D88–D92 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 11 December 2007*
doi:10.1093/nar/gkm964

DBD—taxonomically broad transcription factor predictions: new content and functionality

Derek Wilson[1,*], Varodom Charoensawan[1], Sarah K. Kummerfeld[2] and Sarah A. Teichmann[1]

[1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 0QH, UK and [2]Department of Developmental Biology, Stanford University Medical Center, 279 Campus Drive, Stanford, CA 94305-5329, USA

Received September 14, 2007; Revised October 16, 2007; Accepted October 17, 2007

ABSTRACT

DNA-binding domain (DBD) is a database of predicted sequence-specific DNA-binding transcription factors (TFs) for all publicly available proteomes. The proteomes have increased from 150 in the initial version of DBD to over 700 in the current version. All predicted TFs must contain a significant match to a hidden Markov model representing a sequence-specific DNA-binding domain family. Access to TF predictions is provided through http://transcriptionfactor.org, where new search options are now provided such as searching by gene names in model organisms, searching for all proteins in a particular DBD family and specific organism. We illustrate the application of this type of search facility by contrasting trends of DBD family occurrence throughout the tree of life, highlighting the clear partition between eukaryotic and prokaryotic DBD expansions. The website content has been expanded to include dedicated pages for each TF containing domain assignment details, gene names, links to external databases and links to TFs with similar domain arrangements. We compare the increase in number of predicted TFs with proteome size in eukaryotes and prokaryotes. Eukaryotes follow a slower rate of increase in TFs than prokaryotes, which could be due to the presence of splice variants or an increase in combinatorial control.

INTRODUCTION

Sequence-specific DNA-binding transcription factors (TFs) each recognize a family of *cis*-regulatory DNA sequences described by a consensus motif (1) or position-specific weight matrix (2). They regulate spatial and temporal gene expression by binding to DNA and either activating or repressing action of an RNA polymerase. Like other proteins, TFs are composed of evolutionary units called domains, which belong to families that can occur in many different proteins and various domain combinations. In the DBD database, we define TFs as proteins containing a sequence-specific DNA-binding domain (DBD). Other databases, such as TrSDB (3), or data sets, such as Messina *et al*. (4), include both specific and general TFs. The precise description of TFs as sequence-specific DNA-binding we use is useful in a wide variety of studies. Examples include: improving genome annotation; high-throughput experiments such as ChIP–chip, protein chip or yeast one-hybrid (5); and studies of the evolution of gene regulation comparing multiple genomes (6), or gene regulation networks (7). The DBD database has been used as an annotation tool in the context of the InterPro (8) and FlyTF (http://FlyTF.org) (9) databases.

Access to the DBD database is via http://transcription factor.org, where all data is available for viewing and immediate download. The community can browse predictions for over 700 species (from *Arabidopsis thaliana* to *Zymomonas mobilis*) or DBD family (including helix–turn–helix, zinc-fingers, homeobox and many others); search predictions by sequence identifier or domain family; receive classifications for submitted protein sequences, and download our domain assignments, as well as our manually curated list of DBDs.

The prediction method in the DBD database (10) uses hidden Markov models (HMMs) to identify domains in proteins from two databases: SUPERFAMILY (11) and Pfam (12). From DBD release 2.0 onwards, updated annotation resulted in 303 HMMs from SUPERFAMILY and 145 from Pfam compared to a total of 251 HMMs in the first version of DBD. The HMMs from SUPERFAMILY represent 37 superfamilies and 87 families according to the definitions in the SCOP database (13). This includes 98 new models representing 37 sequence-specific DBD families. This resulted in an increase in additional TF predictions of 4.7%, for the 150 organisms in the original version of DBD.

*To whom correspondence should be addressed. Tel: +44 (0)1223 402479; Fax: +44 (0)1223 213556; Email: dbd@mrc-lmb.cam.ac.uk

The pipeline used to predict TFs begins with a domain annotation of all proteins from completely sequenced genomes with all HMMs from the SUPERFAMILY and Pfam databases (Supplementary Figure 1). A protein is classified as a TF if it has a significant match to a model we annotated as being a DBD, with the significance thresholds for HMM matches taken from the Pfam and SUPERFAMILY databases. This results in an estimated 1–5% of false-positive annotations. The TF predictions are limited to the families in our annotated collection, which means that the coverage is about two-thirds of known TFs. At the same time, up to an additional 50% of proteins are predicted as TFs that have annotations such as 'hypothetical protein', particularly in metazoan genomes. For details of benchmarking, please refer to (10). The prediction method is general and applicable to any proteome or sequence set. In fact, the database has grown to encompass TF repertoires of over 700 publicly available genomes. Predictions for newly sequenced genomes are continuously added to the database.

The current DBD database contains information on over 200 000 predicted TFs. These TFs are distributed across the tree of life. It is not surprising that, we find a greater number of TFs in larger genomes. To investigate the relationship between TF abundance and proteome size in different lineages we graph these variables on a log–log plot as in Kummerfeld and Teichmann (10) (Supplementary Figure 2 in this paper). To illustrate the difference between the eukaryotic and prokaryotic superkingdoms we separately perform a model fitting for these lineages. From the linear relationship on the log–log scale a power law can be inferred. This power law could be due to the underlying distribution of DBDs. A small number of DBDs (such as helix–turn–helix and zinc-finger families) occur in the majority of TFs. Whereas most DBDs occur in only a small number of TFs. In agreement with van Nimwegen (14) and Ranea *et al.* (15), we find a higher proportion of TFs are required to regulate larger proteomes. We also find the TF abundance in archaea and bacteria expands more rapidly than in eukaryotes. Thus, in general, the same number of TFs regulate fewer prokaryotic genes than eukaryotic genes. The higher degree of combinatorial control, where gene expression is regulated by not just one but by a group of TFs, may also contribute to the lower eukaryotic TF requirements. Different combinations of TFs mean the number of gene regulation modes can increase with a reduced increase in TFs. Bacteria and archaea obey the same power law in terms of number of TFs and number of proteins. This is in accordance with their shared repertoire of DBD families, which we will return to below.

Apicomplexa appear not to follow either the prokaryote or typical eukaryote trends, perhaps because they are obligate parasites, and only survive in the nutrient-rich environment of their hosts. Thus, a different mode of gene regulation may be used by this lineage, or it is possible that their TFs are not well characterized by the current model libraries. Below, we will illustrate in more detail how the DBD database provides a consistent framework for comparison of the distribution of DBDs across the tree of life.

NOVEL DEVELOPMENTS

Researchers can use the DBD database in several ways. For instance, all TF predictions are available to download. However, most users are only interested in a small number of TFs, so we have expanded the website search options to allow retrieval of individual TFs and subsets of TFs. New search capabilities include: searching for gene names, for example lacI or P53; listing all TFs that contain either a specified DBD or non-DBD family, for instance all TFs containing the bZIP (leucine zipper) family; retrieving all TFs containing a specified DBD family, which occur in a particular organism, e.g. all homoeodomain-containing TFs in human (Figure 1a and b).

We illustrate the TFs containing a specified DBD family in a particular organism in Figure 1, where a hypothetical researcher is interested in the Homeobox TFs. These TFs are known to regulate vertebrate limb formation amongst other processes (16). Figure 1a depicts the search for TFs in *Homo sapiens* containing the homoeobox domain. A subset of the results of this search are shown in Figure 1b. By selecting the HOXA9 TF from this result set, the researcher can examine one of the new pages containing detailed information on each TF (Figure 1c). The detailed pages include the sequence of the TF, links to external databases containing further information on the protein, domain assignment regions and an indication of the quality of the domain assignment in the form of an Evalue. Links to predicted TFs with similar domain combinations are also provided on these pages. An example of predicted TFs with similar Pfam architectures to the HOXA9 TF (i.e. an N-terminal Hox9 activation region and a C-terminal Homoeobox domain) is shown in Figure 1d.

Using the data on DBD families in different organisms, we compare the occurrence of DBDs (from the Pfam project) across the tree of life. The heatmap in Figure 2 demonstrates the lineage-specific DBD expansions and contractions. The list of species and DBD lists are included in Supplementary Tables 1 and 2. We found the number of occurrences of each DBD in each organism, and then normalized this number by the proteome size of that organism. In order to represent both contractions and expansions, we calculated a *Z*-score for each of the normalized DBD occurrence values. The *Z*-score is calculated from the distribution of normalized DBD occurrence across genomes for a particular DBD family, and has a mean of zero and a standard deviation of one. It is negative when the normalized DBD occurrence is below the mean, and positive when above the mean. In Figure 2, DBD expansions (positive *Z*-scores) are represented using red, and contractions (negative *Z*-scores) using green.

Different sets of DBDs expand in different lineages. There is a clear separation between the DBD occurrence pattern in eukaryotes (in the top section of the heatmap) and prokaryotes. The DBD occurrence in prokaryotes is relatively diverse. For instance, there is a significant overlap between the DBD repertories of the actinobacteria, proteobacteria and firmicutes. This is almost certainly due to the ubiquitous horizontal gene transfer between prokaryotes. The DBD expansion pattern in archaea is

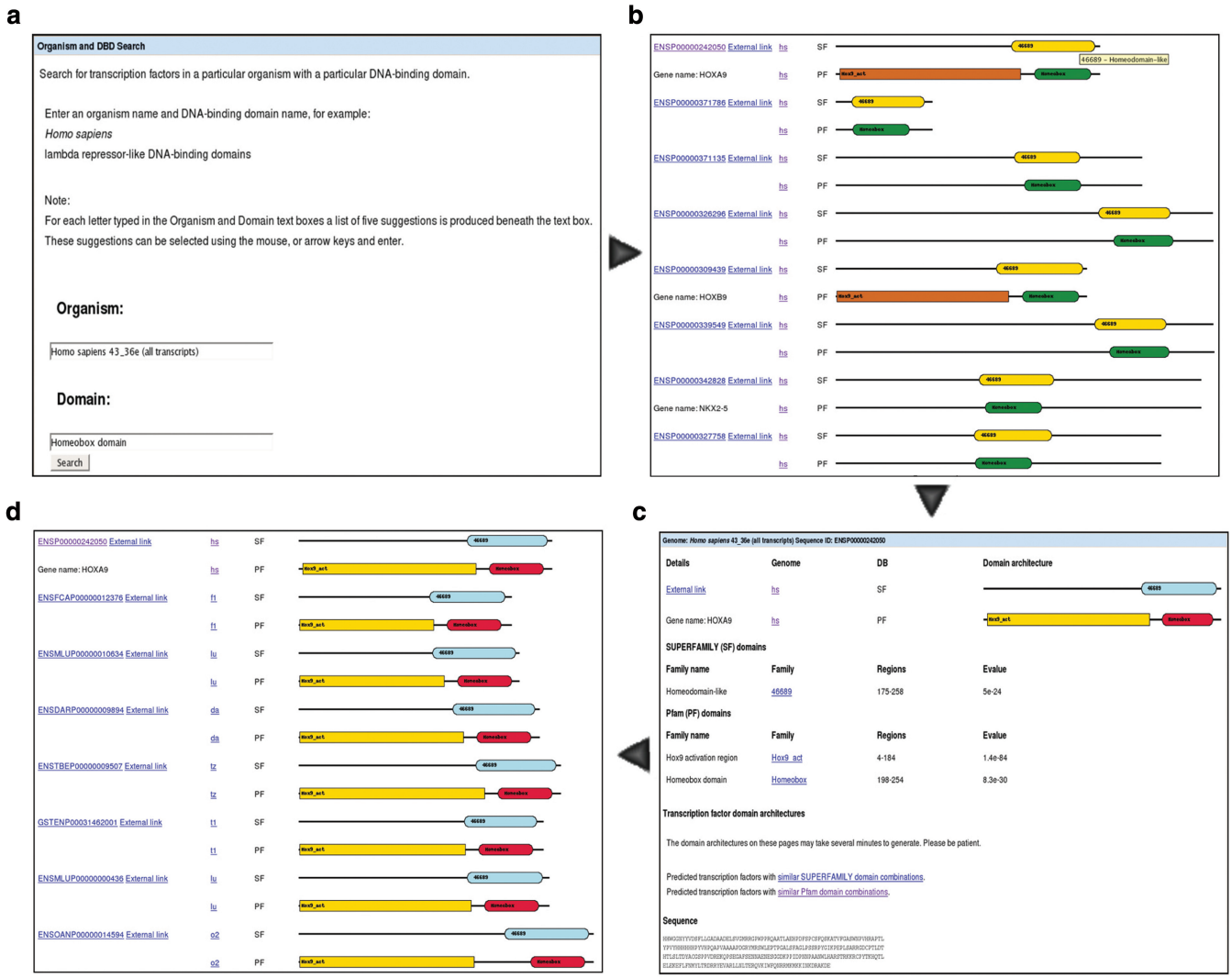

Figure 1. Examples of new search capabilities and content. **(a)** Search for TFs from a particular organism containing a specified DBD. The example used here is TFs from *Homo sapiens* containing the homoeobox domain. **(b)** The search in (a) results in TF predictions from *Homo sapiens* containing the homoeobox DNA-binding domain. **(c)** Selection of HOXA9 from (b) results in a web page with detailed information on this particular TF. **(d)** Clicking on the Pfam domain combination link in (c) retrieves the subset of TF predictions, which have the same two-domain arrangement as the HOXA9 transcription factor.

similar to that in bacteria, despite sharing conserved basal transcriptional machinery with eukaryotes rather than with bacteria. The majority of these prokaryotic DBDs have the helix–turn–helix as part of their structure (17).

The eukaryote-specific DBD expansions have considerably greater variety than the prokaryotic expansions. An increased DBD kingdom-specificity is found in the eukaryotes. The metazoan, fungal and plant kingdoms are clearly distinguishable (Figure 2a). The fungal and metazoan kingdoms share more DBDs than the plant and metazoan kingdoms, which reflects their closer phylogenetic relationship (18). The metazoa, in the top right section of Figure 2a, have the largest kingdom-specific DBD repertoire. This is most likely due to the regulatory overhead of metazoan complexity in terms of cell types.

The significant plant-specific DBD expansion is possibly due to the regulation of a large defence system—which plants have due to their inability to escape toxic environmental conditions. Figure 2b clarifies the nature of the DBD expansions in the viridiplantae lineage. The AP2 family is expanded throughout this lineage, but is believed to also occur in the apicomplexa (19). Figure 2c shows the AP2 domain in complex with DNA. This family is known to bind to the GCC-box pathogenesis-related promoter element (20) and activate defence genes. Several families are specifically expanded in the plant genomes of *A. thaliana*, *Medicago truncatula* and *Oryza sativa* (as opposed to the other viridiplantae, which are algae) including the family of ethylene insensitive 3 (EIN3) DBDs. This family regulates transcription in response to the chemically simplest plant hormone, ethylene (21).

FUTURE DIRECTIONS

Above we described novel developments in the display facilities and search tools, as well as the content of the

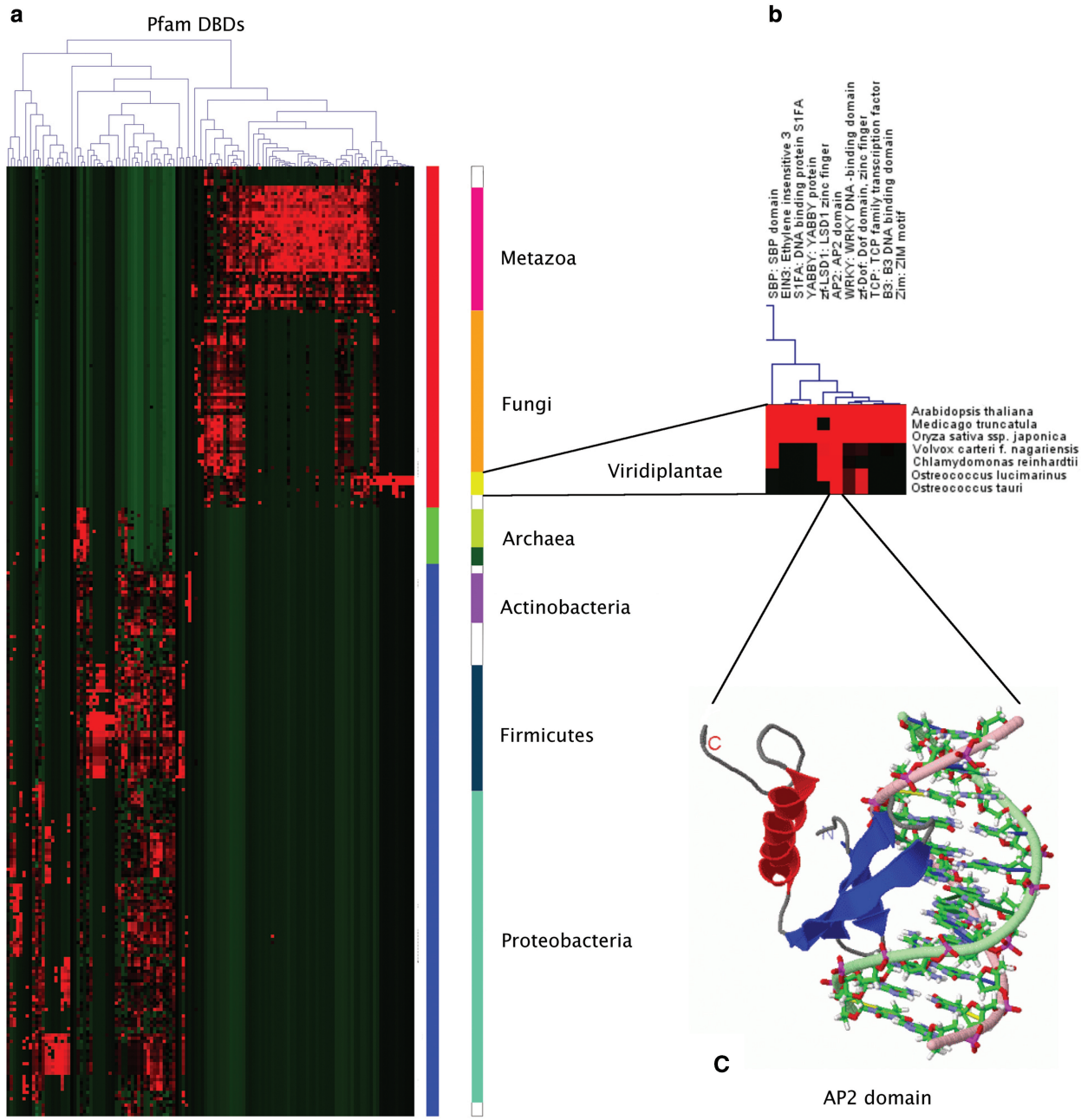

Figure 2. **(a)** Expansion and contraction patterns of DBD occurrence across the tree of life. Each column corresponds to a Pfam DBD. Each row of the heatmap represents a genome, ordered using the NCBI taxonomy. The vertical coloured bars indicate superkingdoms, kingdoms or phyla to which genomes belong. Eukaryotes are indicated using a red bar, archaea using a green bar and bacteria using a blue bar. Other kingdoms are represented using white bars. DNA-binding domain families are clustered using the average linkage method with Pearson correlation distance. Red squares represent an expansion of a DBD family, green squares represent a contraction of that family in a genome relative to other genomes. **(b)** A zoom on DBD expansions in the viridiplantae lineage. **(c)** Illustration of the three-dimensional structure of one of the DBDs specifically expanded in the viridiplantae kingdom, the AP2 domain in complex with DNA. The AP2 family transcription factors are known to be involved in plant pathogen defence response processes.

DBD database, with a few examples of the type of insight this provides. In the future, we will continue to update the HMM libraries, which will result in improvements to the TF prediction coverage. When updating the Pfam HMMs we will make use of, and incorporate, the Pfam clan information (12). We will also continue to add and update predictions for new proteomes. Exciting new eukaryotic proteomes we hope to add soon include higher eukaryotes such as orangutan, marmoset and wallaby, disease vector insects, additional nematodes and several plants.

We have eliminated several eukaryotic genomes (*Xenopus tropicalis*, *Apis melifera* and *Populus trichcarpa*) from our analysis of DBD occurrence due to the presence of uncharacteristically high numbers of bacterial DBDs.

This was a known problem in the *X. tropicalis* (frog) genome (22). The use of lineage-specific information on the occurrence of DBDs is a promising method for reducing false-positive TF classifications in the eukaryotes.

We also plan to refine the TF prediction procedure by taking into account that DBDs have typical patterns of domain repetition or combination with other DBDs or non-DBDs. It may be possible to make use of over-represented domain combinations to further improve our predictions, for instance by including marginal DBD matches if they occur in common TF domain arrangements as indicated by the statistical methods used in (23) and (24).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We gratefully acknowledge comments on the manuscript from Subhajyoti De and Siarhei Maslau. This work was funded by Medical Research Council; Royal Thai Government Scholarship to VC. Funding to pay the Open Access publication charges for this article was provided by Medical Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Robertson,G., Bilenky,M., Lin,K., He,A., Yuen,W., Dagpinar,M., Varhol,R., Teague,K., Griffith,O.L. *et al.* (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res.*, **34**, 68–73.
2. Bulyk,M.L. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol.*, **5**, 201–201.
3. Hermoso,A., Aguilar,D., Aviles,F.X. and Querol,E. (2004) TrSDB: a proteome database of transcription factors. *Nucleic Acids Res.*, **32**, 171–173.
4. 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.
5. Barrasa,M.I., Vaglio,P., Cavasino,F., Jacotot,L. and Walhout,A.J. (2007) EdgeDB: a transcription factor-DNA interaction database for the analysis of C. elegans differential gene expression. *BMC Genomics*, **8**, 21–21.
6. Drosophila Comparative Genome Sequencing and Analysis Consortium. (2007) Evolution of genes and genomes in the context of the drosophila phylogeny. *Nature*, In press, Nov.
7. Amoutzias,G.D., Robertson,D.L., Oliver,S.G. and Bornberg-Bauer,E. (2004) Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *EMBO Rep.*, **5**, 274–279.
8. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, 224–228.
9. Adryan,B. and Teichmann,S.A. (2006) FlyTF: a systematic review of site-specific transcription factors in the fruit fly drosophila melanogaster. *Bioinformatics*, **22**, 1532–1533.
10. Kummerfeld,S.K. and Teichmann,S.A. (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res.*, **34**, 74–81.
11. Wilson,D., Madera,M., Vogel,C., Chothia,C. and Gough,J. (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res.*, **35**, 308–313.
12. Finn,R.D., Mistry,J., Schuster-Böckler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, 247–251.
13. 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.
14. van Nimwegen,E. (2003) Scaling laws in the functional content of genomes. *Trends Genet.*, **19**, 479–484.
15. Ranea,J.A., Buchan,D.W., Thornton,J.M. and Orengo,C.A. (2004) Evolution of protein superfamilies and bacterial genome size. *J. Mol. Biol.*, **336**, 871–887.
16. Cohn,M.J., Patel,K., Krumlauf,R., Wilkinson,D.G., Clarke,J.D. and Tickle,C. (1997) Hox9 genes and vertebrate limb specification. *Nature*, **387**, 97–101.
17. Pérez-Rueda,E., Collado-Vides,J. and Segovia,L. (2004) Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. *Comput. Biol. Chem.*, **28**, 341–350.
18. Baldauf,S.L., Roger,A.J., Wenk-Siefert,I. and Doolittle,W.F. (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, **290**, 972–977.
19. Balaji,S., Babu,M.M., Iyer,L.M. and Aravind,L. (2005) Discovery of the principal specific transcription factors of apicomplexa and their implication for the evolution of the ap2-integrase DNA binding domains. *Nucleic Acids Res.*, **33**, 3994–4006.
20. Ohme-Takagi,M. and Shinshi,H. (1995) Ethylene-inducible DNA-binding proteins that interact with an ethylene-responsive element. *Plant Cell*, **7**, 173–182.
21. Chao,Q., Rothenberg,M., Solano,R., Roman,G., Terzaghi,W. and Ecker,J.R. (1997) Activation of the ethylene gas response pathway in arabidopsis by the nuclear protein ethylene-insensitive3 and related proteins. *Cell*, **89**, 1133–1144.
22. Yang,S., Doolittle,R.F. and Bourne,P.E. (2005) Phylogeny determined by protein domain content. *Proc. Natl Acad. Sci. USA*, **102**, 373–378.
23. Mott,R., Schultz,J., Bork,P. and Ponting,C.P. (2002) Predicting protein cellular localization using a domain projection method. *Genome Res.*, **12**, 1168–1174.
24. Coin,L., Bateman,A. and Durbin,R. (2003) Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proc. Natl Acad. Sci. USA*, **100**, 4516–4520.