D74–D81 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj131

DBD: a transcription factor prediction database

Sarah K. Kummerfeld* and Sarah A. Teichmann

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Received August 12, 2005; Revised and Accepted October 24, 2005

ABSTRACT

Regulation of gene expression influences almost all biological processes in an organism; sequence-specific DNA-binding transcription factors are critical to this control. For most genomes, the repertoire of transcription factors is only partially known. Hitherto transcription factor identification has been largely based on genome annotation pipelines that use pairwise sequence comparisons, which detect only those factors similar to known genes, or on functional classification schemes that amalgamate many types of proteins into the category of 'transcription factor'. Using a novel transcription factor identification method, the DBD transcription factor database fills this void, providing genome-wide transcription factor predictions for organisms from across the tree of life. The prediction method behind DBD identifies sequence-specific DNA-binding transcription factors through homology using profile hidden Markov models (HMMs) of domains. Thus, it is limited to factors that are homologus to those HMMs. The collection of HMMs is taken from two existing databases (Pfam and SUPERFAMILY), and is limited to models that exclusively detect transcription factors that specifically recognize DNA sequences. It does not include basal transcription factors or chromatin-associated proteins, for instance. Based on comparison with experimentally verified annotation, the prediction procedure is between 95 and 99% accurate. Between one quarter and one-half of our genome-wide predicted transcription factors represent previously uncharacterized proteins. The DBD (www.transcriptionfactor.org) consists of predicted transcription factor repertoires for 150 completely sequenced genomes, their domain assignments and the hand curated list of DNA-binding domain HMMs. Users can browse, search or download the predictions by genome, domain family or sequence identifier, view families of transcription factors based on domain architecture and receive predictions for a protein sequence.

INTRODUCTION

The essence of any organism is the spatial and temporal expression pattern of its gene repertoire. While the genome provides the template, it is the way genes are expressed that defines the organism. Consequently, regulation of gene expression influences almost all biological processes in an organism.

Transcription factors (TFs) are often termed the master regulators of gene expression. By binding to the DNA, they tightly control where and when the nearby target gene is expressed. Despite their importance as a fundamental component of biological systems for all organisms across the tree of life, the transcription factor repertoires for many genomes remain largely uncharted. Hitherto transcription factor identification has been largely based on genome annotation pipelines that use pairwise sequence comparisons (1), which detect only those factors similar to known genes, or on functional classification schemes that amalgamate many types of proteins into the category of 'transcription factor' (2). Using a novel transcription factor identification method, our online resources, the DBD transcription factor database provides transcription factor predictions for all completely sequenced genomes.

Databases of transcription factors to date have focused on single or small groups of genomes. They are largely based on manual literature curation, pairwise sequence comparison and functional classification schemes. Genome specific resources include: RegulonDB for *Escherichia coli K-12* (1), DBTBS for *Bacillus subtilis* (3), FlyBase (providing TF as well as other annotation) for *Drosophila* (4), TFdb for mouse (5) and TRANSFAC for eukaryotes (6). RegulonDB and DBTBS are databases of transcription factors and their target genes for their respective genomes (*E.coli K-12* and *B.subtilis*). DBTBS also provides information about the Pfam domains, but this is purely extra information and is not used for prediction. FlyBase is a more general resource for *Drosophila* that compiles information from the fly genome projects and liter-

*To whom correspondence should be addressed. Tel: +44 1223 402479; Fax: +44 1223 213556; Email: skk@mrc-lmb.cam.ac.uk

ature curated annotation, including listings of known transcription factors. TFdb is a database of mouse transcription factors. It is built by: selecting proteins annotated by Gene Ontology (GO) as transcription factors, manual curation and addition of close homologs using pairwise sequence comparison. The manual curation involves addition of known TFs that are missed by GO and removal of those that seem to be erroneously classified by GO. Similarly, TRANSFAC (6) is list of eukaryotic transcription factors based on manual literature curation.

Others have made use of these resources to compile their own lists, for example, Messina *et al*. (7) used TRANSFAC together with GO annotation of UniProt and FlyBase to seed sequence and hidden Markov models (HMMs) searches, followed by manual curation, to identify human factors. Their aim was to produce a rough list of factors as a starting point for array experiments across species. Because the experiments were large-scale surveys, the study favoured over- rather than under-prediction. Only blatant errors, for instance DNA or RNA polymerases were manually removed. This liberal approach to false positives meant that the final set included a range of proteins that are not sequence-specific DNA-binding transcription factors. Another study by Riechmann *et al*. (8) used curated lists of factors in combination with BLAST sequence searches to identify the transcription factors in four eukaryotes: *Arabidopsis*, *Drosophila*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*.

A slightly different approach was taken by Iyer *et al*. (9). They used the multiple sequence comparison tool PSI-BLAST, seeded with known viral regulatory proteins, to identify viral transcription factors. Ravasi *et al*. (10) also used multiple sequence comparisons, focusing on zinc finger transcription factors in mouse. Computational studies of transcriptional regulation have used domain assignments in an *ad hoc* (uncalibrated) way to identify transcription factor proteins for particular groups of genomes (11,12). Finally, TrsDB (13) use position specific scoring matrices describing DNA-binding motifs to identify and classify transcription factors for nine eukaryotic genomes.

All of these resources list transcription factors for an individual or small group of organisms. Their underlying approach is identification by literature review, which means the proteins identified must be known (and published) factors. For most of the datasets the only scope for inclusion of uncharacterized transcription factors is via pairwise sequence searches, capable of identifying close homologs. Some resources also use functional classification schemes; however, these are firstly prone to error (due to inclusion of regulatory but non-DNA-binding factors) and secondly, they too are produced by literature review and pairwise sequence search. These methods are not comprehensive with respect to either the genomes they cover, transcription factor families included or both.

A second group of resources include compilations and predictions of transcription factor binding sites: e.g. MATCH (14), JASPAR (15) and MAPPER (16). While these tools are not directly comparable to our database, they are complementary, providing information about the DNA sequences that transcription factors recognize.

The prediction method described here is applicable to all genomes across the tree of life. It has been quantitatively evaluated and is capable of accurately identifying both known and previously uncharacterized transcription factors that bind specifically to DNA. Even TFs with no obvious sequence homology to known factors may be identified.

The prediction method behind DBD identifies sequence-specific DNA-binding transcription factors through homology using profile HMMs of domains. The collection of HMMs is taken from two existing databases (Pfam and SUPERFAMILY), and is limited to models that exclusively detect transcription factors that specifically recognize DNA sequences. It does not include basal transcription factors or chromatin-associated proteins, for instance. Based on comparison with experimentally verified annotation, the prediction procedure is between 95 and 99% accurate. Between one-quarter and one-half of the genome-wide predictions represent previously uncharacterized proteins.

At present, DBD consists of predicted transcription factor repertoires for more than 150 completely sequenced genomes (to be periodically updated), their domain assignments and the hand-curated list of DNA-binding domain HMMs. Users can browse predictions by genome or domain family, search using sequence identifiers and view TF domain architectures. Protein sequences can be submitted for automatic prediction and all transcription factors lists are available for download grouped by genome.

The potential applications of predictions are broad ranging, from single protein to multi-genome studies. We expect that the main use of our database will be for prediction of transcription factor repertoires for particular genomes. The predictions also provide the starting point for use in high-throughput experiments that characterize the nature of regulation. For example, measuring characteristics of genes such as expression levels across different tissues or identifying DNA-binding sites. Examples of large-scale experiments that have used TF repertoires as a starting point are studies by Messina *et al*. (7). They used microarrays to investigate expression patterns of human transcription factors. Two recent analyses carried out large-scale ChIP-chip experiments of *S.cerevisiae* transcription factors with the aim of identifying transcription factor target genes (17,18). Our predictions may also be of interest to theoretical biologists and are already being used for comparative genomics studies in fungi and insects.

We begin with a detailed explanation of the transcription factor prediction procedure and rationale for its design. The second section discusses a series of tests that were used to evaluate the performance of the method. Finally, we describe the web interface and explore the biological significance of this information.

PREDICTION METHOD

Transcription factors regulate gene expression by binding to DNA near their target genes. Some are sequence-specific, recognizing only particular DNA sequences, while others are basal, binding to a more general promoter (e.g. TATA box or initiator sequence). Here we are concerned only with the sequence-specific DNA-binding transcription factors, because these proteins are important for differential regulation of gene expression. To function as a sequence-specific DNA-binding transcription factor, a protein must contain a domain that binds to DNA in a sequence-specific manner. We exploit this requirement in order to identify transcription factors.

Our approach uses protein structure (through domains) and remote homology recognition, to accurately, and sensitively identify transcription factors. It can be automatically applied to any genome to identify both known and previously uncharacterized factors.

We use profile HMMs from the SUPERFAMILY (19) and Pfam (20) databases to identify proteins that contain sequence-specific DNA-binding domains. The advantage of transcription factor prediction based on HMMs of DNA-binding domains is two-fold. First, it is more sensitive than conventional genome annotation procedures, because it uses the powerful multiple sequence comparison method of HMMs. Secondly, it recognizes only transcription factors that use the mechanism of sequence-specific DNA binding, as opposed to functional classification schemes that amalgamate many types of proteins into the category of transcription factor (e.g. co-activators or co-repressors and chromatin modification enzymes).

The two HMM libraries that we use, SUPERFAMILY and Pfam, both represent domains, but they differ in their method of construction and definition of domains. Briefly, SUPERFAMILY contains HMMs of domains of known three-dimensional structure based on the domain definitions of the Structural Classification Of Proteins (SCOP) database (21). Each SCOP domain is used as a seed to build a model representing its family. In most cases, one SCOP superfamily is represented by a set of models that each recognize a subset of superfamily members.

In contrast, the Pfam HMMs are built from hand-curated multiple sequence alignments. Groups of sequences are identified by manual literature review as belonging to the same family, they are aligned and used as the seed for an HMM. This model is used to search a large sequence database in order to detect more distant or poorly characterized family members. The newly detected sequences are included in a second alignment which is used to build a final, broader HMM representing the family.

The variation in domain definition and method of construction means that Pfam and SUPERFAMILY differ in their coverage. By including both databases in our prediction method, we improve the overall prediction rate as compared to using either database alone. The DBD website indicates the number of transcription factors identified using each database and the TF domain architectures.

We manually inspected all 2537 SCOP (version 1.61) and 7677 Pfam (version 16.0) families, and identified 110 and 141, respectively, that represent sequence-specific DNA-binding domains. From this annotation, we selected the HMMs that represent these families from the SUPERFAMILY and Pfam databases. For Pfam, selection of relevant models was straight-forward because each family corresponds to an HMM, and these models are specifically designed to recognize only members of the family.

For SUPERFAMILY, model selection is less straightforward because SUPERFAMILY models are designed to identify members within a SCOP superfamily rather than a SCOP family. The superfamily level includes highly divergent members that often span different functions. For example, the *Putative DNA-binding domain* superfamily is made up of five families that are involved in: RNA-binding, general (non-sequence-specific) DNA-binding as well as sequence-specific DNA-binding transcription factors. For this reason, our manual curation considered SCOP families rather than superfamilies.

To overcome this problem, we selected models that were seeded by proteins classified in the SCOP database as sequence-specific DNA-binding and assessed their potential to match non-DNA-binding domains using a SCOP all-against-all test. This test involves scoring the seed sequences against the models. For example, in the case of the *Putative DNA-binding domain* superfamily all SCOP sequences were searched against the HMMs. In two cases, one of the DNA-binding family models gave a significant match (or cross-hit) to a non-DNA-binding sequence. To ensure accurate identification of sequence-specific DNA-binding transcription factors, we excluded the cross-hitting models. In total 13 models representing 12 families and 5 superfamilies were excluded (Supplementary Table 1).

Separate from these cross-hits, there are a small number of families where the overwhelming majority of members are sequence-specific DNA-binding domains, but some representatives have other functions (possibly in addition to their DNA-binding role). For example, C2H2 zinc fingers may bind RNA rather than DNA and proteins containing a zinc finger domain carry out multiple functions (including but not exclusively sequence-specific transcription regulation) (10). In these rare cases, we include the domain (and its HMM) and accept that this may generate a small number of false positives.

This process allowed selection of 141 Pfam and 210 SUPERFAMILY models (110 families) representing sequence-specific DNA-binding domains. To make a prediction as to whether a protein is a transcription factor, we search the amino acid sequence against the HMM libraries and designate the protein to be a transcription factor if it has a significant match to a model we annotated as representing a sequence-specific DNA-binding domain. This procedure is illustrated in Figure 1. Note that only transcription factors from these families will be detected. Characterized TFs that are not recognized by the HMMs are not automatically included.

EVALUATION

To evaluate the accuracy of the prediction process, we carried out a series of tests on groups of sequences that had been experimentally annotated as transcription factors. The first test considers only proteins of known structure in order to check our annotation of SCOP domains. The second test considers the largest available set of 1.5 million proteins including sequences from across the tree of life, providing a large-scale assessment of the HMM-based prediction in order to determine our accuracy and coverage statistics. The final set of tests focuses on individual genomes, evaluating performance in comparison to manually curated lists of factors. As discussed above, the primary use of our database is expected to be for prediction of transcription factor repertoires for individual organisms. This final test is designed to directly assess our performance on whole genomes, allowing users to ascertain the level of confidence they should expect for repertoire predictions.

The aim of the first test was to assess the accuracy of the underlying approach (that is, transcription factor identification

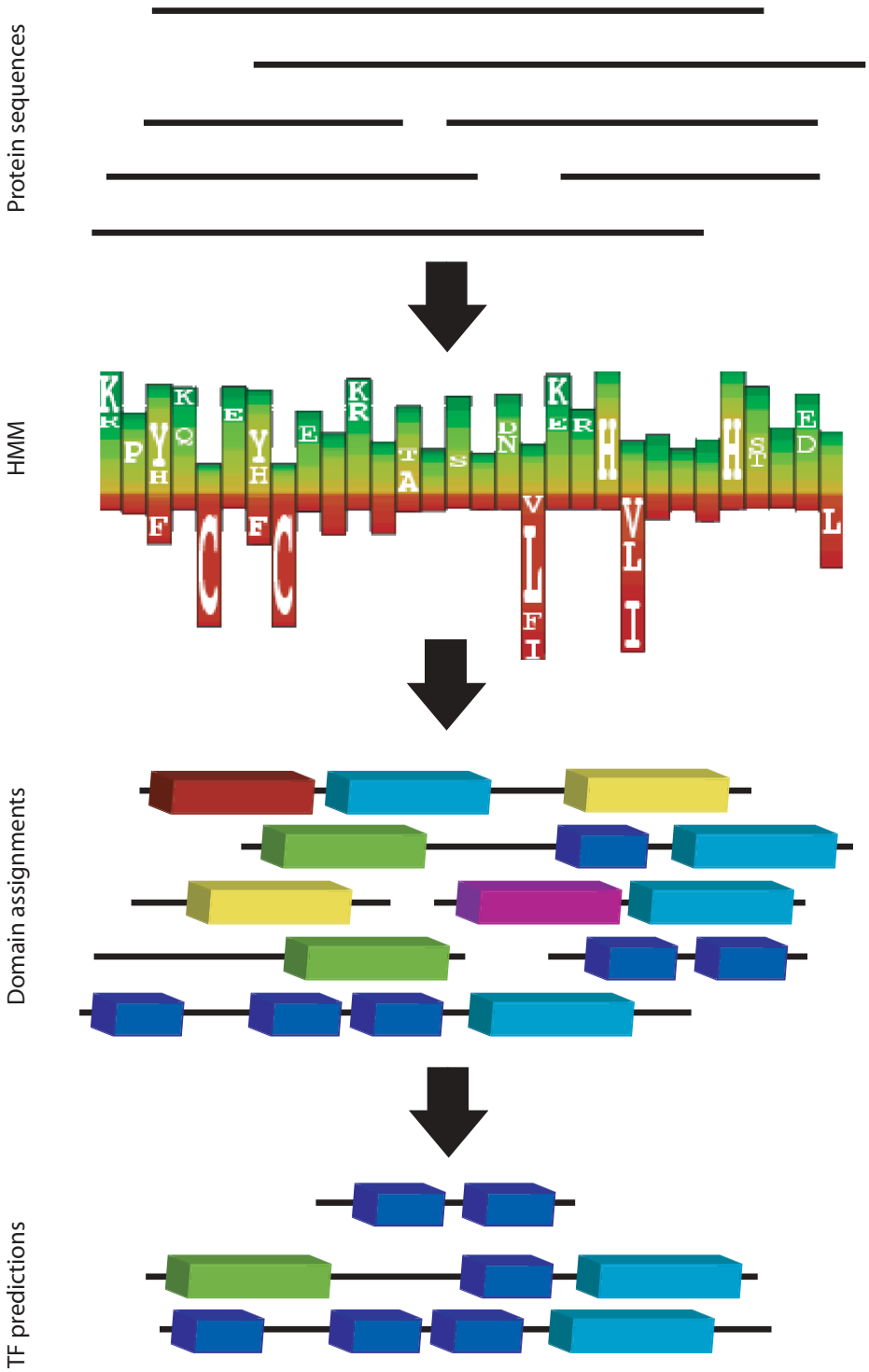

Figure 1. Transcription factor prediction procedure. We begin with a set of proteins, shown as horizontal lines. For example, the initial set of proteins may be a whole proteome. Each sequence is searched against the SUPERFAMILY and Pfam HMM libraries. A domain is assigned to a particular protein when one of the HMMs matches a region of sequence with an *E*-value less than or equal to 0.001 for SUPERFAMILY or greater than or equal to the trusted cutoff for PFAM. Assigned domains are shown as coloured boxes where the colour indicates the family. For example, the small dark-blue boxes represent the Zinc finger C2H2 type DNA-binding domains. Proteins with at least one DNA-binding domain assigned are selected as putative transcription factors. The designation of DNA-binding is based on our manual curation of Pfam and SUPERFAMILY models.

via manual inspection of SCOP), without adding the complexity of domain prediction. The sequence set was from the PDB (22), including only proteins of known structure with curated domain composition from SCOP. By including only proteins with known domain composition, we eliminated any potential error introduced by incorrect assignments by the HMMs.

We used the GO annotation (2) of the PDB proteins as a standard list of known TFs. The GO functional classes that

represent the transcription factors are shown in Table 1 (Supplementary Table 2 provides a comprehensive list, including categories we classified as *expression related*). It should be noted here that when we manually inspected proteins classified by GO as transcription factors, we found that the set also includes some basal (i.e. non-sequence-specific) factors and chromatin remodelling proteins.

When we examined PDB proteins identified by us as containing a sequence-specific DNA-binding domain, we found that more than 99% (393) are classified by GO as TFs. The remaining 1% (4) are classified by GO as nucleic acid binding and have not been allocated to a GO sub-category (Details are shown in Table 2). This test illustrates the validity of both the underlying approach, prediction based on structural domains, and our hand curation of the SCOP domains.

Next, we aimed to evaluate the prediction method as a whole, including the domain assignment step using SUPERFAMILY and Pfam. The sequence set used was from the UniProt database (23), the most comprehensive catalogue of proteins available including more than 1.5 million sequences. As a standard for comparison, we used the experimentally verified GO annotation for UniProt (that is we excluded homology based annotation). We searched the Pfam and SUPERFAMILY HMMs against the UniProt sequence set to derive a set of predicted transcription factors. In order to evaluate the accuracy of our method, we calculated the number of predicted TFs for which GO supported our prediction. That is, GO annotated the protein as being a member of one of the categories shown in Supplementary Table 2. This benchmark established our accuracy to be between 95 and 99% (Table 3). This means that we expect 5 out of 100 of our predictions to be incorrect. Manual inspection and literature search of the false positives suggests that at least one-half are in fact experimentally verified sequence-specific DNA-binding proteins. Many of the remaining putative false positives have little annotation, but any provided is supportive of the suggestion that these proteins are transcription factors. Therefore, a 5% error rate should be considered an upper bound.

Conversely, we calculated the coverage of our method to be between 60 and 67% by counting the number of proteins that GO annotates as a transcription factor but we fail to predict. This suggests that we miss around one-third of transcription factors. Closer inspection of these proteins showed that many are not actually sequence-specific DNA-binding TFs, but are involved in some other expression related process (e.g. basal transcription factors and chromatin proteins). This means that the false negative rate of one-third should be considered an upper bound. We expect to miss some TFs because we rely on HMM domain assignments which are known to give incomplete coverage [depending on the genome, between 30 and 60% of amino acids lack a domain assignment (19,20)]. Closer inspection of the 358 known TFs that we categorized as carrying out some other (non-expression related) function indicates that limitations in the homology detection are likely to be to blame; more than 60% of this set have no domain assignments at all.

At the same time, it must be noted that our predictions encompass as many transcription factors again that are unclassified in GO (37840 novel compared to 20246 known TFs) and these additional predictions are expected to be at least 95% accurate (Table 3). Therefore, despite the incomplete coverage, our method predicts many transcription factors that are unannotated.

Table 1. Sequence-specific DNA-binding transcription factor GO categories

Accession no.	Description
GO:0003700	Transcription factor activity
GO:0003702	RNA polymerase II transcription factor activity
GO:0003709	RNA polymerase III transcription factor activity
GO:0016563	Transcriptional activator activity
GO:0016564	Transcriptional repressor activity

These five categories are from the molecular function ontology and have been selected because they include sequence-specific DNA-binding transcription factors. We use these categories to evaluate our predictions.

Table 2. GO annotation of our predictions for PDB proteins

GO annotation	Number of proteins
Annotated as TF	393
Expression related	4
Other function	0
Unclassified	88

To evaluate the prediction method for proteins with known three-dimensional structures, we compared our results with the experimentally derived GO annotation of the PDB database. The first column of numbers indicates the GO annotation of proteins in our predicted TF set. 99% of predictions are corroborated by GO. (Annotated as TF based on experimentation rather than homology in GO.) The remaining 1% are classified as *nucleic acid binding*. This means they may be transcription factors, but there is insufficient functional annotation to make a sub-categorisation.

Table 3. UniProt benchmark

	Accuracy: Our TF list as annotated by GO	Coverage: GO annotated TFs as annotated by us
Full GO annotation		
Annotated as TF	20 246	20 246
Expression related	44	2414
Other function	2	3698
Unclassified	8774	987
Absent from GO or no domain assigned	29 066	2632
Experimentally derived GO annotation		
Annotated as TF	1010	1010
Expression related	27	88
Other function	22	168
Unclassified	1226	98
Absent from GO or no domain assigned	35 451	314

To evaluate the prediction method, we compared our predictions with the experimentally derived GO annotation of the UniProt database. The first column of numbers indicates the GO annotation of proteins in our predicted TF set. 99.8% of predictions are corroborated GO (annotated as TF), giving a false positive rate of 0.2% (Expression related and other function). Based on only experimentally derived annotation, 95% of predictions are validated by GO. The final column is our annotation of all the proteins GO annotates as transcription factors. Proteins with domains of unknown function are counted in the 'unclassified row'. This shows that we identify 67% of known transcription factors, or conversely, we miss about one third. Manual inspection suggests that some of the missed proteins may in fact be basal factors and therefore have been correctly excluded from our set. It should also be pointed out that we predict 37840 transcription factors that are not classified or of unknown function according to GO.

The final group of tests involved comparison with curated lists of transcription factors for individual genomes. First, we considered *S.cerevisiae*, using a list of 160 factors curated from literature by Luscombe *et al*. (24) as our standard. In total we predicted 169 transcription factors: 125 (74%) of these were known, 5 (3%) seemed to be false positives and 39 (23%) were novel, previously unannotated proteins. For the 160 known transcription factors, we correctly predicted 78% (125). We failed to predict the remaining 22% (35). Of these, one-half had no domain assignments and the remaining one-half had some domains assigned but no DNA-binding domain. These results confirm that our annotation provides good coverage (78%).

A second manual analysis of predicted transcription factors for the mouse genome identified an even higher proportion of proteins of unknown function. Shown in Table 4, our method identifies more than 600 currently unannoted proteins as being TFs. This corresponds to a 90% increase in the known mouse TF repertoire.

These examples illustrates the power of our method for identifying previously uncharacterized transcription factors. Almost one-quarter of our predictions were for uncharacterized proteins. Assuming the same false-positive rate as for the known factors, this means we have identified 590 new mouse transcription factors, increasing the size of the repertoire by more than 90%.

In summary, we have developed an automatic, broadly applicable method for predicting sequence-specific DNA-binding transcription factors. Based on an evaluation using a large set of annotated protein sequences, we find that it is accurate (95 to 99% correct) and has good coverage (between 60 and 78% identification rate). Most importantly, many previously unannotated transcription factors are reliably predicted.

Table 4. Functional annotation of predicted mouse transcription factors

Annotation for predicted TFs in mouse	Genes	% Mouse TF predictions
Known transcription factor	671	39
Transcription factor homolog	51	3
Previously known to contain a DBD	360	21
Other function	10	0.6
Unknown function	608	36

Annotation was taken from the MGD (27). Our prediction method identifies 608 genes of unknown function as transcription factors. This amounts to more than a 90% increase over the known factors.

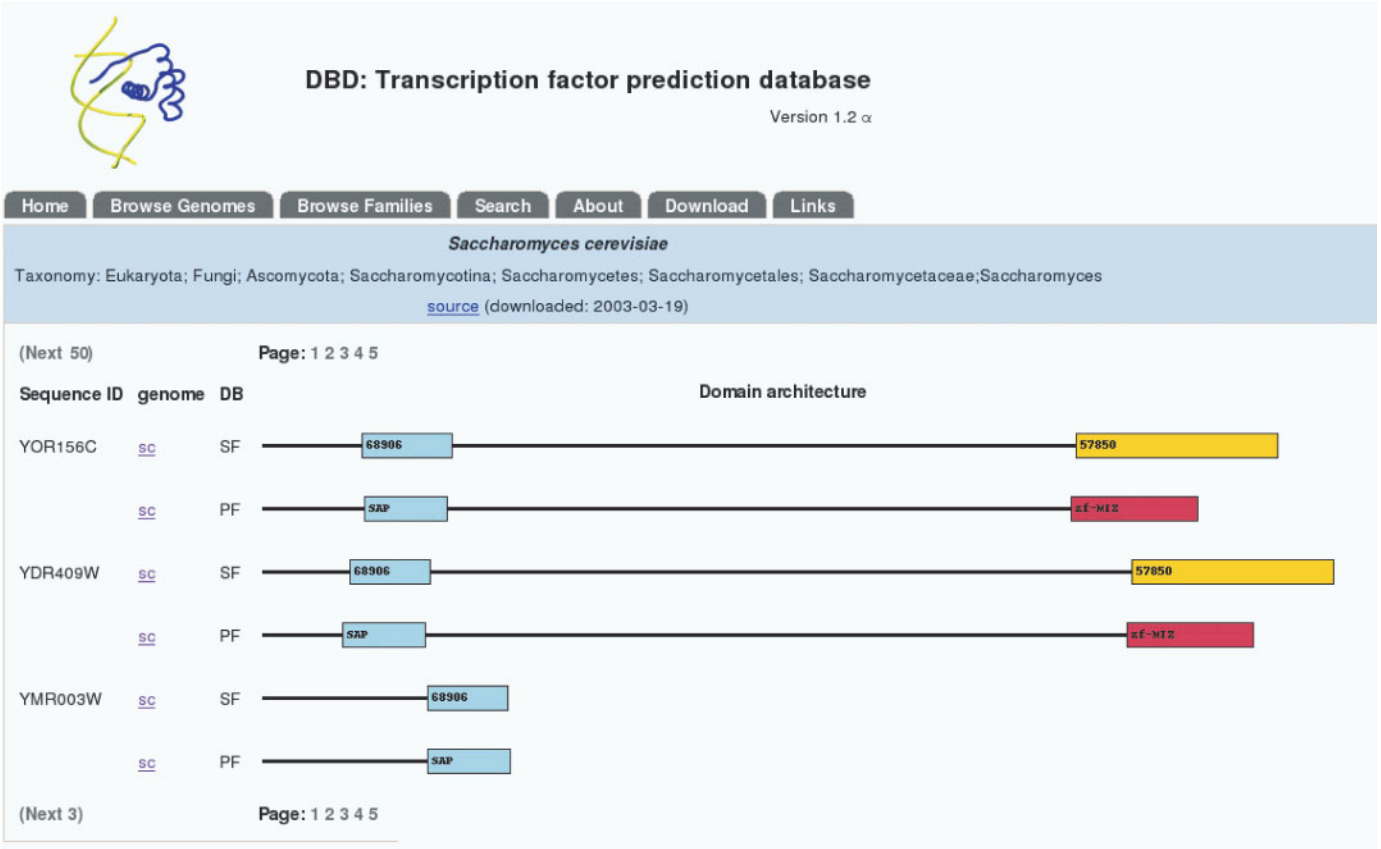

Figure 2. DBD: Yeast predictions screen-shot. Each predicted transcription factor is listed with two rows for the SUPERFAMILY and PFAM domain architectures. Domains are represented as rectangles, coloured according to their family and horizontally located based on their position in the amino acid sequence. Clicking on a domain takes the user directly to that family in the relevant domain database. Proteins are ordered based on their domain architecture. For ease of navigation (in particular for large genomes), the list of transcription factors is split into pages with 50 entries per page by default. Users can navigate between pages using previous/next or clicking on a page number.

Table 5. Transcription factor predictions in eukaryotes

Genome	Genome size	Transcription factors	% of proteins
Mus musculus 15.30	32 911	3240	9.8
Homo sapiens 15.33	32 035	3022	9.4
Drosophila melanogaster 3.1	18 484	936	5.0
Fusarium graminearum 1	11 640	453	3.9
Candida glabrata	5271	187	3.5
Candida albicans	6165	208	3.4
Kluyveromyces lactis	5331	176	3.3
Debaromyces hansenii	6869	230	3.3
Ashbya gossypii 1.0	4726	155	3.3
Saccharomyces cerevisiae	6356	201	3.2
Yarrowia lipolytica	6659	198	3.0
Kluyveromyces waltii	5214	155	3.0
Schizosaccharomyces pombe	5005	132	2.6

Genome size and number of predicted transcription factors are indicated for thirteen eukaryotes including 10 unicellular fungi and three multicellular animals. Splice variants are included in these counts. Organisms have been arranged according to the percentage of their proteins that are transcription factors. The unicellular fungi have been shaded in grey.

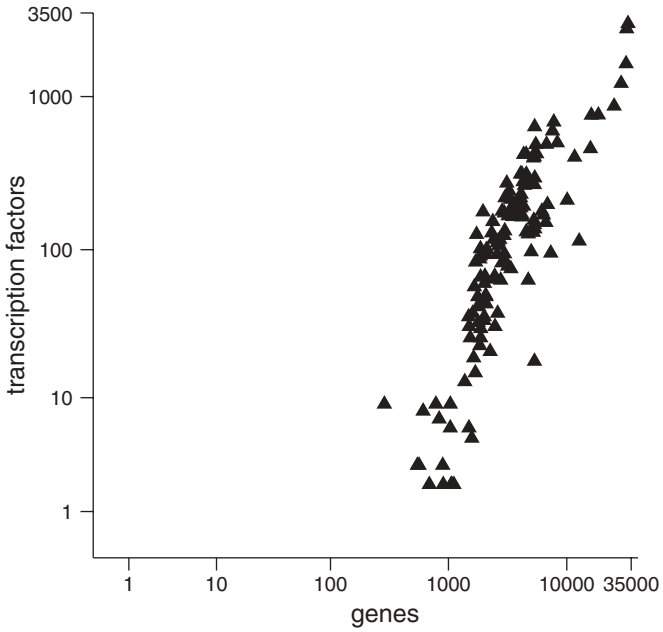

Figure 3. Number of genes in each of 151 genomes versus transcription factor predictions. The Number of genes (*x*-axis, log-scale) is plotted against the number of predicted transcription factors (*y*-axis, log-scale). Each splice variant is counted independently. (See database website for a list of genomes considered.)

THE DATABASE: DBD

The transcription factor prediction method described above is broadly applicable to any genome or sequence set. As we have shown, the results are both reliable and have good coverage. This means that by applying the method to complete genomes, it is possible to predict transcription factor repertoires for organisms. This type of information is invaluable to both bioinformaticians and biologists interested in gene regulation or expression.

In order to make our method accessible to the scientific community, we have developed an online database (www.transcriptionfactor.org) with pre-computed predictions for more than 150 completely sequenced genomes. Users can: browse predictions by genome or DNA-binding domain, search for particular sequence identifiers or domains and submit their own amino acid sequence for prediction. The web interface also allows users to download the domain assignments and list of DNA-binding domain HMMs as text files. The SUPERFAMILY and Pfam annotation as well as predictions for all genomes are available as text files.

Figure 2 shows part of the result page for *S.cerevisiae*. SUPERFAMILY and Pfam domain architectures are illustrated for each transcription factor, with colour indicating the domain family. Users can click on a domain to link directly to the relevant domain database.

Examples of the number of transcription factors we identify across eukaryotic genomes is shown in Table 5. The proportion of proteins that are transcription factors increases from fungi to insects to mammals. That is, between 2.6 and 3.9% of the unicellular eukaryotes' proteins are transcription factors compared to 5% for fly and almost 10% for mouse and human. Figure 3 shows the number of transcription factors in each genome compared to their total number of genes. This exponential increase in transcriptional regulatory proteins compared to genome size has been observed previously, based on GO functional categories, for bacteria (25) and genomes across all three kingdoms of life (26).

CONCLUSION

We have developed a broadly applicable method for automatically predicting sequence-specific DNA-binding transcription factors. The procedure uses HMMs from the SUPERFAMILY and Pfam databases to identify proteins that contain sequence-specific DNA-binding domains. A thorough evaluation showed that the method is both accurate (95 to 99% correct) and has good coverage (between 60 and 78% of known factors were identified). However, the most exciting feature of our method is that we also predict many novel, unannotated transcription factors. For example for mouse we find over 600 new factors amounting to more than a 90% increase in the TF repertoire.

We have applied our prediction method to more than 150 completely sequenced genomes from across the three kingdoms of life and implemented a web interface to make the data publicly accessible.

While the method described here represents a significant advance in the field of transcription factor prediction, adapting our system to use profile-profile methods (rather than profile-sequence) for remote homology detection is almost certain to increase sensitivity and coverage. Aside from generally improving domain assignments, profile-profile comparison could be used to make direct family level assignments for SCOP.

Until now, any researcher hoping to study transcriptional regulation would need to devise a list of putative factors for consideration. This database provides the first quantitatively evaluated transcription factor prediction set for all completely sequenced genomes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank M. Babu, N. Luscombe and S. Maslov for their suggestions and advice. S.K.K. is supported by an LMB Studentship, Overseas Research Student Award and University of Sydney Travelling scholarship. Funding to pay the Open Access publication charges for this article was provided by the MRC.

Conflict of interest statement. None declared.

REFERENCES

1. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Diaz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. and Collado-Vides,J. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.*, **32**, D303–D306.
2. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
3. Makita,Y., Nakao,M., Ogasawara,N. and Nakai,K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res.*, **32**, D75–D77.
4. Drysdale,R.A. and Crosby,M.A. (2005) FlyBase: genes and gene models. *Nucleic Acids Res.*, **33**, D390–D395.
5. Kanamori,M., Konno,H., Osato,N., Kawai,J., Hayashizaki,Y. and Suzuki,H. (2004) A genome-wide and nonredundant mouse transcription factor database. *Biochem. Biophys. Res. Commun.*, **322**, 787–793.
6. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T.Pr.M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Genome Res.*, **28**, 316–319.
7. Messina,D.N., Glasscock,J., Gish,W. and Lovett,M. (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res.*, **14**, 2041–2047.
8. Riechmann,J.L., Heard,J., Martin,G., Reuber,L., Jiang,C., Keddie,J., Adam,L., Pineda,O., Ratcliffe,O.J., Samaha,R.R. *et al.* (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.
9. Iyer,L.M., Koonin,E.V. and Aravind,L. (2002) Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors. *Genome Biol.*, **3**, RESEARCH0012.
10. Ravasi,T., Huber,T., Zavolan,M., Forrest,A., Gaasterland,T., Grimmond,S., Hume,D., Group,R.G. and Members,G. (2003) Systematic characterization of the zinc-finger-containing proteins in the mouse transcriptome. *Genome Res.*, **13**, 1430–42.
11. Madan Babu,M. and Teichmann,S.A. (2003) Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.*, **31**, 1234–1244.
12. Babu,M.M., Luscombe,N.M., Aravind,L., Gerstein,M. and Teichmann,S.A. (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, **14**, 283–291.
13. Hermoso,A., Aguilar,D., Aviles,F.X. and Querol,E. (2004) TrSDB: a proteome database of transcription factors. *Nucleic Acids Res.*, **32**, D171–D173.
14. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
15. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
16. Marinescu,V.D., Kohane,I.S. and Riva,A. (2005) The MAPPER database: a multi-genome catalog of putative transcription factor binding sites. *Nucleic Acids Res.*, **33**, D91–D97.
17. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
18. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.-B., Reynolds,D.B., Yoo,J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
19. Madera,M., Vogel,C., Kummerfeld,S.K., Chothia,C. and Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.*, **32**, D235–D239.
20. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
21. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
22. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
23. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
24. Luscombe,N.M., Babu,M.M., Yu,H., Snyder,M., Teichmann,S.A. and Gerstein,M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.
25. Ranea,J.A.G., Buchan,D.W.A., Thornton,J.M. and Orengo,C.A. (2004) Evolution of protein superfamilies and bacterial genome size. *J. Mol. Biol.*, **336**, 871–887.
26. van Nimwegen,E. (2003) Scaling laws in the functional content of genomes. *Trends Genet.*, **19**, 479–484.
27. Blake,J.A., Richardson,J.E., Bult,C.J., Cadin,J.A. and Eppig,J.T., and the members of the mouse Genome Database Group (2003) MGD: The Mouse Genome Database. *Nucleic Acids Res.*, **31**, 193–195.