Available online at www.sciencedirect.com
SCIENCE DIRECT
Biochemical and Biophysical Research Communications 347 (2006) 141–144
BBRC
www.elsevier.com/locate/ybbrc

Automatic transcription factor classifier based on functional domain composition

Ziliang Qian [a,b], Yu-Dong Cai [c,d,*], Yixue Li [a,e,*]

[a] Bioinformatics Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China
[b] Graduate School of the Chinese Academy of Sciences, 19 Yuquan Road, Beijing 100039, China
[c] CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, China
[d] Biomolecular Sciences Department, University of Manchester, Institute of Science and Technology, P.O. Box 88, Manchester M60 1QD, UK
[e] Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, 200235 Shanghai, China

Received 7 June 2006
Available online 21 June 2006

Abstract

To understand the transcriptional regulatory mechanism, it is indispensable to identify transcription factors (TF) from the whole genome and to classify transcription factors into different classes. New computational approaches have been developed to identify TFs/non-TFs, and furthermore to classify TFs into four different classes, based on the protein functional domain composition [K.C. Chou, Y.D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, J. Biol. Chem. 277 (2002) 45765–45769]. We trained and tested our method on a non-redundancy dataset consisting of 74 transcription factors collected from TRANSFAC v7.0 [V. Matys, O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel, E. Wingender, TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes, Nucleic Acids Res. 34 (2006) D108–D110] and 1558 non-transcription factors from UniProtKB/Swiss-Prot Release 49.3 of 21-Mar-2006. The overall success rates of jackknife cross-validation tests reached 98.4% for TF/non-TF identification and 97.2% for classifications of TF classes: basic domains, zinc-coordinating DNA-binding domains, helix-turn-helix, and β-scaffold factors.

Keywords: Transcription factors; Functional domain composition; Intimate sorting classifier; Jackknife cross-validation test

Transcription factor (TF) is often termed as the major regulator of transcription. Most of them (or dimers) bind to specific DNA fragments using their DNA-binding domain and modulate nearby genes' transcription through their trans-activating/repressing domains. Generally, transcription factors can be classified into four major classes [2–4] (cf. Fig. 1): (1) Basic domains. (2) Zinc-coordinating DNA-binding domains. (3) Helix-turn-helix. (4) β-Scaffold factors with Minor Groove Contacts. Given a newly identified protein with poor prior knowledge, the following two questions are often raised: (1) Is it a transcription factor? (2) Which class it belongs to? Both are very important to understand its transcriptional regulatory function.

Previous works on TF identification/classification were remarkably based on manual annotations [4], e.g., experimental data on transcription regulatory activity, protein structure and whole sequence homology, etc. However, collecting protein annotations manually is a time-consuming task. To overcome this problem, in this research we developed an automatic method to discriminate TFs from non-TFs and furthermore to classify TFs into four categories mentioned above based on protein functional domain composition, which has already been successfully used to predict protein–protein interaction [5], protein structure [6], protein function [7–9], protein subcellular location [1], etc. The classifier we built got a fairly good performance

* Corresponding authors.
E-mail addresses: cyd@picb.ac.cn (YD Cai), yxli@sibs.ac.cn (YX Li).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.06.060

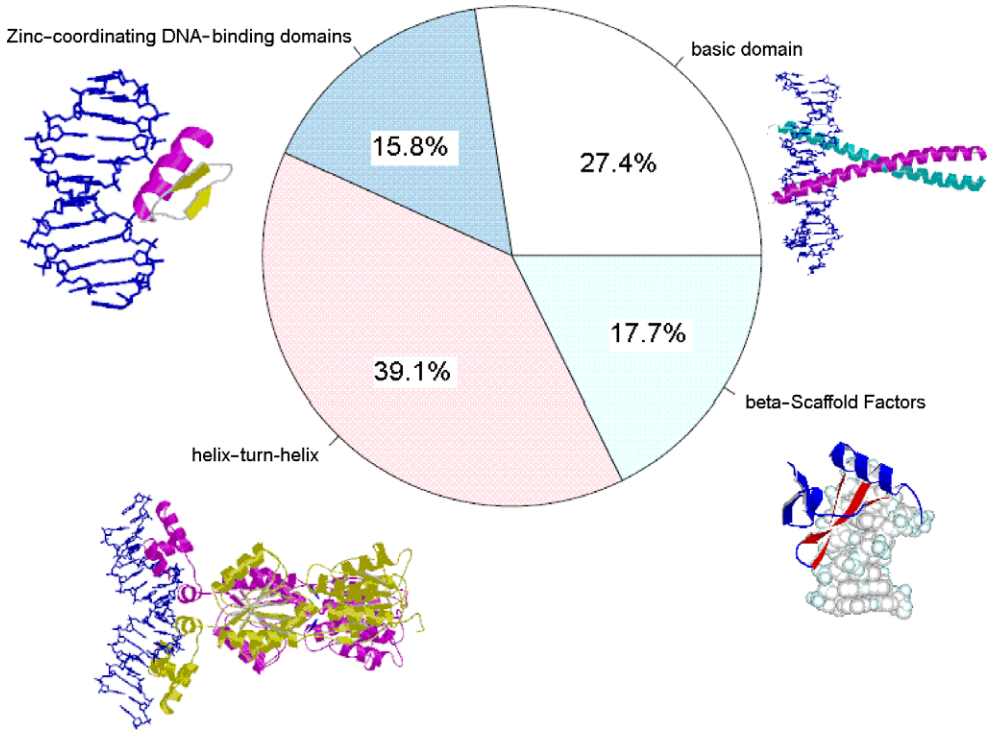

Fig. 1. Transcription factor classification. Transcription factors are generally classified into four distinct classes. Left-up, zinc-coordinating DNA-binding domains. Right-up, basic domains. Left-bottom, helix-turn-helix. Right-bottom, β-scaffold factors. 3D structures of protein–DNA complexes were adapted from [18]. Proportion was calculated based on TRANSFAC(v7.0) [2].

with overall success rates of 98.4%, 97.2% for TF/non-TF identification and TF classification, respectively.

Materials and methods

TF/non-TF datasets. First, for transcription factors (TFs) with classification information, the dataset came from TRANSFAC v7.0 [2], and for non-TF, the dataset was randomly selected from UniProtKB/Swiss-Prot Release 49.3 of 21-Mar-2006 by using keyword "membrane," "secretory," "antigen," "transferase," "kinase." All together, a dataset with 1176 transcription factors and 29,295 non-transcription factors was built. And then we refined this dataset as follows: (1) Filter out proteins with a length over 5000 aa or less than 50 aa and those without SwissProt accession number. (2) Remove the redundancy against homology bias using the programs cd-hit [10,11] and PISCES [12]. As a consequence, none of the sequences investigated have more than 25% sequence identity. Finally, a positive dataset with 84 TFs with known classification information and a negative dataset with 2167 non-transcription factor proteins were obtained (Table 1).

Functional domain composition feature vector. To facilitate a feasible statistical classifier, each transcription factor (TF) must be expressed in terms of a set of discrete numbers instead of whole amino-acid sequence to catch the core features intimately related to biological functions. Because TFs are classified according to their structures and functions, it is anticipated that the prediction quality will be enhanced if we can find a feasible approach to use the knowledge of structural and functional domains to define a transcription factor sample, such as DNA-binding domain(s), oligomerization domain(s), and trans-activating domain. This can be realized through the integrated domain and motif database, or the interPro databases at [http://www.ebi.ac.uk/interpro] through the following steps.

Table 1
The original dataset

Dataset		Size
Transcription factors (TF)	Basic domain	21
	Zinc-coordinate	15
	Helix-turn-helix	36
	β-Scaffold	12
	Overall	84
Non-TF		2167
Total		2251

Step 1. Extract domains information of a protein from InterPro by using the Protein2ipr mapping provided. For our TF/non-TF dataset, Protein2ipr release 12.0 on Friday November 18th 2005 [ftp://ftp.ebi.ac.uk/pub/databases/interpro/] was used. The result totally covered 8151 InterPro entries with well-known structural and functional domain types.

Step 2. With each of the 8151 functional domain patterns as a vector-base, the sample of a TF can be represented in a 8151D (dimensional) vector as: If there is a hit, e.g., transcription factor P49716 contains IPR004827 which is the 1970th record of the 8151 domains, then the 1970th component of the transcription factor P49716 in the 8151D feature space is set to 1; otherwise 0.

Step 3. Then feature vector T for a given TF can thus be explicitly formulated as

$$T = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_i \\ \vdots \\ t_{8151} \end{pmatrix}, \quad (1)$$

where,

$$t_i = \begin{cases} 1, & \text{hit found,} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Defined in this way, each transcription factor will correspond to a 8151D vector T with each of the 8151 functional domains as the base for the vector space. In other words, rather than the amino acid composition approach or pseudo-amino acid composition as often used by previous investigators [7,13,14], a TF is now represented in terms of the functional domain composition. By doing so, not only some sequence-related features but also some function-related features are naturally incorporated in the representation.

Intimate sorting (ISort) classifier. The prediction was performed with the ISort classifier, which can be briefly described as follows. Suppose there are N transcription factors (TFs) $T_1, T_2, \ldots, T_n, \ldots, T_N$ which have already been classified into categories $1, 2, \ldots, c(n), \ldots, \mu$, $c(n)$ is the category of T_n. Now, for a query TF T, how can we predict which class it belong to? To deal with this problem, let us define the following scale to measure the similarity between T and T_i $(i = 1, 2, \ldots, N)$

$$\Lambda(T, T_i) = \frac{T \cdot T_i}{\|T\| \cdot \|T_i\|}, \quad (3)$$

where $T \cdot T_i$ is the dot product of T and T_i, and $\|T\|$ and $\|T_i\|$ are their moduli, respectively. Obviously, when $T \equiv T_i$, we have $\Lambda(T, T_i) = 1$, meaning they have perfect 100% similarity. Generally speaking, the similarity is within the range of 0 and 1, $0 \leqslant \Lambda(T, T_i) \leqslant 1$. Accordingly, the ISort predictor can be formulated as follows. If the similarity between T and T_k $(k = 1, 2, \ldots, N)$ is the highest, i.e.,

$$\Lambda(T, T_k) = \max\{\Lambda(T, T_1), \Lambda(T, T_2), \ldots, \Lambda(T, T_N)\}, \quad (4)$$

where the operator max means taking the maximum one among those in the brackets, then the transcription factor T is predicted belonging to the same category as of T_k. If there is a tie, i.e.,

$$\Lambda(T, T_{k_1}) = \Lambda(T, T_{k_2}) = \max\{\Lambda(T, T_1), \Lambda(T, T_2), \ldots, \Lambda(T, T_N)\}, \quad (5)$$

the query transcription factor T cannot be uniquely determined. In these rare cases, T will be randomly classified into category as of either T_{k_1} or T_{k_2}.

Results and discussion

Two 8151D ISort classifiers were built, one for identifying TF/non-TFs and another for further classifying TFs into four different categories: basic domains, zinc-coordinating DNA-binding domains, helix-turn-helix, and β-scaffold factors. According to step 1, step 2, and step 3 mentioned above, we obtain the following results. (1) For TF/non-TF identification, with the exclusion of proteins that have no functional domain annotation and except orphans that have domains occurring only once in our original dataset, 8151D feature vectors were built for 74 TFs and 1558 non-TFs (cf. Table 2, Supplement file). (2) For TF classification, three more TFs were filtered because of orphans, thus 8151D feature vectors were built for 71 TFs (cf. Table 3, Supplement file).

Jackknife cross-validation test was adopted to examine the performance of our predictor. In statistical prediction, the single independent dataset test, sub-sampling test, and the Jackknife test are the three cross validation approaches often used to examine the power of a predictor. Of these three, the Jackknife test is deemed as the most objective and rigorous one and hence adopted by more and more investigators [15–17]. In our implementations, Jackknife cross-validation tests were operated as follows:

For identifying TF/non-TF, for each protein T in the dataset consisting of 74 TFs and 1558 non-TFs, we applied the first ISort classifier to predict T's property (TF/non-TF) using the rest proteins excluding T. Classifier succeeded if it correctly predicted the property of T. Then the success rate for TF/non-TF identification was given according to the following formulas:

$$\begin{cases} \text{Success rate for TF} = \dfrac{\text{Correctly predicted TF}}{\text{True TF}}. \\ \text{Success rate for non-TF} = \dfrac{\text{Correctly predicted non-TF}}{\text{True non-TF}}. \\ \text{Success rate for overall} = \dfrac{\text{Correctly predicted}}{\text{Total}}. \end{cases} \quad (6)$$

For classifying TFs into four different classes, for each protein T in the dataset consisting of 74 TFs, we applied the second ISort classifier to predict T's classification using the rest proteins excluding T. Classifier succeeded if it correctly predicted the classification of T. Then the success rate for TF classification was given according to the following formulas:

$$\begin{cases} \text{Success rate for ``basic domain''} = \dfrac{\text{Correctly predicted ``basic domain''}}{\text{True ``basic domain''}}. \\ \text{Success rate for ``zinc-coordinating''} = \dfrac{\text{Correctly predicted ``zinc-coordinating''}}{\text{True ``zinc-coordinate''}}. \\ \text{Success rate for ``helix-turn-helix''} = \dfrac{\text{Correctly predicted ``helix-turn-helix''}}{\text{True ``helix-turn-helix''}}. \\ \text{Success rate for ``beta-scaffold''} = \dfrac{\text{Correctly predicted ``beta-scaffold''}}{\text{True ``beta-scaffold''}}. \\ \text{Success rate for overall} = \dfrac{\text{Correctly predicted}}{\text{Total}}. \end{cases} \quad (7)$$

Tables 2 and 3 give the success rates of Jackknife cross-validation test for TF/non-TF identification and TF

Table 2
The performances of TF/non-TF identification

Category	Jackknife test success rate
TF	66/74 = 89.2%
Non-TF	1540/1558 = 98.8%
Overall	1606/1632 = 98.4%

Jackknife test successful rates in identifying TF/non-TFs. 74 out of 84 transcription factors (TF) and 1558 out of 2167 non-TF left after removing proteins which have no functional domain and removing orphans which have domains that occurred only once in our dataset.

Table 3
The performances of TF classification

Classification	Jackknife test success rate
Basic domain	20/20 = 100%
Zinc-coordinate	10/11 = 90.9%
Helix-turn-helix	33/33 = 100%
β-Scaffold	6/7 = 85.7%
Overall	69/71 = 97.2%

Jackknife test successful rates in classifying TFs into four different classes. 71 out of 84 transcription factors left after removing proteins which have no functional domain and removing orphans which have domains that occurred only once in our dataset.

classification, respectively. Our predictors got a very good performance. As shown in Table 2, the success rates were 89.2%, 98.8% for TF and non-TF identification, respectively, and 98.4% overall. As shown in Table 3, the success rates reached 100%, 90.9%, 100% and 85.7% for basic domain TFs, zinc-coordinating TFs, helix-turn-helix TFs, and β-scaffold TFs, respectively, meanwhile 97.2% overall. These cheerful results demonstrate that domain composition is a very effective means to characterize the features of TF for classification.

The computation was performed in a Silicon Graphics IRIS Indigo workstation (Elan 4000).

Conclusion

To understand transcription regulation mechanism, there is a great demand to identify TFs and further classify TFs into functional categories. Therefore, we built two automatic classifiers, respectively, in this contribution. We got a fairly good result, 98.4% for TF/non-TF identification and 97.2% for four TF classifications, which means that our classifiers can be used as a good support in TF annotations as the amount of protein sequences increases rapidly in the post-genomic era.

Acknowledgments

This work was funded by National "973" Basic Research Program (2004CB518606, 2003CB715900, and 2001CB510209) and National "863" High-Tech R&D Program (2004BA711A21).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2006.06.060.

References

[1] K.C. Chou, Y.D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, J. Biol. Chem. 277 (2002) 45765–45769.
[2] V. Matys, O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel, E. Wingender, TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes, Nucleic Acids Res. 34 (2006) D108–D110.
[3] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A.E. Kel, O.V. Kel-Margoulis, D.U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, E. Wingender, TRANSFAC(R): transcriptional regulation, from patterns to profiles, Nucleic Acids Res. 31 (2003) 374–378.
[4] E. Wingender, Classification scheme of eukaryotic transcription factors Molekularnaya, Biologiya 31 (1997) 483–497.
[5] J. Wojcik, V. Schachter, Protein–protein interaction map inference using interacting domain profile pairs, Bioinformatics 17 (2001) S296–S305.
[6] K.C. Chou, Y.D. Cai, Predicting protein structural class by functional domain composition, Biochem. Biophys. Res. Commun. 321 (2004) 1007–1009.
[7] Y.D. Cai, K.C. Chou, Predicting enzyme subclass by functional domain composition and pseudo amino acid composition, J. Proteome Res. 4 (2005) 967–971.
[8] Y.D. Cai, A.J. Doig, Prediction of *Saccharomyces cerevisiae* protein functional class from functional domain composition, Bioinformatics 20 (2004) 1292–1300.
[9] X.J. Yu, J.C. Lin, T.L. Shi, Y.X. Li, A novel domain-based method for predicting the functional classes of proteins, Chinese Sci. Bull. 49 (2004) 2379–2384.
[10] W. Li, L. Jaroszewski, A. Godzik, Tolerating some redundancy significantly speeds up clustering of large protein databases, Bioinformatics 18 (2002) 77–82.
[11] W. Li, L. Jaroszewski, A. Godzik, Clustering of highly homologous sequences to reduce the size of large protein databases, Bioinformatics 17 (2001) 282–283.
[12] G. Wang, R.L. Dunbrack Jr., PISCES: a protein sequence culling server, Bioinformatics 19 (2003) 1589–1591.
[13] G.P. Zhou, An intriguing controversy over protein structural class prediction, J. Protein Chem. 17 (1998) 729–738.
[14] K.C. Chou, A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space, Proteins: Struct. Funct. Genet. 21 (1995) 319–344.
[15] Y.D. Cai, K.C. Chou, Predicting membrane protein type by functional domain composition and pseudo-amino acid composition, J. Theor. Biol. 238 (2006) 395–400.
[16] Y.D. Cai, K.C. Chou, Predicting membrane protein type by functional domain composition and pseudo-amino acid composition, J. Theor. Biol. 238 (2006) 395–400.
[17] X. Yu, C. Wang, Y. Li, Classification of protein quaternary structure by functional domain composition BMC, Bioinformatics 7 (2006) 187.
[18] N.M. Luscombe, S.E. Austin, M. Helen, J.M. Thornton, An overview of the structures of protein-DNA complexes, Genome Biol. 1 (2000), reviews 001.