

Systematic DNA-Binding Domain Classification of Transcription Factors

Philip Stegmaier¹

pst@biobase.de

Alexander E. Kel¹

ake@biobase.de

Edgar Wingender^{1,2}

e.wingender@med.uni-goettingen.de

¹ BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbüttel, Germany

² Department of Bioinformatics, Medical School, University of Göttingen, Goldschmidtstr. 1, D-37077 Göttingen, Germany

Abstract

Based on the manual annotation of transcription factors stored in the TRANSFAC database, we developed a library of hidden Markov models (HMM) to represent their DNA-binding domains and used it for a comprehensive classification. The models constructed were applied on the UniProt/Swiss-Prot database, leading to a systematic classification of further DNA-binding protein entries. The HMM library obtained can be used to classify any newly discovered transcription factor according to its DNA-binding domain and, thus, to generate hypotheses about its DNA-binding specificity.

Keywords: transcription factors, DNA-binding domains, HMM, classification, TRANSFAC database

1 Introduction

Transcription factors (TF) are proteins that control the first step of gene expression, the transcription of DNA into RNA sequences. Most of them do so by recognizing specific DNA-sequence features of so-called *cis*-regulatory elements in promoters, enhancers, and other regulatory regions in (eukaryotic) genomes. Each TF is thus part of the primary decoding machinery reading out the regulatory information that is laid down in the genomic nucleotide sequence and defines when, where and under which conditions a gene becomes active.

It is therefore of great importance to have information about the (usually relaxed) DNA-binding specificity of transcription factors, and to predict their binding sites and, thus, their target genes. About 10% of the genes in the human genome encode TFs (rough estimate from Swiss-Prot annotation and [19]), but the functionality of only about one third has been characterized (Swiss-Prot and TRANSFAC¹ annotation), and for only half of them we have some knowledge about their DNA-binding specificity. These figures look similar for the mouse system, and are much worse for other eukaryotes.

To get a first idea of the DNA-binding specificity of an as yet uncharacterized TF, we need to assign it to a systematic classification of DNA-binding domains (DBDs) that provide sequence-specific protein-DNA interactions. A first attempt for classifying TFs on the basis of their DNA-binding domains was published several years ago [21] and updated later on [13]. However, more members of the previously suggested classes and families were identified since then, many additional TFs were found which could not yet be assigned to these groups at all, new insights into the structural features of many DBDs required re-arrangements of the affected groups, and the increasing knowledge about the complexity of TF domain composition required to build up a DBD classification first before applying it onto a TF classification.

¹TRANSFAC is a registered trademark (®) of BIOBASE GmbH, Wolfenbüttel, Germany

In this contribution, we propose a comprehensive DBD classification scheme. The top-most levels are defined by structural considerations, defining four distinct superclasses with 31 classes, whereas a fine-classification was achieved by building up a library of hidden Markov models (HMM). The basis for this was provided by the manual domain annotation given by the TRANSFAC database, but was then extended by retrieving model-matching UniProt entries, refined and applied for a comprehensive classification of a maximal number of these UniProt entries. As a result, a system was obtained which can be used for the automatic annotation of newly discovered TF genes and the classification of the encoded TFs.

2 Databases and Methods

2.1 Databases

We used TRANSFAC Professional r7.2 [13], INTERPRO release 7.1 [1], InterProScan package 3.1, and UniProt databases Swiss-Prot release 42 and TrEMBL release 24 [5] and further updates. In addition to TRANSFAC annotation, structural models from PDB [4] as well as Pfam [3] and SMART [16] predictions were considered in the process of domain border definition.

2.2 Alignments and Phylogenetic Analyses

De novo alignments were computed with MAFFT [11] and CLUSTAL W [18], where MAFFT was the primary method of choice because of its superior accuracy and speed compared to CLUSTAL W.

Automatically derived multiple alignments of either kind commonly require manual improvement such as residue rearrangement or exclusion of sequences. This editing was conducted with SEAVIEW [10] and Jalview [7].

Relationships among domains were explored with two phylogenetic methods, the neighbor-joining implementation of CLUSTAL W [18] and Bête [17]. CLUSTAL W and Bête trees were further analyzed with ATV [22].

2.3 HMM Generation

We used the HMMER package version 2.3.1 [8, 23] for building profile-HMMs, computation of multiple alignments, assembly and maintenance of HMM libraries, sequence database searches with HMMs and HMM library searches with protein sequences.

Two types of hidden Markov models were developed. The first type, which we denote class-HMMs, is used to annotate structural domains corresponding to TRANSFAC classes. Models of the second type, subtype-HMMs, were created to specifically assign domains to lower hierarchical levels, families and subfamilies.

The workflow for the development of class-HMMs is shown in Figure 1. In the first step, sequences of factors and of their domains were extracted by their TRANSFAC class entry and by INTERPRO accession (Fig. 1, red gradient box). For the definition of domain boundaries, an extended alignment of domains was revised in the context of TRANSFAC and INTERPRO annotations as well as structural data, if possible.

In each round a new training set was derived (Fig. 1, yellow box) from which an HMM was built and used to detect new homologs in TRANSFAC and UniProt (blue gradient box). Additionally, one

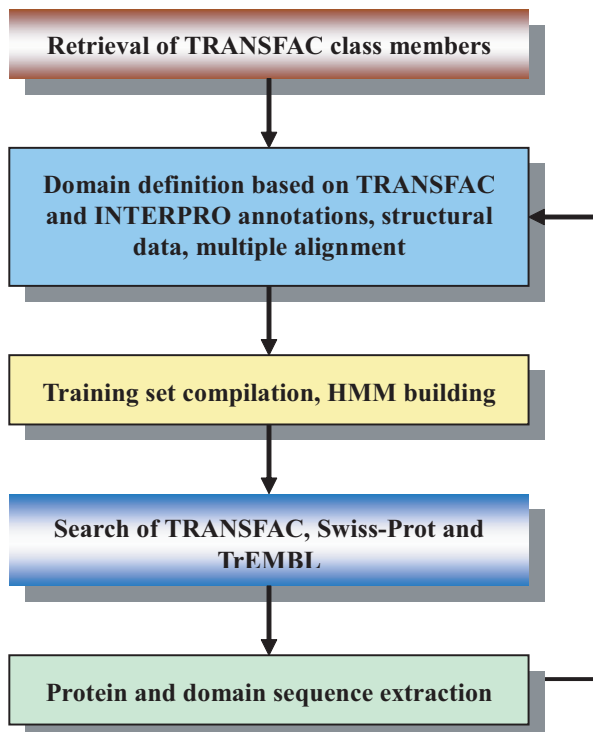


Figure 1: Workflow for the development of class-HMMs.

domain from subsets of the major alignment with eighty to 90% identity served as query in convergent PSI-BLAST searches with a conservative E-value threshold of maximally 0.001. New proteins from either HMM and PSI-BLAST searches were added to a library of TRANSFAC and UniProt homologs (green box). This library was used to compile a new large alignment with extensions, fragments were removed, domain definitions were reviewed and applied to a new model.

The procedure finally arrives at a stage where no more new homologs are found and a large alignment with trusted domains exists. The accepted HMM was optimized towards accurate reproduction of a maximal number of these domains. While this optimization primarily targeted the TRANSFAC set, a great coverage of UniProt sequences was usually achieved simultaneously.

Subtype-HMMs were constructed in a semi-automatic procedure. A program was developed that recursively seeks to represent a target family or subfamily by one model or by several models if necessary. In the first step a non-redundant training set is extracted from a trusted alignment based on the classification data. An HMM is built and tested against all domains of the class. If some domains are not correctly distinguished from those of other branches, more conservative training sets are retrieved from subsequent child levels, models are built and tested. The recursion continues until all domains of the target family or subfamily are covered. The output of the program was manually reviewed and modified if necessary. Finally, thresholds were set manually as well.

2.4 Domain Classification

The classification task was aided by specialized tools for data integration and representation. Classifications were developed manually in the context of the TRANSFAC classification as the prior guideline and phylogenetic data from CLUSTAL W and Bête, with the principal aim to identify functionally meaningful groups by their DNA-binding domain signature. Functional interpretation was primarily sought through TRANSFAC family and subfamily descriptions as well as factor feature annotations.

3 Results

3.1 Classification Principles

The primary goal in the development of a class-HMM was the construction of a model that is capable of detecting domains whose positions are consistent with the manually defined boundaries. Secondly, coverage of TRANSFAC domain representatives was sought with one or several models that meet the first requirement. Finally, coverage was extended over a maximal set of domains in public databases. Hence, class-HMMs juxtapose the specificity aspect of subtype-HMMs with the property to generalize domain definitions over a possibly large sequence space. The choice to assign the task of consistent domain annotation to class-HMMs is therefore derived from the constraint that subtype-HMMs are trained with TRANSFAC sequences exclusively, because only those were classified. It also supports the development of an automatic classification system that firstly identifies structural features and then proceeds with this information at lower hierarchy levels.

The resulting model or model library was applied to extract domains for classification and a multiple alignment was built for subsequent analyses. In contrast to the original TF classification, which was restricted to a maximally 6-level hierarchy including some optional levels, it seemed more appropriate to allow for a much deeper classification of individual DBDs. Generally, the resulting classification was a strictly tree-structured hierarchy, assigning a decimal classification number to each node: the first for the superclass, the second for the class, and the third for the family assignment. Different from the earlier TF classification [21], all further levels are denoted here “subfamilies”. Some of the levels, in particular the family level nodes, were occasionally skipped (assigned a “0” in the decimal classification number) when we had the feeling that the subnodes will, but could not yet be grouped together to higher-level taxa (“missing links”). In some branches of the hierarchy, the subclassification seems to reflect the phylogenetic relations of the underlying biological species since (roughly) vertebrate, insect, plant and fungal members nicely differentiate against each other.

However, in many other cases, they mix up even at very low hierarchical levels thus justifying such a comprehensive DBD classification based just on the protein features.

3.2 The Superclasses

On the topmost level, we have grouped all TF DNA-binding domains according to the structures which are known or can at last reasonably hypothesized by homology, into four superclasses: 1. basic domains, 2. zinc-coordinating domains, 3. helix-turn-helix domains, 4. beta-scaffold domains with minor groove contacts [13, 21, 24]. A fifth “superclass” (numbered “0”) comprises all those DBDs which may be grouped into families by sequence homologies, but for which no superclass-assignment can be done yet because of lack of structural information. Having revisited the classes within each superclass, the subdivision sometimes changed considerably compared with the old classification (Table 1). For this, however, it should be noticed that we did not yet revise the C2H2 and GATA zinc finger classes and obtained only preliminary results on the homeo domain classes. The changes introduced are due to (1) more recent information about the structure of some TFs; (2) significantly more members of the existing classes appeared, leading to new insights into family relationships; (3) new types of DBDs have been reported that make up new classes or families; (4) the fact that we attempted to establish a pure DBD classification instead of a DBD-based TF classification [21].

Table 1: Superclass sizes in the new TF DBD classification scheme.

No.	Superclass	New classification		Old classification	
		# of classes	# of families	# of classes	# of families
1	Basic domains	3	20	5	20
2	Zinc-coordinating domains	5 ^a	7	5	12
3	Helix-turn-helix domains	6	21	6	15
4	Beta-scaffold domains with minor groove contacts	13	24	12	24
0	Others	4	6	5	7

^a not yet including the C2H2 and GATA zinc finger classes and their families

3.3 Database

The data produced during the classification efforts were stored in a relational database. They fall into three categories: (1) domains annotated in TRANSFAC, (2) their classification in the form of hierarchical relationships and (3) profile-HMMs representing groups at different levels of the hierarchy.

At its current status the database contains 2367 domains in 30 classes. 540 HMMs were constructed for class and subtype representation providing reasonable profile coverage of the classified items.

3.4 Superclass 1: Basic Domains

These are DBDs which are characterized by a large excess of positive charges, preventing them from being structured when free in solution, but becoming α -helically folded when interacting with DNA (e.g., [20]). Usually, they appear in tight connection with a dimerization domain, a leucine zipper (ZIP), a helix-loop-helix (HLH) or a helix-span-helix (HSH) domain. Since dimerization of these factors is a prerequisite for their DNA-binding and largely contributes to their DNA-binding specificity, these two regions have to be analyzed together.

The first class of this superclass is inhabited by “basic region + leucine zipper” motifs (bZIP). Because of the heterogeneity of manual data concerning the definition of basic region starts, a consensus boundary was chosen for basic regions based on a multiple alignment of all TRANSFAC bZIP proteins analyzed. Leucine-zipper C-termini were adapted from TRANSFAC and also manually reviewed. While leucine zippers are elaborately maintained in the TRANSFAC feature source, they are covered through one-size-fits-all models in InterPro leading to great discrepancies between predicted and manually curated leucine zipper C-termini due to this simplification. Therefore, automatically generated data about leucine zippers cannot be trusted, since these regions are highly variable in length and lack a terminal signal. The construction of HMMs for the bZIP class underlines the priority of predictive

precision over coverage and generalization. Unlike other classes, classified bZIP motifs are represented by a large number of HMMs (88 class-, 43 (sub-)family-HMMs; Table 2), each trained with a conservative set of sequences typically derived from a single subfamily. Thresholds were set for the spectrum of precisely reproduced bZIP regions for each HMM individually. Hence, the resulting bZIP HMM library generalizes poorly compared to other models and is likely to yield many false negatives, but is capable of accurately recovering the whole set of bZIP regions defined during this work. Because of the problems of finding a comprehensive and accurate HMM for the leucine zipper region, we omitted the “ZIP-only” family (i.e., dimerization region without the DNA-contacting basic region) from the classification.

In contrast to bZIP regions, bHLH and bHSH regions were found to have well conserved N-terminal and C-terminal boundaries. For both of them, only a few class-HMMs were sufficient to generally describe these domains, where HLH domains of inhibitory factors without a basic region are distinguished from bHLH domains by their own model (Table 2). Applying these models, we found that changes had to be introduced by combining the previous classes 1.2 (“basic region + helix-loop-helix”, bHLH) and 1.3 (“basic region + helix-loop-helix + leucine zipper”, bHLH-ZIP), as well as by combining the class of NF-1-like factors (TRANSFAC class 1.4, until release 7.4) with that of the SMAD factors (4.12). The previous modification was done because of the above-mentioned problem of leucine zipper prediction. The latter modification was to be introduced because of the structural homology between SMAD and NF-1 DNA-binding domains that was recently revealed [15]. - The last class contains “basic regions + helix-span-helix” domains (bHSH) (now 1.3, previously 1.6), still populated by only one family (AP-2 factors).

3.5 Superclass 2: Zinc-Coordinating Domains

Five classes currently populate the superclass of zinc coordinating domains. Three are newly introduced to the hierarchy. Classification and HMM data of 1518 C2H2 domains from 314 factors and of 86 GATA domains from 54 factors are not yet inserted into the database, so that these are not considered to be covered by this work. Additionally, the TRANSFAC scheme contains the small class “Zinc fingers of alternating composition” (2.5) which was not adopted as well, because its members are likely to be discussed in separate classes in future. For these factors, in particular the C2H2 zinc finger domains, we may also refer to the classifications reported for certain phylogenetic groups [6, 9].

The class of Cys4 zinc fingers of nuclear receptor type contains 233 domains from the same number of factors in two families and 195 subfamilies (Table 2). One HMM is linked to the class and 51 HMMs were constructed for family and subfamily assignment. The tree was extensively reconstructed to capture higher order relationships between nuclear receptor subfamilies based on their DNA-binding domain. The resulting classification model is in strong compliance with the clusters of the nuclear receptor nomenclature [14] below the family level, whereas the two families reflect the previous TRANSFAC hierarchy with the exception that the estrogen receptor(-related) DNA-binding domains have been placed now with thyroid hormone receptor-like domains (2.1.2.13). Before, it was classified together with the other steroid hormone receptors more because of the chemical and functional similarity of the ligands rather than the DNA-binding domains, and its placement near the thyroid hormone receptor-like domains now also reflects much better its binding preference for a TGACCT rather than a TGTYCT core sequence.

The remaining three classes of zinc-coordinating domains, DM, GCM and WRKY, are newly introduced to the hierarchy. DM and GCM factors appear to be restricted to animals. Both types are involved in developmental processes such as sex differentiation (DM) and gliogenesis (GCM). In contrast, WRKY factors seem to be abundant among plant organisms only, where they participate in a number of physiological programs such as senescence, pathogen defense or secondary metabolite biosynthesis. While GCM factors carry a single DNA-binding unit, DM and WRKY domains occur with one or two copies per polypeptide.

Table 2: Subclassification of the TF superclasses.

Class	Members	Factors	Families*	Subfamilies*	HMMs	
					Class	(Sub-) Family
bZIP, 1.1	266	266	7	267	88	43
bHLH, 1.2	302	302	12	193	2	30
bHSH, 1.3	12	12	1	7	1	0
Nuclear receptors, 2.1	233	233	2	195	1	51
C6 zinc clusters, 2.3	50	50	1	88	1	9
DM, 2.4	2	2	1	2	1	0
GCM, 2.5	4	4	1	7	1	0
WRKY, 2.6	35	24	2	54	1	7
Homeo box, 3.1	1031	1007	10	≥ 800	12	≥ 60
Paired box, 3.2	90	90	1	79	1	9
Forkhead/winged helix, 3.3	134	134	2	161	2	23
HSF, 3.4	27	27	3	35	1	7
Tryptophan clusters, 3.5	451	302	3	528	14	84
TEA domain, 3.6	9	9	2	9	1	0
RHR, 4.1	28	28	2	35	1	11
STAT, 4.2	17	17	3	19	1	11
p53-like, 4.3	5	5	1	6	1	0
MADS, 4.4	274	274	3	241	1	59
β -Barrel α -helix domains, 4.5	1	1	1	2	1	0
TBPs, 4.6	22	11	2	17	1	2
HMG, 4.7	119	88	5	110	1	23
Histone fold, 4.8	30	30	1	42	1	3
Grainyhead, 4.9	5	5	1	6	1	0
Cold-shock domain, 4.10	13	13	1	5	1	0
Runt-like domain, 4.11	26	26	1	19	1	0
SMAD/NF-1, 4.12	100	100	2	81	1	14
T-Box domain, 4.13	43	43	1	46	1	10
Copper fist, 0.1	4	4	1	5	1	0
Pocket domain, 0.2	7	7	1	12	1	0
AP2/EREBP-related, 0.3	55	47	3	74	1	12
SAND domain, 0.4	3	3	1	3	1	0

*including missing link nodes. In each row a class is identified by its name and its hierarchy number. For each class the number of classified domains and the number of source factors is given in the Members and Factors columns, respectively. The numbers of HMMs representing a class are shown in the class-HMMs column. Numbers of family and subfamily nodes are compiled in the Family and Subfamily columns, whereas the size of the HMM libraries representing a class at these levels is given in the (sub-)family-HMMs column. The figures for class 3.1 (homeobox factors) are preliminary estimates.

3.6 Superclass 3: Helix-Turn-Helix Domains

At the current database status, there are five classes in the helix-turn-helix superclass which are arranged in accordance with the previous TRANSFAC classification. The classification of 1031 Homeobox domains from 1007 factors and data of the 12 HMMs covering these are not annotated in the relational database, yet. First analyses, however, have revealed that the resulting family structure may correspond well to the classification reported on the Homeodomain page of T. Bürglin [24], except that the Prd and the LIM family are much more heterogeneous than the prd and the LIM class defined there, and that at least four additional families had to be defined. Results of further analyses conducted on this superclass are summarized in Table 2.

The set of Paired domains contains 90 sequences with a low degree of divergence between most of them. Families in the existing TF classification distinguish proteins with an additional homeo domain (3.2.1.) from those which carry only a Paired domain (3.2.2.). However, this distinction is not expressed at the level of Paired domain sequence similarity, so that there is significant overlap between both groups in the pure domain hierarchy constructed here. Moreover, Paired domains are currently not classified into families since it is unclear which features determine the relationship between 89 out of the 90 sequences, showing very little divergence, and NPX1 from *Caenorhabditis elegans* which is much more remotely related. Therefore, the family assignment is left open by the current solution and Paired domains are just grouped into subfamilies where the type of relationship between members of the set is obvious. As in other cases, a pure DBD classification may differ from the classification of TFs. Having classified the individual domains first separately, TFs comprising several of them in distinct combinations may have to be classified based on these specific combinations. This will apply also to C2H2 zinc finger as well as to Myb-domain factors (see below for the latter).

The Forkhead/winged helix class (3.3) and the class of Tryptophan clusters (3.5) are treated in a special way. Although DNA-binding domains of their families, Forkhead (3.3.1) and RFX (3.3.2) in the first case, Myb (3.5.1), ETS (3.5.2), and IRF (3.5.3) in the latter, are structurally related, they are not robustly accessible with HMMs, which is primarily due to different domain boundaries. Each family is treated separately in domain definition and prediction processes so that both classes, Forkhead/winged helix as well as Tryptophan clusters, are covered by different family-specific HMMs. This treatment is also supported by the observation that unusual interference between families of either class cannot be found during model construction. In the case of Tryptophan cluster domains, which are characterized mainly by a regular pattern of mostly three tryptophan residues in a spacing of 12-21 residues, the large number of HMMs constructed for class and (sub-)family nodes correlates with the great sequence divergence among Myb domains and the large number of Myb domain subtypes, since Myb factors usually harbor several such domains. Then, mammalian c-Myb and plant MYBA2 exhibit 3 domains of subfamilies (3.5.1.1) (3.5.1.1) (3.5.1.2), whereas others such as plant Myb-1, MYBAS1 or JAMyB have a (3.5.1.1) (3.5.1.2) and plant CDC5 has a (3.5.1.1) (3.5.1.1) composition. Such a comprehensive picture of differential domain assemblies is likely to accelerate a whole protein-based classification since putative functional factor families can be selected by their domain subtype composition.

3.7 Superclass 4: β -Scaffold Domains with Minor Groove Contacts

It is (still) very difficult to find any characteristic which is common to all TF, or DBDs, respectively, in this superclass. Any pair or subgroup may share some characteristics, but even the feature “ β -scaffold” is not shared by all of them. For instance, the DBD HMG-type comprises just α -helical segments, but their overall-mode of DNA-interaction resembles very much that of others in this superclass (such as TBP) by inserting into the minor groove and causing a sharp kink in the DNA double helix [12]. Many of the classes in here comprise only a single family.

T-Box domains are introduced to the transcription factor hierarchy as a new class with ten subfamilies. The family level is undefined, because higher order relationships are expected, but no clear decision could be deduced from the available information. The order of subfamilies, however, reflects

the consensus of phylogenetic methods and manual inspection, so that putative relationships are implied. T-Box domains are assigned to “ β -scaffold factors with minor groove contacts” because of their relationship to p53, STAT (4.2), RHR (4.1) and Runt DNA-binding motifs (4.11) proposed by structural classifications of CATH [25] and SCOP [25].

There are two classes whose domains may occur as multiple copies per factor. One, two, five or six HMG boxes are observed in HMG factors analyzed here, where sets of five or six domains are restricted to UBF factors. UBF motifs are positioned in one family containing all UBF segments (4.7.4), although they are not connected by homology exclusive to other domains of the class. Yet, a significant relationship to any other HMG box type is not indicated either, so that at this point the decision was made to form a family of UBF domains in order to increase the simplicity of the tree. Besides minor differences, family definitions for HMG boxes are in compliance with the previous TRANSFAC hierarchy. Secondly, the DNA-binding region of TBP comprises two strongly conserved repeats which contact DNA individually. As both segments are well distinct from each other, they are classified in two families (4.6.1., 4.6.2.).

3.8 Automatic Domain Annotation and Classification

The relational data model, annotated domains, their classification and HMMs as representatives of class, family and sub-family nodes provide the necessary resources for automatic domain annotation and assignment to known groups. This requires a procedure that concertes the available information in order to yield reasonable results for an unknown spectrum of sequences. The workflow of the program is shown in Figure 2. Classification of a query sequence proceeds through four stages. Domain and class prediction (red frame, 1), subtype classification by HMMs (blue frame, 2) and sequence based classification by BLAST (green frame, 3) finally yield a decision at which point of the hierarchy a match belongs (dark red frame, 4).

When we applied this scheme to all 6664 eukaryotic Swiss-Prot entries with the keywords “DNA-binding” OR “transcription regulation”, we could classify 2045 of these Swiss-Prot entries with 2214 features, since some of them have multiple DBD motifs. In some cases, such as the p53-related proteins or the WRKY domains, we could retrieve 7- or 2.5-fold as many proteins from Swiss-Prot as are already functionally annotated in the TRANSFAC database, whereas for others, such as the Tryptophan-cluster or the MADS box factors, TRANSFAC has 2-4-fold more annotated and classifiable factors than Swiss-Prot. The remaining 4619 Swiss-Prot entries that could not be classified by our HMM library are either false-positives of the Swiss-Prot query (e.g., representing unspecific DNA-binders like most HMG-proteins or histones), or non-DNA-binding transcription regulating proteins such as coactivators. Entries annotated in Swiss-Prot as transcription regulators make up 2588 of the 4619 unclassified sequences. Furthermore, homeo domains, C2H2

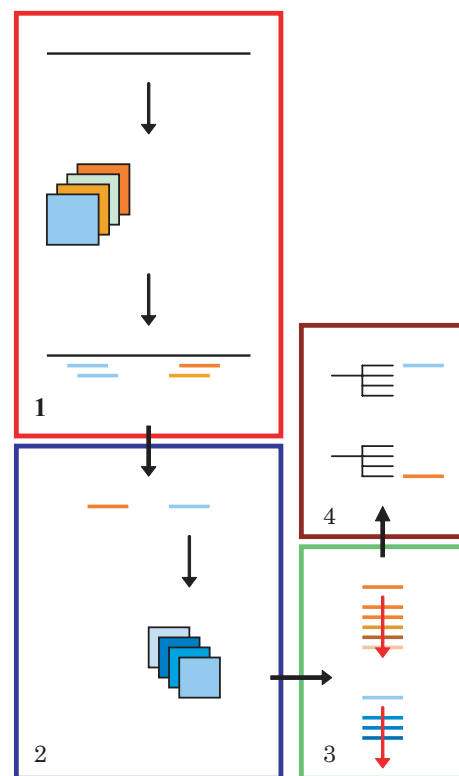


Figure 2: Workflow for automatic classification. In the first step, a general domain and class prediction is performed for the protein sequence to be classified (red frame, 1; black line - query sequence, colored squares - class-HMMs, colored lines - HMM matches). In the second step, the identified DBDs are sub-classified using the HMM library developed (blue frame, 2; colored lines - domain sequences, colored squares - subtype-HMMs). Third, a sequence-based classification is achieved by BLAST-ing query domains against classified TRANSFAC domains (green frame, 3; colored lines under red arrows - classified domains), which leads to positioning of query domains within in the hierarchy (dark red frame, 4).

and GATA zinc fingers, altogether present in 1125 unclassified sequences, were not assessed in this screening as well as other DNA-binding domain classes not yet described in our database. In contrast, HMM searches for domains which are contained in our classification reached false negative rates of 0% (74/74) for MADS box domains and roughly 1% for bHLH (3/310) as well as for nuclear receptor type zinc fingers (4/353). A comparably high value of 9% was obtained for bZIP domains (14/164) which is most likely due to the special conception of the bZIP-specific HMMs (see above). Estimating the false positive rate of the classification is difficult because of incomplete knowledge about the proteins under study. Among the 2045 classified entries, 303 were not annotated in Swiss-Prot as transcription regulators, just as DNA-binders. Out of them, 301 were not assigned to individual transcription factor families by the classification procedure. The remaining two sequences, MYB_DROME and MYBH_DICDI, exhibit an arrangement of Myb domain subtypes typically found in c-Myb transcription factors (see above). Therefore, the majority of the 303 entries presumably either belongs to not transcription-specific DNA-binding domain classes and, thus, were correctly not assigned to transcription factor families, or the involvement in transcription regulation is still to be investigated in vivo or in vitro. Comparable results were obtained for a set of 9600 CATH r2.5.1 domains representing 95% identical sequences. Automatic classification yielded 121 matches to all classes except for bHSH (1.3), WRKY (2.5), TEA (3.6) and Grainyhead (4.9) for which no structural model exists. To our knowledge, these matches cover nearly all corresponding domains present in the test set.

4 Discussion

The system developed here is an initial step towards a comprehensive classification of transcription factor DNA-binding domains, complementing the previous TRANSFAC classification of TFs and providing means for its revision in a couple of instances. As stated before, proteins are commonly composed of multiple domains required, for instance, for DNA-binding, protein interaction or catalytic activities. Their full functional description should therefore be aware of the configuration of domains present in a polypeptide. Hence, this database is assumed to be an important facility in order to arrive at a functional description of transcription factor gene products. Relationships that cannot be treated appropriately in a classification of whole proteins, for example intricate assemblies of multiple DNA-binding units, are possibly accessible at the domain level.

Up to now, class-HMMs as well as classifications were constructed manually which is not only a time-consuming process, but also poses a source of error and discontinuity in spite of the stereotypical paradigm of conserved residue patterns. Manual curation depends on observing such patterns in the context of several samples and in contrast to other clusters in order to make a decision over their significance.

Difficulty in the estimation of significance also affects consecutive hierarchy levels. Some groups may be “overclassified”, which is reflected by very deep branches. Since this is also due to the lack of information whether an observed pattern is functionally meaningful, an appropriate algorithmic treatment appears to be complicated. However, development of classification standards and more sophisticated software applications for the support of manual curation and model construction may constitute achievable devices for improvement on these issues, whereas extended automatization of class-HMM development is possibly more easily accessible than automatic subtype classification.

The classification of transcription factor DNA-binding domains described in this work has been developed to enable automatic retrieval, annotation and classification of TFs for which related domains are already known. In those cases where established classifications have already been published, our hierarchy is usually in good agreement as was specified above for the nuclear receptor class [14] and the homeobox factors [24], but is also true for our bHLH classification in spite of some difference in detail [2].

Up to now, a systematic application on the whole UniProt database has been done in depth for the bHLH class. The attempt to classify 1858 UniProt sequences which were identified by InterPro

as (b)HLH proteins led to the assignment of 1742 (94%) bHLH and HLH-only domains to the nodes in the classification scheme proving that at least for this class, our models provide a satisfying coverage. Out of the non-assignable proteins, at least some are not transcription factors (e.g. Q8F5I5, a triosephosphate isomerase, or Q82SU6, a Glycerol-3-phosphate dehydrogenase). Thus, our system may even help to disclose some erroneous annotations, or at least discriminate well between (b)HLH domains of transcription factors and others. A more detailed analysis of the classification of all Swiss-Prot entries representing DNA-binding or transcription regulating proteins is still to be done, and an extension to all UniProt polypeptides is on the agenda for the near future.

After completion of this work for homeo domains and C2H2 zinc finger motifs, we will initiate a systematic classification of the DNA-binding profiles of the individual DBD groups, as documented by the TRANSFAC database. Some first attempts made us confident that it will be possible to correlate the features that are reflected in the hierarchical DBD classification with pronounced features of the cognate DNA-binding sites. For instance, the mammalian factors Mlx/MondoA and the fungal factors PHO4/NUC-1, though exhibiting very remote sequence similarity, have been classified into one family by our system. It turned out that this fits nicely with the reported consensus sequences of CACGTG and CACGT(G/T) for the Mlx/MondoA complex and for yeast PHO4, respectively. Systematic extension of these studies would be an interesting step towards deciphering the DNA-protein recognition code.

References

- [1] Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D.R., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J.A., and Zdobnov, E.M., InterPro - an integrated documentation resource for protein families, domains and functional sites, *Bioinformatics*, 16(12):1145–1150, 2000.
- [2] Atchley, W.R. and Fitch, W.M., A natural classification of the basic helix-loop-helix class of transcription factors, *Proc. Natl. Acad. Sci. USA*, 94(10):5172–5176, 1997.
- [3] Bateman, A., Birney, E., Cerruti, L., Durbin, R., Eddy, S.R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E.L.L., The Pfam protein families database, *Nucleic Acids Res.*, 30(1):276–280, 2002.
- [4] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P., The protein data bank, *Nucleic Acids Res.*, 28(1):235–242, 2000.
- [5] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M., The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res.*, 31(1):365–370, 2003.
- [6] Böhm, S., Frishman, D., and Mewes, H.W., Variations of the C2H2 zinc finger motif in the yeast genome and classification of yeast zinc finger proteins, *Nucleic Acids Res.*, 25(12):2464–2469, 1997.
- [7] Clamp, M., Cuff, J., Searle, S.M., and Barton, G.J., The Jalview Java alignment editor, *Bioinformatics*, 20(3):426–427, 2004.
- [8] Eddy, S.R., Profile hidden Markov models, *Bioinformatics*, 14(9):755–763, 1998.

- [9] Englbrecht, C.C., Schoof, H., and Böhm, S., Conservation, diversification and expansion of C2H2 zinc finger proteins in the Arabidopsis thaliana genome, *BMC Genomics*, 5(1):39, 2004.
- [10] Galtier, N., Gouy, M., and Gautier, C., SEAVIEW and PHYLO_WIN : Two graphic tools for sequence alignment and molecular phylogeny, *Comput. Appl. Biosci.*, 12(6):543–548, 1996
- [11] Katoh, K., Misawa, K., Kuma, K., and Miyata, T., MAFFT : A novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Res.*, 30(14):3059–3066, 2002.
- [12] Lebrun, A. and Lavery, R., Modeling DNA deformations induced by minor groove binding proteins, *Biopolymers*, 49(5):341–53, 1999.
- [13] Matys, V., Fricke, E., Geiffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.-U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E., TRANSFAC: Transcriptional regulation, from patterns to profiles, *Nucleic Acids Res.*, 31(1):374–378, 2003.
- [14] Nuclear Receptors Committee, A unified nomenclature system for the nuclear receptor subfamily, *Cell*, 97(2):161–163, 1999.
- [15] Stefancsik, R. and Sarkar, S., Relationship between the DNA binding domains of SMAD and NFI/CTF transcription factors defines a new superfamily of genes, *DNA Sequence*, 14(4):233–239, 2003.
- [16] Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P., SMART, a simple modular architecture research tool: Identification of signalling domains, *Proc. Natl. Acad. Sci. USA*, 95(11):5857–5864, 1998.
- [17] Sjölander, K., Bayesian evolutionary tree estimation, *Proc. 11th International Conference on Mathematical and Computer Modelling and Scientific Computing, Computational Biology Session “Computing in the Genome Era”*, Washington, D.C., 1997
- [18] Thompson, J.D., Higgins, D.G., and Gibson, T.J., CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, 22(22):4673–4680, 1994.
- [19] Waterston, R.H., *et al.*, Mouse genome sequencing consortium, initial sequencing and comparative analysis of the mouse genome, *Nature*, 420(6915):520–62, 2002.
- [20] Weiss, M.A., Ellenberger, T., Wobbe, C. R., Lee, J.P., Harrison, S.C., and Struhl, K., Folding, transition in the DNA-binding domain of GCN4 on specific binding to DNA, *Nature*, 347(6293):575–578, 1990.
- [21] Wingender, E., Classification scheme of eukaryotic transcription factors, *Mol. Biol. (Mosk.)*, 31(4):483–497, 1997.
- [22] Zmasek, C.M. and Eddy, S.R., ATV: Display and manipulation of annotated phylogenetic trees, *Bioinformatics*, 17(4):383–384, 2001.
- [23] <http://hmmer.wustl.edu/>
- [24] <http://homeobox.biosci.ki.se/>
- [25] <http://scop.mrc-lmb.cam.ac.uk/scop/>
- [26] http://www.biochem.ucl.ac.uk/bsm/cath_new/
- [27] <http://www.gene-regulation.com/pub/databases/transfac/cl.html>