF (erroneous) entries were present in 2% of the cases.

1 Introduction

Recent years' advance in genome research has yielded thousand hundreds protein and gene sequences accumulated in genomic databases such as Swiss-Prot [1] and Entrez Gene [2], which, being annotated carefully, provide valuable knowledgebase for further research. Meanwhile, effort has been put in protein and gene classification. Researchers attempted to categorize proteins/genes into functional or structural associated groups. To name a few, Pfam [3] intends to cluster proteins into various families based on functional domains. The Gene Ontology [4] project provides three set of controlled vocabulary to describe gene or gene product in any organism, includes three ontologies, namely, molecular function, biological process and cellular component.

The classification mechanisms provide convenient ways for biologists looking for particular groups of proteins or genes, e.g., proteins contain Homeo-Box domain (PF00046), or genes expressed in nucleus (GO:0005634). However, when search criteria is less distinct, the query may not render satisfactory results. For example, when looking for "transcription factor", the search engine of the query database usually applies a pattern matching to text fields – as a result, real transcription factors without explicit notes will be overlooked. Moreover, when a none-TF entry was annotated such as "regulated by transcription factor", it will be extracted by the query.

J.C. Rajapakse, L. Wong, and R. Acharya (Eds.): PRIB 2006, LNBI 4146, pp. 133–139, 2006.

In this study, we use Transcription Factor as an example, to illustrate our intelligent extraction method to find a particular group of proteins / genes from public databases based on prior knowledge and available annotation, as versus to common query method.

1.1 Background

Transcription factors (TFs) form a key regulatory family of proteins that control transcriptional activation of genes. Knowledge of TFs and their activities relative to different genes and gene products is necessary for deciphering transcriptional gene regulatory networks. The definition of TF may vary in different forms, in this study we adopted the broader definition which allows the TFs to be

a/ a protein that regulates transcription (after nuclear translocation) by sequence-specific interaction with DNA or
b/ by stoichimetric interaction with a protein that can be assembled into a sequence-specific DNA-protein complex

It is a challenge to identify which entries belongs to TF even in curated databases. Under Gene Ontology categories, TF genes are categorized into various classes, e.g., *Transcription Factor Activity* (GO:0003700), *Regulation of transcription, DNA-dependent* (GO:0006355), *regulation of transcription*(GO:0045449) and many more. A quick inspection shows that transcription factor genes scatter across tens of GO term in *Molecular Function* category alone, the relationship between these terms, in terms of ontology tree, may be parent-child, siblings, or even unrelated. In the context of protein family classified by Pfam [3] schema, transcription factors present in hundreds of families, such as *Homeobox* (PF00046), *zinc finger* (PF00096) etc. Furthermore, the annotation from databases lack of a standard way to label transcription factors. Researchers have attempted to identify putative TFs based on their Pfam domain information. For example, a method to predict a group of putative TFs purely based on whether they contain DNA-binding domains was adopted by Zupicich *et al.* [5] . In a recent work of Stegmaier *et al.* [6] , the group developed a library of specific hidden Markov models to represent TF DNA-binding domains based on the annotated TRANSFAC [7] entries.

In this study, we proposed a intelligent extraction method to identify a large portion of transcription factors based on annotation data, incorporating gene ontology terms, Pfam domains as well as keywords. The method made use of current knowledge on protein classification to build a classifier recognizing transcription factors from genomic databases.

2 Method

2.1 Dataset Preparation

The study of model genomes such as *Saccharomyces cerevisiae* provides genome-wide identification of protein functional categories, which can serve as a blueprint

Table 1. Datasets used to build TF classifier

Datasets for SVMs training	TF set	None-TF set
Human Protein Reference Database	1614	7144
MIPS Yeast Genome Database	180	764
TRANSFAC	3489	0
Total	5283	7908

for predicting protein families in other genomes. In this study, we used protein classification from yeast [9], human [8] and TRANSFAC [7] to compose our training data. Protein families such as *GTP-binding proteins*, *molecular chaperones*, *protein kinases* etc. were used to compose the non-TF dataset. Detailed classification of proteins from yeast and human were listed in the supplementary document. (http://research.i2r.a-star.edu.sg/svm_tf)

Table 1 shows the datasets we compiled to train the classifier. Each TF is assigned to the positive class and each non-TF entry to the negative class. For each entry, a list of GO terms, Pfam domains as well as keywords were extracted from database and used to form the feature vectors for that entry.

GO terms were extracted from GOA database [10] and Pfam domains were acquired by combining cross references from Swiss-Prot, Entrez Gene to Pfam database. Keywords were taken from Swiss-Prot annotation.

2.2 Build TF Classifier Using SVMs

We used the SVMlight software [11] to implement our method. SVMlight is an implementation of Vapnik's Support Vector Machine [12] for the problems of pattern recognition.

Our method forms the feature space by three groups of features. F_{domain} – Pfam domain features, provide knowledge about the functional unit (conserved motif) of the protein. F_{GO} – Gene ontology features, represent the ontology assignment of the encoded gene; and $F_{keyword}$ – keywords features offered current knowledge about the gene. The value of the features was defined as follow:

$$F_{domain} = \begin{cases} N_{domain} \\ 0 \end{cases} \tag{1}$$

$$F_{GO} = \begin{cases} D_{max} \\ 0 \end{cases} \tag{2}$$

$$F_{domain} = \begin{cases} 1 \\ 0 \end{cases} \tag{3}$$

For Pfam domain features, we defined the occurrence of the domain (how many times the domain motif presents in the sequence) as the feature value. For GO terms, we looked at the depth of the GO node, which is, how far the GO node descent away from the ontology root, as the depth implies the specificity of the ontology annotation. For GO term with multiple parents, various paths may have different value of depth, we choose the greatest one.

We used radial-basis function kernel for SVM and inductive training. We performed a 10-fold cross-validation experiment which obtained average sensitivity of 93.4% and average positive predictive value of 97.5% in recognition of TFs. Then we retrained our SVM with the whole set.

3 Result

The trained SVM TF classifier was applied to Swiss-Prot and Entrez Gene entries to extract TFs. The results are summarized in Table 2. The overlap of entries between TFs extracted from Swiss-Prot and from Entrez Gene is 2097, meaning that we have identified in total 13826 unique TF entries from these two databases.

3.1 Compare Our Method with Database Queries

To evaluate our method, we compare our extraction result with comprehensive queries. To do this, we made several queries to Swiss-Prot and Entrez Gene to identify groups of TF. Table 2 illustrates various queries made in this study. (Quries made on date 20/12/2005)

Table 2. Summary of various queries and extraction results

	query term	result	error rate
Sws_text	*transcription Factor* (full text search)	6448	21.2%
Sws_GO	GO:0003700	824	-
Sws_trans	*crosslink to TRANSFAC*	2522	-
Sws_all	merge Sws_text, Sws_GO and Sws_trans	7344	-
EG_text	*transcription factor* (full text search)	14093	28.6%
EG_GO	GO:0003700	6895	-
	Entries recognized by out method	result	error rate
Sws_svm	Swiss-prot TFs predicted by our SVMs	10975	3.2%
EG_svm	Entrez TFs predicted by out SVMs	4948	2.2%
All_svm	combined total unique TF entries	13826	-

We calculate the overlaps from each set. Figure 1 illustrates the overlaps for TFs predicted from Swiss-Prot. From the graph we can see that our extraction method can pick largest amount of TFs.

3.2 Accuracy Evaluation

We manually inspected how accurate were the queries as well as our methods. To perform the evaluation, we randomly picked 500 entries from each group for manual check. This manual checking was done by a biologist, based on the

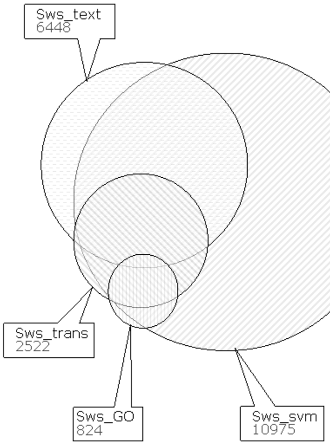

Fig. 1. Overlapped, missed and additional predicted entries based on different queries and our methods applied in Swiss-Prot

description of the entry available, as well as PubMed literature. The error rate was calculated as:

$$E = \frac{N_{none\_TF}}{N_{none\_TF} + N_{TF}} \tag{4}$$

Table 2 lists the error rate for different queries and for our method. We can see that the error rate of text search is quite high. The reason is that search engine simply does pattern matching to detect existence of "transcription factor" without semantic interpretation. Out method achieved good result and the prediction is highly reliable. The visualization representation about the number of TFs identified via various methods can be found in the supplementary file (as above).

4 Discussion

In this study we addressed the question on how to identify protein-specific information from genomic databases with incomplete functional annotation. We have done this in a context of recognition of TF entries in Swiss-Prot and Entrez Gene.

In the process of annotation of proteins by GO categories through GOA project, protein domain information is utilized through InterProScan [13] engine in the annotation process. However, while that system collects domain information from various databases that provide them, it makes no sophisticated assessment of whether the listed domains should or should not confer the implied protein functionality. The annotation uses only information about the presence of particular domains in the protein. Moreover, GO classification may classify TF-producing genes into different categories, making it impossible to use any particular GO term or a combination of GO terms to identify all TFs. For example, only around 50% of TF proteins in TRANSFAC were categorized into ”transcription factor activity”.

Also, one should be aware that the functions assigned to the proteins or genes are those that are most well known at the moment of annotation. Thus, although Swiss-Prot, Entrez Gene and GO are manually curated, this does not imply that every aspect of protein and gene functionality is captured in the entry information. For example, Q01525, a protein identified as TF by TRANSFAC, was annotated as *protein domain specific binding* (GO:0019904) from which one can not directly infer that it is a TF. Also, from Table 2 and Figure 1 we can observe that specific queries relative to GO terms are missing many existing entries that are known TFs.

On the other hand, description of protein functionality through GO categories, although not necessarily explicitly suggest TF activity for the protein, may describe aspects that are related to TF activity. For example, genes categorized by *DNA binding* (GO:0003677) have 35% co-annotation in *regulation of transcription, DNA-dependent* (GO: 0006355). Also, since TFs are a group of proteins that contains different combination of domains and perform various functions (of which many characterize protein as TFs and its activity as TF activity), simply by constructing a query (even if it’s a well-constructed one) to search Swiss-Prot and Entrez Gene one will not be able to get a satisfactory TF lists.

For all these reasons, we have considered a combination of GO category descriptions associated with an entry in Swiss-Prot or Entrez Gene, and Pfam domains, as a valuable basis that can reveal the essential knowledge on proteins/gene product activity.

Since in our method the positive predictive value is greater than 97%, we can expect approximately one wrongly identified non-TF entry among 40 entries identified as TF by our system. However, one should be careful with such generalizations and note that this is an optimistic one since not all non-TF families have been used in the training of our system.

On the other hand, the sensitivity obtained is rather high, 93.4%. However, again, one should be careful in interpreting this score, since it is given for the entries that had either GO category ascribed, or protein domains, or both. In general, the expected (absolute) sensitivity should be lower.

In spite of these considerations, we did show that our method allows for efficient accurate extraction of TF entries from the two considered public resources. With a total of over 13826 unique entries extracted, we were able to extract considerably more entries than contained in TRANSFAC Professional v.8.4 that contains 5919 TF entries. Although it is not possible to directly compare the number of entries in TRANSFAC (since they were manually curated) and our extracted entries, we observe that our method potentially allows extraction of very high quality (putative) TF entries, with 2% error rate. Also, one should note that the predictions made in this study are biased as they relate toward eukaryotic species, since the training data were gathered from eukaryotic organisms. However, the same method should work for prokaryotic transcription factors, if the data is available, although the features should be regenerated and the system should be retrained.

Finally, we can apply our method also to the predicted genes and proteins, as long as they contain the relevant TF domains. This may help in provisional association of TF function in some cases.

References

1. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M., Martin,M.J., Natale,D.A., O'Donovan,C., Redaschi,N., Yeh,L.S. The Universal Protein Resource (UniProt). Nucleic Acids Res. **33** (2005) D154-159.
2. Maglott,D., Ostell,J., Pruitt,K.D., Tatusova,T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. **33** (2005) D54-8.
3. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L. et al. The Pfam protein families database. Nucleic Acids Res. (2002) **30** 276-280.
4. Harris,M.A., Clark,J., et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. (2004) **32** D258-61.
5. Zupicich,J., Brenner,S.E., Skarnes,W.C. Computational prediction of membrane-tethered transcription factors. Genome Biol. (2001) 2:0050.
6. Stegmaier,P., Kel,A.E., Wingender,E. Systematic DNA-Binding Domain Classification of Transcription Factors. Genome Inform Ser Workshop (2004) 15(2):276-86.
7. Matys,V., Wingender,E., et al. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. (2003) **31** 374-8.
8. Peri,S., Navarro.J.D., Pandey,A. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. (2003) **13** 2363-71.
9. Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Gldener,U., Mannhaupt,G., Mnsterktter,M., Pagel,P., Strack,N., Stmpflen,V., Warfsmann,J. and Ruepp,A. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res. (2004) **32** D41-4.
10. Camon, E., Magrane, M., Barrell, D., Lee V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., Apweiler R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res. (2004) **32** D262-D266.
11. Scholkopf,B., Burges,C., Smola,A. Advances in Kernel Methods - Support Vector Learning, MIT-Press. (1990)
12. Vapnik,V.N. The Nature of Statistical Learning Theory. Springer. (1995)
13. Zdobnov E.M. and Apweiler R. InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics, (2001) **17** 847-8.