Molecular Biology, Vol. 31, No. 4, 1997, pp. 483–497. Translated from Molekulyarnaya Biologiya, Vol. 31, No. 4, 1997, pp. 584–600.

UDC 577.214

Classification Scheme of Eukaryotic Transcription Factors

E. Wingender

Gesellschaft für Biotechnologische Forschung mbH, Mascheroder Weg 1, D-38124 Braunschweig, Germany
E-mail: ewi@gbf-braunschweig.de

Received October 31, 1996

Abstract—Transcription factors are proteins that regulate transcription by interacting either directly with specific regulatory signal sequences in the genome, or indirectly with factors binding to these sequence elements. In this paper, a first attempt is made to provide a classification scheme for eukaryotic transcription factors. Such a systematic classification may serve as a basis for detecting class-specific properties. For instance, systematic investigation of the DNA-binding specificity could shed light some light on the question about the existence of a protein–DNA recognition code.

Key words: transcription, transcription factors, gene regulation, protein classification, domains

INTRODUCTION

To our present knowledge, transcription is the main level at which gene expression is regulated. Transcriptional regulation is achieved by a functionally defined large family of proteins, the transcription factors [1, 2]. They interact with the DNA of promoters and enhancers in a more or less sequence-specific manner, recognizing defined sequence patterns and/or structural features. In contrast to prokaryotes, where the major control mechanism is to repress the normally active transcription machinery, eukaryotes have to meet much more complex requirements to coordinate the execution of genetic programs. This is achieved by directed activation of genes whose products are needed under certain cell conditions, in general only a few percent of all genes in the genome.

Once bound to DNA, these factors may influence transcription through several mechanisms:

(i) in most cases studied so far, they enhance the formation of the preinitiation complex at the TATA-box/initiator element through interaction of a *trans*-activation domain with components of the basal transcription complex either directly or through coactivators/mediators;

(ii) some transcription factors cause alterations in the chromosomal architecture, rendering the chromatin more accessible to the RNA polymerase(s);

(iii) some are auxiliary factors, optimizing the DNA conformation for the activity of another transcription factor;

(iv) some factors exert repressing influences, either directly by an active inhibiting domain, or by disturbing the required ensemble of transcription factors within a regulatory array (promoter, enhancer);

(v) finally, there is a group of transcription factors that do not directly bind to DNA but rather assemble into higher order complexes through protein–protein interactions.

At this point, it is necessary to attempt a definition of what we may designate a "transcription factor." On the basis of the functions listed above, we would propose the following definition: a transcription factor is a protein that regulates transcription after nuclear translocation by specific interaction with DNA or by stoichiometric interaction with a protein that can be assembled into a sequence-specific DNA–protein complex.

In this framework, proteins such as hsp90 that trap a transcription factor in the cytosol are excluded. Also excluded are regulatory enzymes, which exert their influence through catalytic rather than stoichiometric activities, and DNA-binding proteins such as histones, which do not interact with DNA in a sequence-specific manner. Among the high-mobility-group family of nuclear proteins, there are nonspecifically binding proteins such as HMG 1 and 2, and real transcription factors such as SRY or the Sox family (see below).

Most transcription factors are modularly composed. They may comprise

(a) a DNA-binding domain (DBD);

0026-8933/97/3104-0483$18.00 © 1997 МАИК Наука/Interperiodica Publishing

Table 1. Classification scheme for eukaryotic transcription factors

Level	Pattern of classification no.	Group designation	Criterion	Example
1	N	Superclass	General topology of DBD	Zinc-coordinating domains
2	N.N	Class	Structural blueprint of the DBD	Zinc finger nuclear receptors
3	N.N.N	Family	Functional criteria such as protein–DNA complex formation (DNA-binding specificity, multimerization behavior) or biological effect	T_3R/RAR (in contrast to steroid hormone receptors)
4	N.N.N.N	Subfamily	Mainly according to sequence similarity of the DBDs	RAR (retinoic acid receptor)
5	N.N.N.N.N	Genus	According to factor gene	RAR-α, RAR-β
6	N.N.N.N.N.N	Factor "species"	Initiation/splice/processing variants	RAR-β1, RAR-β2

Note: At each level, a decimal classification number is assigned as shown (with N for a real number). Each level was given a name chosen arbitrarily with some analogy to the taxonomy of biological species.

(b) an oligomerization domain (most factors bind to DNA as dimers, some also as higher order complexes) which in most cases forms a functional unit with the DBD;

(c) a *trans*-activating (or *trans*-repressing) domain, which is frequently characterized by a significant overrepresentation of a certain type of amino acid residues (e.g., glutamine-rich, proline-rich, serine/threonine-rich, or acidic activation domains);

(d) a modulating region which often is the target of modifying enzymes, mostly protein kinases;

(e) a ligand-binding domain.

CLASSIFICATION

The original signal causing the transcriptional regulation of a gene is a short DNA-sequence. Therefore, the nature of the DNA-binding domain of a transcription factor primarily determines its interaction with such a signal. Additionally, most factors bind to DNA as dimers (multimers), and the composition of these complexes may also influence the DNA-binding specificity. For these reasons, the classification scheme outlined in Table 1 is mainly based on the properties of the DNA-binding domain, wherever appropriate in conjunction with the dimerization domain. Subdivisions can be defined according to functional and structural criteria. One of the lower ranks in the hierarchy is given by the transcription factor genes, the lowest being the individual peptides derived from one gene by alternative splicing. In this scheme, a number is assigned to each category, and thus an unambiguous decimal classification code is assigned to each individual transcription factor.

Wherever it is reasonable to assume that two proteins from different biological species are functionally and/or structurally homologous, they are considered as one factor "genus"/"species." This is generally difficult or even impossible to assess, however, for factors of very different origin, e.g., vertebrate and insect proteins. In most cases, we therefore consider factors of vertebrates, insects, higher plants, and fungi separately, just to mention the most studied biological organisms. Although these categories reflect highly different levels in the taxonomy of biological species, a coarse separation of factors into these groups is of practical use.

According to this scheme, we recognize four large superclasses of DNA-binding domains, as well as several smaller and very small ones, some containing only very few if not a single representative. As shown in Table 2, the four major superclasses are

(1) factors that have just a stretch of mainly basic amino acid residues as the DNA-contacting motif;

(2) factors whose DNA-contacting surface is brought to a defined conformation by coordinated zinc ion(s);

(3) proteins that make use of the DNA-binding principle developed already by prokaryotes, the helix-turn-helix motif;

(4) factors whose DNA-contacting interface is a scaffold of suitably arranged β-strands fitting the minor groove.

Further, there is a large group of transcription factor DBDs whose three-dimensional structure has not yet been determined, or which even has not yet been mapped to a defined region within the protein molecule. They are grouped into a provisional "superclass 0."

The designation of the subcategories has been arbitrarily chosen in a way similar to the classification of biological species (Table 1). First, classes are defined according to the basal structural blueprint of the DNA-binding domains. These classes are already

Table 2. Classification of eukaryotic transcription factors

1. Superclass: Basic Domains

1.1. Class: Leucine zipper factors (bZIP)

1.1.1. *Family*: AP-1(-like) components

1.1.1.1. *Subfamily*: Jun

1.1.1.2. *Subfamily*: Fos

1.1.1.3. *Subfamily*: Maf

1.1.1.4. *Subfamily*: NF-E2

1.1.1.5. *Subfamily*: fungal AP-1-like factors

1.1.1.6. *Subfamily*: CRE-BP/ATF

1.1.1.0. *Subfamily*: Others

1.1.2. *Family*: CREB

1.1.3. *Family*: C/EBP-like factors

1.1.4. *Family*: bZIP / PAR

1.1.5. *Family*: Plant G-box-binding factors

1.1.5.1. *Subfamily*: Opaque-2 ("V")

1.1.5.2. *Subfamily*: EmBP-1 ("E")

1.1.5.3. *Subfamily*: HBP-1a ('Q")

1.1.5.4. *Subfamily*: TGA1a ("L/M")

1.1.5.5. *Subfamily*: TGA1b ("R")

1.1.6. *Family*: ZIP only

1.1.0. *Family*: Other bZIP factors

1.2. Class: Helix-loop-helix factors (bHLH)

1.2.1. *Family*: Ubiquitous (class A) factors

1.2.2. *Family*: Myogenic transcription factors

1.2.3. *Family*: Achaete-Scute

1.2.4. *Family*: Tal/Twist/Atonal/Hen

1.2.4.1. *Subfamily*: Lymphoid factors

1.2.4.2. *Subfamily*: Mesodermal Twist-like factors

1.2.4.3. *Subfamily*: HEN

1.2.4.4. *Subfamily*: Atonal

1.2.4.5. *Subfamily*: Pancreatic factors

1.2.5. *Family*: Hairy

1.2.5.1. *Subfamily*: Hairy

1.2.5.2. *Subfamily*: Esp

1.2.5.3. *Subfamily*: Fungal regulators

1.2.6. *Family*: Factors with PAS domain

1.2.7. *Family*: INO

1.2.8. *Family*: HLH domain only

1.2.0. *Family*: Other bHLH factors

1.3. Class: Helix-loop-helix/ leucine zipper factors (bHLH-ZIP)

1.3.1. *Family*: Ubiquitous bHLH-ZIP factors

1.3.1.1. *Subfamily*: TFE3

1.3.1.2. *Subfamily*: USF

1.3.1.3. *Subfamily*: SREBP

1.3.1.4. *Subfamily*: AP-4

Table 2. (Contd.)

1.3.2. *Family*: Cell-cycle controlling factors

1.3.2.1. *Subfamily*: Myc

1.3.2.2. *Subfamily*: Mad/Max

1.3.2.3. *Subfamily*: E2F

1.3.2.4. *Subfamily*: DRTF

1.4. Class: NF-1

1.4.1. *Family*: NF-1

1.5. Class: RF-X

1.5.1. *Family*: RF-X

1.6. Class: Heteromeric CCAAT factors

1.6.1. *Family*: Heteromeric CCAAT factors

1.7. Class: Grainyhead

1.7.1. *Family*: Grainyhead

1.8. Class: Cold-shock domain factors

1.8.1. *Family*: csd

2. Superclass: Zinc-coordinating DNA-binding domains

2.1. Class: Cys4 zinc finger of nuclear receptor type

2.1.1. *Family*: Steroid hormone receptors

2.1.1.1. *Subfamily*: Corticoid receptors

2.1.1.2. *Subfamily*: Progesterone receptor

2.1.1.3. *Subfamily*: Androgen receptor

2.1.1.4. *Subfamily*: Estrogen receptor

2.1.2. *Family*: Thyroid hormone receptor-like factors

2.1.2.1. *Subfamily*: Retinoic acid receptors

2.1.2.2. *Subfamily*: Retinoid X receptors

2.1.2.3. *Subfamily*: Thyroid hormone receptors

2.1.2.4. *Subfamily*: Vitamin D receptor

2.1.2.5. *Subfamily*: NBGF1-B

2.1.2.6. *Subfamily*: FTZ-F1

2.1.2.7. *Subfamily*: PPAR

2.1.2.8. *Subfamily*: EcR

2.1.2.9. *Subfamily*: ROR

2.1.2.10. *Subfamily*: Tll/COUP

2.1.2.11. *Subfamily*: HNF-4

2.1.2.12. *Subfamily*: CF1

2.1.2.13. *Subfamily*: Knirps

2.2. Class: Diverse Cys_4 zinc fingers

2.2.1. *Family*: GATA factors

2.2.1.1. *Subfamily*: Vertebral GATA factors

2.2.1.2. *Subfamily*: Fungal metabolic regulators

2.2.2. *Family*: Trithorax

2.3. Class: Cys_2His_2 zinc finger domain

2.3.1. *Family*: Ubiquitous factors

2.3.2. *Family*: Developmental/cell cycle regulators

2.3.2.1. *Subfamily*: Egr/Krox

2.3.2.2. *Subfamily*: Krueppel-like

Table 2. (Contd.)

2.3.2.3. *Subfamily*: GLI-like
2.3.2.0. *Subfamily*: Others
2.3.3. *Family*: Metabolic regulators in fungi
2.3.4. *Family*: Large factors with NF-κB-like binding properties
2.3.5. *Family*: Viral regulators
2.4. Class: Cys_6 cysteine–zinc cluster
2.4.1. *Family*: Metabolic regulators in fungi
2.5. Class: Zinc fingers of alternating composition
2.5.1. *Family*: $Cx_7Hx_8Cx_4C$ zinc fingers
2.5.2. *Family*: $Cx_7Hx_4Cx_4C$ zinc fingers
3. Superclass: Helix-turn-helix
3.1. Class: Homeo domain
3.1.1. *Family*: Homeo domain only
3.1.1.1. *Subfamily*: AbdB
3.1.1.2. *Subfamily*: Antp
3.1.1.3. *Subfamily*: Cad
3.1.1.4. *Subfamily*: Cut
3.1.1.5. *Subfamily*: DllI
3.1.1.6. *Subfamily*: Ems
3.1.1.7. *Subfamily*: En
3.1.1.8. *Subfamily*: Eve
3.1.1.9. *Subfamily*: Prd
3.1.1.10. *Subfamily*: HD-ZIP
3.1.1.11. *Subfamily*: H2.0
3.1.1.12. *Subfamily*: HNF1
3.1.1.13. *Subfamily*: Lab
3.1.1.14. *Subfamily*: Msh
3.1.1.15. *Subfamily*: NK-2
3.1.1.16. *Subfamily*: Bcd
3.1.1.17. *Subfamily*: XANF
3.1.1.18. *Subfamily*: PBC
3.1.1.0. *Subfamily*: Not assigned
3.1.2. *Family*: POU domain factors
3.1.2.1. *Subfamily*: I
3.1.2.2. *Subfamily*: II
3.1.2.3. *Subfamily*: III
3.1.2.4. *Subfamily*: IV
3.1.2.5. *Subfamily*: V
3.1.2.6. *Subfamily*: VI
3.1.3. *Family*: Homeo domain with LIM region
3.1.3.1. *Subfamily*: Homeo domain with LIM region
3.1.3.2. *Subfamily*: LIM-only transcription (co-)factors
3.1.4. *Family*: Homeo domain plus zinc finger motifs

Table 2. (Contd.)

3.2. Class: Paired box
3.2.1. *Family*: Paired plus homeo domain
3.1.3. *Family*: Paired domain only
3.3. Class: Fork head/winged helix
3.3.1. *Family*: Developmental regulators
3.3.2. *Family*: Tissue-specific regulators
3.3.0. *Family*: Other regulators
3.4. Class: Heat shock factors
3.4.1. *Family*: HSF
3.5. Class: Tryptophan clusters
3.5.1. *Family*: Myb
3.5.1.1. *Subfamily*: Myb-factors
3.5.1.2. *Subfamily*: Myb-like factors
3.5.2. *Family*: Ets type
3.5.3. *Family*: Interferon-regulating factors
3.6. Class: TEA domain
3.6.1. *Family*: TEA
4. Superclass: β-Scaffold factors with minor groove contacts
4.1. Class: RHR (Rel homology region)
4.1.1. *Family*: Rel/ankyrin
4.1.2. *Family*: Ankyrin only
4.1.3. *Family*: NF-AT
4.2. Class: p53
4.2.1. *Family*: p53
4.3. Class: MADS box
4.3.1. *Family*: Regulators of differentiation
4.3.1.1. *Subfamily*: MEF-2
4.3.1.2. *Subfamily*: Homeotic genes
4.3.1.3. *Subfamily*: Yeast regulators
4.3.2. *Family*: Responders to external signals
4.3.3. *Family*: Metabolic regulators
4.4. Class: β-Barrel/α-helix transcription factors
4.4.1. *Family*: E2
4.5. Class: TATA-binding proteins
4.5.1. *Family*: TBP
4.6. Class: HMG
4.6.1. *Family*: SOX
4.6.2. *Family*: TCF-1
4.6.3. *Family*: HMG2-related
4.6.4. *Family*: UBF
4.6.5. *Family*: MATA
4.6.0. *Family*: Other HMG box factors

widely accepted in the literature. They are used by the TRANSFAC database and are explained in its CLASS table [3].

The assignment of a particular transcription factor to a class is, in most cases, very clear-cut. Over the last years, there were only a few changes, one of the major alterations was to combine Myb-like and Ets-like factors into the class of factors with a tryptophan cluster (see below). In most classes, subgrouping into families is evident. Thus, within the bZIP class the distinction between AP-1-, CREB-, and C/EBP-like components is obvious from structural and functional considerations. In many cases, they should be classified further into subfamilies, whereas in other cases no such subdivision makes sense. This is therefore an optional level, which in the decimal classification is expressed either by consecutive numbering of the subfamilies or by the number 0. The latter is also used along with the ordinals (1, 2, ...) for those factors (or factor "genera") which cannot yet be assigned to any of the subfamilies known so far, rather than alloting a special subfamily (number) to each of them. To reflect the dynamics of the whole system, undefined (as yet) grouping may be generally characterized by the number "0" at any level (Table 2). For sake of simplicity, here the 5th and 6th positions (factor "genus"/"species") are omitted.

1. Basic Domain Superclass

The common characteristic of the proteins belonging to this superfamily is that they contact the DNA through a basic region which is unordered in solution but becomes α-helically folded upon binding to DNA [4–8]. The most prominent classes of this group are the bZIP and bHLH proteins, but we also tentatively assigned some other, much smaller classes such as the NF-1-like factors to this category.

There is no "consensus" motif for all basic domains of this superfamily. Even the basic regions of bZIP and bHLH factors diverge greatly, although their mode of interacting with DNA is strictly homologous: a specific α-helical dimerization domain provides linkage between two DNA-contacting basic regions which adopt a helical conformation when their positively charged side chains are neutralized by the phosphate backbone of the DNA.

No information is available about the three-dimensional structure of the DBD of the proteins that belong to the classes of NF-1-, RF-X-, CP1-, CP2-like or cold shock domain factors. However, all of them share the feature that they have certain small clusters of 2–4 basic amino acid residues with individual positively charged residues interspersed. The spacing of these "micromotifs" is highly variable (figure).

1.1. Basic–leucine zipper class. Functionally, four large subclasses can be distinguished: (i) components of the AP-1 factors, (ii) proteins that bind to cAMP-responsive elements (CRE) and similar sequences, (iii) C/EBP-like factors, and (iv) a subclass of plant factors that is frequently referred to as G-box-binding factors, mainly for historical reasons (see Table 2).

This classification illustrates that in this context, "function" mainly means DNA-binding specificity. Interestingly, this parallels the function of the factors in the cell physiology, at least to some extent. "Classical" AP-1 and CREB are involved in signal transduction processes: AP-1 is the factor that binds to TREs, TPA-responsive elements triggered by protein kinase C (PKC)-mediated pathways [9]; CREB mediates gene responses to enhanced cell cAMP levels, and thus the cAMP-dependent protein kinase A (PKA) is a key enzyme in the activation of this factor [10, 11]. However, this simplified view holds true only for the best studied family members, which are heterodimers of c-Fos/c-Jun as the most common AP-1 composition, whereas other components (FosB, Fra-1, Fra2, or JunB, JunD, respectively) may respond in a different and sometimes even antagonistic manner [12]. Additionally, factors that constitute "NF-E2 factors" are assigned to this group owing to their preferential binding to extended AP-1-like sequences [13]. They comprise the subgroups of NF-E2 p45-like and the Maf-like proteins [14, 15]. Finally, many ATF proteins such as the CRE-BP's also have to be assigned to this group since they preferably heterodimerize with Jun and, in some cases, with Fos factors, directing them to CRE-like DNA sequences (see below) [16]. Their idealized binding site is TGACGTAA. Some fungal factors have been grouped separately, since their homology with animal AP-1 components is difficult to assess.

Within the CREB/ATF-like factor group, there are only few members that really respond to cAMP with enhanced transcriptional activation. These are CREB (and some of its splice variants), ATF-1, and CREM-τ [17–19]. Some others are constitutive activators, and some of them act as repressors. In particular, most splice variants of the CREM family (cAMP responsive element modulators) are repressors, except the testis-specific CREM-τ [20, 21]. However, all of them bind to CRE-like sequences, the canonical motif of which is the palindromic TGACGTCA. The consensus of some ATF proteins is restricted to TGACG. In addition to vertebrate CREB and CREM, some insect and fungal homologs appear to belong to this group.

The third large group of the bZIP class comprises C/EBP and related factors. They were first described as distinct factors binding to some viral enhancers or to CCAAT boxes (enhancer- or CCAAT-binding pro-

```
basic region mMyoD                                  KRKTTNADRRKAATMRERRR
basic region mE12                                   QKAEREKERRVANNARERLR
basic region hc-Myc                                       KRRTHNVLERQRR
basic region hc-Jun                                RIKAERKRMRNRIAASKCRKRKLERIAR
basic region hc-Fos                                 KRRIRRERNKMAAAKCRNRRR
basic region hCREB                                  RKREVRLMKNREAARECRRKKKEYVK
basic region rC/EBP                       KAKKSVDKNSNEYRVRRERNNIAVRKSRDKAKQR
DBD of HAP2                          RILKRRYARAKRERKPYLHESRHKHAMRRPRGEGGRF
mCBF-B                     HRILKRRQARAKLEAEGKIPKERRKYLHESRHRHAMARKR
grainyhead              RDTAVFHRGYCQIKVFCDKGAERKTRDEERRAAKRKMTATGRKKLDELYHPVTDR
rNF-1/L                                                   RKRKYFKKHEKRMSKEEERAVK
hDbpB csd                                                  KKVIATKVLGTVKWFNVR
hDbpB csd                                        KKNNPRKYLR / KKNNPRKYLR
RF-X1                                              KLIRSVFMGLRTRRLGTRGNSKYHYYGLRIK
```

Basic domains of some superclass I transcription factors. Shown are the basic regions of the DNA-binding domains of murine MyoD (MYOD_MOUSE, 108–121), murine E12, human c-Myc (MYC_HUMAN, 355–367), human c-Jun (AP1_HUMAN, 252–279), human c-Fos (FOS_HUMAN, 139–159), human CREB (CREB_HUMAN, 284–309), rat C/EBPα (CEBA_RAT, 273–306), yeast HAP2 (HAP2_YEAST, 170–214), mouse CBF-B (CBFB_MOUSE, 272–311), Drosophila Grainyhead (ELF1_DROME, 784–838), rat NF-1/L (NFIL_RAT, 32–53), two parts of the cold shock domain of human DbpB (CBFX_HUMAN, 52–69 and 92–101, the second region in two alternative "alignments"), and human RF-X1 (RFX1_HUMAN, 482–512). Basic residues are shown in bold. The basic regions has been arranged according to an optimal match of two "clusters" of basic residues flanking a (relatively) hydrophobic center.

tein, EBP or CBP), later recognized as identical proteins [22, 23]. C/EBP factors bind to DNA with very broad sequence specificity, neither a consensus string nor a sufficiently restrictive weight matrix can be deduced from the experimentally proven binding sites. As an idealized palindromic recognition motif, the decamer ATTGCGCAAT has been suggested [24]. There is little cross-dimerization with members of the two bZIP families described above, except CREB-2 which interacts with C/EBP variants α, β, γ, and ε [25], and the newly discovered C/ATF which through heterodimerization with C/EBPs can direct these factors to CRE sites [26].

Very similar to the C/EBP proteins are the DNA-binding domains of DBP factors (referred to as "D-element binding proteins" because of the D-site within the albumin promoter as the first described sequence to interact with these factors) [27]. However, they exhibit some significant differences which confer a much more restrictive DNA-binding specificity on these proteins. Most of them do not heterodimerize with C/EBP factors, and if so, these complexes do not productively bind to DNA [25].

The fourth group of bZIP factors comprises exclusively plant transcription factors, which are collectively referred to as G-box-binding proteins. However, their real recognition motif rather comprises an ACGT core, which in many cases is embedded in G/C-rich environments, such as CCACGTGG for GBF-1 [28]. Conventional tree construction for their bZIP domains reveals two big and three small subgroups. When we tried to construct consensus sequences for these subgroups, it turned out that there is one position near the N terminus of the basic region (position 5) that is indicative for all members. Therefore, the subgroups can be labeled according to the amino acid residue that is found thereat.

1.2. Basic–helix-loop-helix class. Similarly to the bZIP factors, the bHLH proteins contact DNA through their basic region (b), and do so as homo- or heterodimers that form through the helix-loop-helix domain as the interface [6–8]. Not included in this class are bHLH factors that also have a leucine zipper as an additional or augmenting dimerization interface C-terminal to the HLH domain (see below, 1.3). There are some leucine zipper-like motifs in other factors as well, but they are in the region N-terminal to the bHLH domain, and their functional impact is doubtful.

Two well-recognizable groups within this class are (.1) the E-box-binding factors encoded by the E2A, E2-2, and related genes, and (.2) the myogenic factors with bHLH domains. Well-known members are E12/E47 and MyoD, respectively. A third group is constituted by the Achaete–Scute proteins and their vertebrate homologs.

However, other groups are more difficult to identify with conventional alignment approaches: different relationships are found depending on whether the whole bHLH domains, the isolated basic regions, or the first or the second α-helices are analyzed. We therefore applied an additional algorithm which exploited positional correlation properties of the bHLH domains. With this approach, we identify three additional groups of bHLH factors. One big group comprises the subgroups of lymphoid factors (like Tal-1), mesodermal developmental regulators (like Twist), HEN factors, Atonal-related proteins, and pancreatic factors. In general, they have an aromatic residue (Phe, Tyr, or His) in position 40 and a loop length of 10 or 11 residues.

Another group is that of Hairy/Enhancer of split factors. They have in common an arginine residue in

position 14 and a loop length of 14–16 residues. Based on this simple criterion, the yeast factor PHO4 was assigned to this group as well, and also Nuc-1 from *Neurospora crassa* because of the general sequence similarity of its bHLH domain with that of PHO4.

Some factors with only remote similarity to other bHLH factors have been grouped together since they share a so-called PAS domain [29]. In addition to the Drosophila protein Single-minded (Sim), the nuclear receptor AhR is a known member of this group. There are also some other families not considered here.

1.3. Helix-loop-helix/leucine zipper class. Here we can distinguish between two families from a functional point of view. The first family comprises several subfamilies of ubiquitous or at least widespread transcription factors, most of them constitutively present and active. Prominent members of this family are TFE-3 and USF. In contrast, four subfamilies of cell-cycle controlling factors are assembled into the second family, showing a pairwise functional relation: factors of the c-Myc and Mad/Max subfamilies heterodimerize with each other, as do those of the E2F and DRTF subfamilies. Particularly divergent and thus lacking an obvious bHLH "consensus" are sequences of the E2F and DRTF factors; they have been assigned to this group exclusively on the basis of topological considerations.

1.4. NF-1-like proteins. There are hitherto no structural data available about NF-1 or other factor classes that we tentatively arranged in this superclass. However, they share a basic DNA-contacting domain which reveals a high α-helical probability when analyzed by programs for prediction of secondary structures (data not shown). NF-1 transcription factors have a high degree of homology to each other, but do not exhibit significant sequence similarity with any other factor group known so far. As one of their most prominent features, the proline-rich *trans*-activation domain has been described [30], but this characteristic is independent of the DNA-binding domain and thus does not contribute to the classification criteria applied here. NF-1 factors are encoded by at least four to six different genes, producing numerous splice variants [31–34].

1.5–1.7. Overview of classes. The information on these classes is limited. RF-X factors (1.5.1) have been disclosed as proteins that bind to the so-called X box of the MHC genes, or to certain viral enhancers, then being referred to as EF-C (enhancer factor C). It has been reported that the RF-X factors may exhibit a weak homology with the bHLH consensus [35], but an equally vague similarity with NF-1 proteins can be observed as well. Four known mammalian genes give rise to at least five different splice variants of transcription factor RF-X [36].

Factors of the CP-1 family (1.6) had been characterized from human, murine and rat sources. Since they bind to CCAAT-like sequences, they were designated CCAAT-protein 1 (CP-1) from human cells and CBF (CCAAT box factor) from rat cells, respectively [37, 38]. In murine cells, these factors were discovered first as binding activity that interacts with the Y box of MHC genes and thus was designated NF-Y (nuclear factor Y) [39]. It was recognized very early that they consist of three subunits, one of them (CP-1B = CBF-B = NF-YA, the yeast homolog being HAP2) containing the basic DNA-binding domain which can be assigned to this superclass.

The CP-2 transcription factor (1.7) also binds to CCAAT(-like) sequences and was originally described as probably being similar to CP-1, but cloning approaches revealed that it is homologous to Grainyhead of Drosophila [40]. Though binding to DNA on its own, probably as a homodimer, it may be part of a larger complex [41].

1.8. Cold shock domains. Factors of this class have been shown to stimulate or to repress transcription, but it is not completely clear whether their real task in the cell may be instead to control translation by interacting with RNA. These proteins reveal several basic regions with some homology to protamines, i.e., a high content of positive side chains with interspersed proline residues [42]. However, their real DNA-binding domains have been reported to be the "cold shock domains" [43]. Proteins with these domains are known in bacteria to induce certain genes when the organism is exposed to low temperatures. While protamines have been reported to adopt a helical conformation upon binding to the DNA and concomitant neutralization of positive charges, the cold shock domains mostly exhibit β-strands [43]. If further physicochemical investigations confirm this prediction, these factors may have to be reassigned to superclass 4 of the scheme proposed here.

2. Zinc-coordinating DBD Superclass

It was known since 1983 that the Pol III transcription factor TFIIIA requires zinc as cofactor for its DNA-binding activity [44]. After successful cloning, the repetitive pattern of cysteine and histidine residues within the polymerase III transcription factor TFIIIA was discovered and lead to the model of "zinc fingers" of the Cys_2His_2 type [45]. Shortly after, sequences coding for the estrogen receptor were cloned and revealed a somewhat similar arrangement of cysteines only (Cys_4), with no additional homologies [46, 47]. Nevertheless, a zinc finger model was proposed for these factors as well and was subsequently proven, with a minor alteration in the Cys pattern [48].

In addition to these two large classes of zinc finger proteins, there are many factors revealing Cys_4 zinc

fingers of variable composition, such as the GATA factors, some fungal regulators, and adenovirus E1A. Finally, the yeast GAL4-like regulators are grouped together as Cys_6 "zinc clusters." We did not include the tumor suppressor gene product p53 into one of these classes, since its zinc-coordinating domain is not part of the DNA-contacting surface of this protein (see below).

2.1. Nuclear receptors with Cys_4 zinc fingers. All these proteins comprise two zinc finger motifs, the first being responsible for specific DNA-binding, the second for stabilization [49]. In the DNA-bound complex, they dimerize through the region around the first cysteine doublet of the second zinc finger [49]. They are activated by binding a low-molecular-weight ligand which, however, has not yet been identified in many cases, and thus the general validity of this statement remains to be proven. The major difference used for further classification within this class is the nature of ligands and the heterodimerization behavior.

The first family comprises the steroid hormone receptors, which interact with the cytoplasmic heat shock protein 90 (hsp90). As shown in detail for the glucocorticoid receptor, this interaction retains the receptor in the cytoplasm and imposes a conformation suitable for binding the ligand [50, 51]. The complex subsequently dissociates and allows the receptor to translocate into the nucleus. Most of the steroid receptors possess a Gly-Ser dipeptide between the cysteine doublet at the end of the first finger. Only the estrogen receptor has a Glu-Gly dipeptide in this position, like the factors of the second family. This group comprises the receptors for thyroid hormone, retinoic acid and retinoids, for vitamin D3, the insect hormone ecdysone, and many "orphan" receptors. In general, they are nuclear factors that are activated by ligand-binding. To our present knowledge, they do not interact with hsp90 [52]. Some of them may bind to DNA as homodimers, but in most cases, high-affinity DNA-binding is achieved by heterodimerization with retinoid receptors RXR.

2.2. Diverse Cys_4 zinc finger factors. First of all, the factors that exhibit a GATA-like structure have to be mentioned in this class. GATA-1 was discovered as an erythroid-specific factor binding to regulatory elements in the globin genes [53]. Subsequently, several related factors were found and cloned from vertebrate as well as from fungal sources. All of them contain one or two zinc finger-like motifs [54]. The consensus for the first finger is $CXNCx_4TPLWRRX_3GXXLCNACgl$.

If a second zinc finger of the GATA type is present, it has P10T and L21V substitutions. Interestingly, up to now only double zinc fingers have been discovered in vertebrates (and nematodes), whereas in fungi only single-GATA-finger proteins have been isolated. Nevertheless, this motif has been conserved remarkably in evolution. The structure of the GATA-type zinc finger motif has been resolved by NMR studies [55]. It turned out that in those factors that contains two GATA-like zinc finger motifs, the C-terminal one performs specific DNA-contacts, whereas the N-terminal contributes to the overall stabilization of the protein–DNA complex [55]. The C-terminal zinc finger exhibits a topology similar to the DNA-binding finger of nuclear receptors (where it is the N-terminal one), an α-helix being exposed to the major groove of the DNA; moreover, the C-adjacent sequence of the GATA-1 molecule which exhibits an extended structure is also involved in determining the DNA-binding specificity [55].

The second family tentatively grouped into this class, Trithorax (Trx), does not reveal appreciable similarity with the GATA-type zinc finger motif beyond the cysteine pattern. This factor, as well as a Trx-like human protein (Hrx), possesses several zinc finger-like motifs of different type and length, some of them with either cysteine of the second doublet replaced with histidine, and not exhibiting a recognizable consensus sequence. More experimental data are required for the final classification of Trx-like proteins.

2.3. Cys_2His_2 zinc finger domains. This is a very large family of proteins with different function, many but not all being true transcription factors. Some zinc finger motifs of this type may rather serve as RNA-binding or dimerization domains. Those for which a role in transcriptional control is evident or at least highly suggestive have been classified according to the following scheme. Since for these factors no general view is as yet emerging that would link their mode of building protein–DNA complexes to their biological role, the proposed scheme defines families according to the major biological effect their members may exert. Thus, the Cys-His-zinc finger factors that are ubiquitous and play a role in housekeeping gene expression have been compiled in the first family: the (classical) Pol III factor TFIIIA, the Sp1 family, and the repressing/activating protein YY1.

Among the second family, which comprises a large number of developmental regulators or factors involved in cell cycle control, the zinc finger motifs have been subclassified according to their gross structure, i.e., the number of amino acids between the two cysteine residues, the two histidines, and between the Cys and the His doublet. Since all factors of this family contain more than one zinc finger motif, often of different types. In all these cases, however, a predominant type of zinc finger motif could be identified and was taken for the classification of the factor. Thus, we established the following subfamilies: Egr/Krox factors with three adjacent DNA-binding zinc fingers of 2/4,12,3 type ($CX_{2,4}C–X_{12}–HX_3H$), where "adjacent'

generally means a distance of 6–7 residues; the Krueppel-like factors which comprise 4–13 zinc finger motifs mainly of 2,12,3 type ($CX_2C–X_{12}–HX_3H$); the GLI-like factors with 4-5 adjacent zinc fingers, mainly of 4,12,3–4 type ($CX_4C–X_{12}–HX_{3-4}H$); and a less defined subfamily of factors that possess 2–3 zinc fingers, at least one of them being of 2,12,4–5 type ($CX_2C–X_{12}–HX_{4-5}H$); they may be scattered over the molecule.

These subfamilies may also contain yeast factors which, however, may slightly deviate from the given criteria.

In addition, a family of fungal regulators seems to be justified because of the sequence similarity of its members. The paradigm of this family is ADR1 from *Saccharomyces cerevisiae*. And finally, not fitting into any of these families and therefore constituting a family of its own is the large T antigen of SV40.

2.4. Cys_6 zinc cluster. The most prominent member of this class is GAL4, and up to now, only fungal transcription factors have been identified that possess this motif. In contrast to the zinc finger motifs described above, the number of zinc-coordinating residues is not a multiple of 4. Rather, six cysteines are arranged as a binuclear cluster of the type Zn_2Cys_6, comprising two tetrahedrons that share one edge [56]. Thus, two of the six cysteines complex with two zinc ions each. The outside edges of the double tetrahedron are α-helically folded, which together with some turns may provide DNA contacts [57–60].

3. Helix-Turn-Helix Superclass

This is a particularly large and heterogeneous family of transcription factors. Their DNA-binding motif appeared very early in evolution, since it is found in prokaryotic and bacteriophage regulators as well as in mammalian transcriptional activators and repressors [61–63]. Presumably owing to this property, they fulfill in eukaryotes elementary functions such as developmental regulation and determination of differentiation processes.

They act by activating or frequently by repressing genes, many of which encode other developmental regulators. They constitute a highly complex regulatory network of positive and negative activities and feedbacks. In some cases they just occupy arrays of repetitive binding sites, starting at a few high-affinity sites and, using these as nucleation centers, progressively occupying low-affinity sites as well [64, 65]. In the case of certain repressing molecules, the may even cover a whole regulatory region.

Basically, we divide the factors that exhibit this structural motif within their DNA-binding domain into six big classes: the homeo domain, the paired box, the fork head/winged helix domain, the heat shock factors, the tryptophan clusters, and the TEA domain. Finally, the DBD of yeast Adf-1 may contain a helix-turn-helix motif, but since it does not match any of the other classes, it constitutes a class of its own.

3.1. Homeo domain. Proteins of this class can first be distinguished according to the absence or presence of additional functional domains that may be either directly or indirectly involved in DNA-binding. We thus find (.1) the homeo-only group, (.2) POU factors, (.3) LIM factors, and (.4) homeo domain factors with zinc finger motifs.

POU factors have a second DNA-binding motif, the POU-specific domain (POU_s), which also exhibits a helix-turn-helix motif [66, 67]. They have nevertheless been assigned to the class of homeo domain factors, since the homeo domain appears to be the predominant DNA-binding principle and no factors have yet been found to possess a POU_s domain but no homeo domain. It remains to be investigated whether the zinc finger motifs found in some homeo domain factors still have the potential to act as DNA-binding domains or whether they have adopted another role in these proteins, e.g., as a protein–protein interaction interface. Their final location within the classification scheme depends on the answer to this question.

The same appears to be true for the cysteine-histidine-rich LIM domain which might have developed from a zinc finger ancestor and serve now as a module for interprotein contacts [68, 69]. Of the LIM-only proteins, only a few have been shown to influence transcription, and thus constitute a distinct subgroup in the group of LIM factors, whereas most of them exhibit other functions.

The homeo domain factors have already been classified according to a scheme developed by Bürglin [70]. We adopted this scheme for the group of homeo domain factors. Thus, two big subgroups of related homeo domain transcription factors can be distinguished, the Abd-B (Abdominal-B) and the Antp (Antennapedia) group. Among the 15 additional groups, there is a group of Prd (Paired)-like homeo domains which should not be confused with the Paired Box class (see Table 2).

Another family is constituted by the POU factors. They are subgrouped according to the classification scheme proposed by He *et al.* [71] who originally identified four families, numbered I–IV. Later, additional "classes" were established when more nonassignable POU factors were discovered. Thus we now use the classification scheme proposed in Table 2 disclosing six clearly distinguishable subfamilies. It should be noted, however, that the members of some subfamilies (such as 3) are much more similar to each other than those of others (e.g., subfamily 5).

Table 3. Properties of superclass 4 factors

	DBD class			
Properties	Rel homology region	MADS box	TBP	HMG domain
Embracing the DNA	+	–	+	+
DNA bending/kinking	+	+++	+++	+++
		DNA is wrapped around the protein kernel	Insertion of the protein into the minor groove distorts DNA	Insertion of the protein into the minor groove distorts DNA
β-Scaffold	+	+	+++	–
	β-Barrel exposing DNA-contacting loops	β-Sheet exposing a DNA-contacting α-helix and an extended region	DNA-contacting β-sheet	L-shaped DBD of three α-helices and an extended region
DNA-binding by loops/extended regions	+	+	–	+
Minor groove contacts	(+)	+	+++	+++
	Potential, depending on the site architecture	In addition to major groove contacts	Exclusively	Exclusively

3.2. Paired box. These transcription factors (PB) have been assembled into a separate class, since the predominant DNA-binding domain of these factors appears a paired box rather than the paired-type homeo domain (HD) above.

3.3. Fork head/winged helix. The Drosophila *fork head (fkh)* gene was originally also detected as a development-regulating gene [72]. Subsequently, it was discovered that some tissue-specific factors such as the liver-enriched HNF-3 revealed certain sequence similarity with Fkh [73], and after the three-dimensional structure of HNF-3γ had been solved it became clear that these proteins belong to the same structural class [74]. Moreover, it was evident from the crystal structure of HNF-3γ that the fork head domain folds in a helix-turn-helix conformation whose detailed structure was then described as a "winged helix."

3.4. Heat shock factors. While exhibiting some unusual properties, like homotrimerization through a leucine zipper-like domain [75], the heat shock factors possess a DNA-binding domain with a topology which is very similar to that of "canonical" helix-turn-helix proteins, in particular to that of the prokaryotic catabolite activator protein (CAP) [76]. However, it has a longer turn (five instead of three amino acid residues) and the second α-helix in a somewhat displaced position.

3.5. Tryptophan clusters. This class comprises three important families, the Myb-like proteins, the interferon-regulatory factors, and the Ets-like proteins. They are characterized by a definite tryptophan pattern. In Myb-like factor molecules, one to three such motifs are found, which comprise three Trp residues with 17–20 (mostly 18-19) intervening amino acids; occasionally, a tyrosine, phenylalanine, histidine or isoleucine may substitute for one of the tryptophan residues.Factors of the Ets family generally harbor one Trp_3 cluster, the spacing is 17–21 amino acids, one tryptophan may occasionally be replaced with a tyrosine. The IRF-related factors possess one Trp cluster of 5 residues which are separated by 11–19 residues.

As demonstrated for Myb-like DNA-binding domains, the tryptophan residues constitute a hydrophobic core, around which a helix-turn-helix fold is formed [77, 78]. A helix-turn-helix structure has also been evidenced for an Ets-like factor [79], whereas for IRF-like proteins a similar topology may be assumed only by homology.

3.6. TEA domain. This domain has been identified as a region conserved among transcription factors TEF-1, TEC1, and abaA. This domain in TEF-1 has been shown to interact with DNA, although two additional regions may also contribute to DNA-binding. It is predicted to fold into three α-helices, with a random region of 16–18 residues between helices 1 and 2, and a short stretch between helices 2 and 3 (3–8 residues) [80]. Although the latter region does not exhibit the characteristics of known helix-turn-helix motifs, we tentatively assigned this class to the superclass of helix-turn-helix factors.

4. Superclass of β-Scaffold Factors with Minor Groove Contacts

In this superclass, we grouped together six transcription factor classes for which it is difficult to define one common characteristic, but a series of characteristics can be listed (Table 3).

Thus most of these DNA-binding domains have a scaffold of β-strands (β-barrel, β-sandwich, β-sheet) which serves as a structural template to expose DNA-contacting secondary structure elements, be it α-helices, loops, or β-strands themselves. These factors

make extensive DNA-contacts either by embracing the DNA for at least half a turn, as shown for a member of the Rel class [81, 82], for p53 multimers [83], for TBP [84–86], and for HMG domains [87, 88] (classes 4.1, 4.2, 4.5 and 4.6, respectively), or wrap the DNA around the protein core. The latter was shown for a MADS-box factor [89] and for the viral regulator E2 [90] (classes 4.3 and 4.4).

A feature common for many of these factors is that they bind to A/T-rich cores with C/G-rich flanks (4.1–4.5), the central interactions made up by minor groove contacts (4.1–4.6). In the case of the Rel protein NF-κB p50, this central minor groove contact depends on the precise architecture of the binding site [81, 82]. These minor groove interactions with the DNA give rise to drastic distortions of the double helix, with the possible exception of the Rel class.

4.1. RHR (Rel homology region). In this class, we find activators and inhibitors; these factors may have Rel-like DNA-binding domains, ankyrin repeat protein–protein interfaces, and both types of domains. We encounter the hitherto unique case that one gene codes for different proteins of highly divergent structure and function: the NFKB1 gene encodes the NF-κB p105 precursor, which is an inhibitor of nuclear translocation comprising both a Rel domain and ankyrin repeats (see [91] for review). This precursor is processed to a nuclear protein, p50, a DNA-binding subunit of NF-κB, which as a homodimer is basically transcriptionally inert, whereas it is a potent *trans*-activator when heterodimerized with p65. However, the same gene also gives rise to an IκB-like inhibitor (IκB-γ) by usage of a downstream initiator codon on an alternative transcript [92].

We grouped all Rel-only and Rel/ankyrin-factors together since they share the presence of a certain DBD-type, thus having p105 and p50 still together, but separated IκB-γ into the second group of this class, the ankyrin repeat-only proteins. This is up to now the only case in the proposed classification scheme where two factors encoded by the same gene appear in distinct families.

The structure of the Rel-type DBD has been resolved for the NF-κB1 p50 homodimers. They exhibit a bipartite subdomain structure, each subdomain comprising a β-barrel with five loops that form an extensive contact surface to the major groove of the DNA [81, 82]. Particularly, the first loop of the N-terminal subdomain (the highly conserved "recognition loop") performs contacts with the recognition element on the DNA, but other loops are involved. The fact that the main DNA-contacts are made through loops has been suggested to provide a high degree of flexibility in binding to a range of different target sequences. Augmenting interactions are achieved by two α-helices within the N-terminal part that form strong minor groove contacts to the A/T-rich center of the κB-element. In p65, the sequence between both α-helices is much shorter and even helix 2 is truncated. This may explain why HMG proteins interact with the central nucleotides of a κB site that is bound by a p50/p65 heterodimer, since here less intensive minor groove contacts can be performed. The second, C-terminal domain is necessary mainly for protein dimerization. It has been noticed that binding sites for Rel- and bZIP-class factors frequently constitute "composite elements," i.e., short regulatory regions that comprise binding sites for two or more transcription factors, closely cooperating in a synergistic or antagonistic manner and thus conferring qualitatively new characteristics onto the gene under control [93]. A particularly intimate cooperation is found between an AP-1-like component and a distinct group of the Rel-class, constituting a complex which originally was even described as one factor, NF-AT [94]. This was first discovered in activated T-cells as a DNA-binding activity which bound to certain enhancer elements of, e.g., the interleukin 2 gene [95, 96]. NF-AT was disclosed to consist, at least in some cases, of a Fra-1/JunB heterodimer [97], while several peptides with a Rel-like domain were cloned that may take over the part of the additional component.

4.2. The p53 class. The RHR-type DBD reveals some similarity with that of p53 in that both interact with DNA through a loop at the N-terminal end of a "β-sandwich." However, p53 also makes additional major groove contacts through an α-helix, which is part of a loop-sheet-helix motif at the C-proximal end of the DBD [83]. The loop belonging to this motif also binds to the DNA, but in the minor groove. While Rel-domain factors interact with a sequence comprising an oligoG at its 5' and an oligoC at its 3' end with an A/T-rich center, the p53 recognition site follows the consensus RRRCWWG. In the crystallized p53/DNA complex, the sequence GGGCAAG was used and, similar to the NF-κB p50/DNA complex, the minor groove contacts have been mapped to the AA dinucleotide. It is noteworthy that one residue within the minor groove-contacting loop, Arg-248, is the most frequently mutated residue in the p53 tumor suppressor gene.

4.3. MADS box. DNA-bound dimers of the MADS-box protein SRF (serum response factor) also reveal a scaffold made of β-strands which expose two α-helices (one per subunit) to the DNA. They are arranged nearly parallel to the minor groove in the center of the CCW_6GG consensus, where they form base contacts, while a randomly coiled N-terminal extension of either helix performs major groove contacts to the flanking C/G nucleotides [89]. Up to now, these structural properties have been resolved for SRF only. Because of sequence homologies, several other transcription factors can be easily assigned to this

class as well, and we grouped them according to their biological function: (.1) Regulators of differentiation, some of them (the MEF2-like proteins) exhibit additional conserved sequences (the MEF2 box) in addition to the MADS box, (.2) responders to external signals such as SRF itself, and (.3) metabolic regulator(s) of yeast.

4.4. β-Barrel/α-helix factors. Of this class, we know only two viral examples, E2 from bovine papilloma virus (BPV), and EBNA-1 from Epstein–Barr virus (EBV) [90, 98]. In the homodimeric DBD of these regulators, eight β-strands form a β-barrel that exposes two α-helices on its surface. These α-helices bind the DNA through specific major groove contacts. E2 bends the DNA smoothly over the β-barrel scaffold, thus somewhat resembling the DNA complex of the MADS-box factor SRF where the DNA is also wrapped around the protein core. EBNA-1 has hitherto been reported as an replication-stimulating protein that binds to the origin of replication, but since E2 has been described to be essential for both replication and transcription processes, we tentatively assigned EBNA-1 to this class of transcription factors.

4.5. TATA-binding protein(s). Principally, only one kind of this class is known, the TATA-binding protein. However, there are big species-specific differences in the N-terminal parts of this molecule from, e.g., human and Drosophila, and some organisms have more than one TBP gene (e.g., *Arabidopsis thaliana* and wheat). An extensive β-pleated sheet forms a quasisymmetric saddle-like structure which embraces the DNA by inserting into the minor groove, thereby imposing a significant distortion of the DNA towards the major groove [84–86], at a nearly right angle.

4.6. High mobility group (HMG) domain. This class of proteins is highly divergent when considering their function. Some of them are pure architectural components without the capability to recognize specific DNA sequences, such as HMG1, and are not considered as transcription factors. Some other HMG factors, such as SRY, exhibit some sequence-specific binding, but also preferably recognize and bind to certain structural characteristics such as four-way-junctions [99]. Apparently, their role is mainly to help adjusting the most suitable DNA conformation for transcriptional stimulation. Others, such as LEF-1, may be true transcriptional activators, with a recognizable *trans*-activation domain [100]. All of them have a typical L-shaped HMG domain. It should be noted that there are factors which are also referred to as high-mobility-group proteins such as HMGI(Y). They also play an important augmenting role for the action by other transcription factors such as NF-κB [101]. However, they do not posses a HMG domain nor do they reveal significant homology with other regions of HMG domain factors and thus are not members of this class.

The L-shaped HMG domain consists of three α-helices and an extended N-terminal extension of the first helix [87, 88]. The latter together with helix 1, which contains a kink, form the long arm of the "L," whereas helices 1 and 2 form the short arm. Binding to the minor groove induces a sharp bending of the DNA by more than 90E, away from the bound protein. The overall topology of the DNA–protein complexes resembles somewhat that of the TBP–TATA box complex. Mainly for this reason, the HMG domain proteins are included in this class of transcription factors.

In addition to several HMG factors which could not be assigned to a specific family, the other proteins of this class were grouped according to sequence similarity into five families: HMG domain factors of Sox, TCF1, HMG2, UBF, and MATA types. One of the best-studied members of the first family is SRY, the putative sex-determining factor. The TCF family comprises several factors mainly found in lymphoid cells, the HMG2 family (although HMG2 itself is not considered as a transcription factor) contains proteins like SSRP1 (structure-specific recognition protein 1), possibly also involved in DNA recombination events. The UBF family comprises this Pol I factor (UBF1, UBF2), and the MATA family up to now consists of only one member, yeast mat-Mc.

DISCUSSION

A classification scheme for transcription factors like that proposed here may be of practical use for handling data about these proteins, for instance in databases. It may also help to assess the function of newly discovered proteins which obviously fall in one of the defined categories. Moreover, it may give some more general clues to the structure/function relationship of the proteins the scheme comprises, and it may give hints on evolutionary relations within that group.

However, any classification attempt encounters some general problems. First of all, subgrouping of some classes or families is immediately evident, since the members are obviously related and clearly distinct from those of the other subgroups. In other groups, however, the subgrouping is much less clear but nevertheless reasonable, and again in others may be impossible at all, at least on the basis of present knowledge. Moreover, not all subgroupings may follow the same or even comparable criteria, due to the intrinsic structural and functional peculiarities of the members of these groups: sometimes, pure sequence similarity of the DBDs is sufficient for a statisfying classification, sometimes it is necessary to focus on particular residues, in other cases additional sequences/domains have to be taken into account as well which may be responsible, e.g., for certain

dimerization behavior. The classification scheme proposed here makes an attempt to consider these particular features of each class or family, but has to be considered as a dynamic system anyway, since our present knowledge about the objects of this classification is far from being complete.

The proposed classification scheme for eukaryotic transcription factors may become part of a more comprehensive scheme for all proteins. However, the general problems arises, how proteins can be classified which are composed of distinct modules. In fact, classification schemes can be established for these modules. For those proteins considered here, a completely different scheme would be obtained when choosing transcription activation domains instead of DNA-binding domains. Therefore, and on a long term, the only comprehensive classification scheme of proteins has to reflect the hierarchical levels of protein structures: the lowest level has to describe the "modules" (structural/functional domains), and a higher level will have to classify the modular composition of polypeptide chains. Additional levels may then consider subunit association and other increasingly complex features.

ACKNOWLEDGMENTS

I am indebted to O.V. Kel' and A.E. Kel' for numerous stimulating discussions, and to J. Collado-Vides and K.H. Seifart for critically reading the manuscript. This work was supported by grants of the German Ministry of Education, Science, Research and Technology (project no. 01 1B 306 A) and by the European Commission (BIO4-CT95-0226).

REFERENCES

1. McKnight, S.L. and Yamamoto, K.R., *Transcriptional Regulation*, Cold Spring Harbor Laboratory, 1992.
2. Wingender, E., *Gene Regulation in Eukaryotes*, Weinheim: VCH, 1993.
3. Wingender, E., Dietze, P., Karas, H., and Knüppel, R., *Nucleic Acids Res.*, 1996, vol. 24, pp. 238–241.
4. Patel, L., Abate, C., and Curan, T., *Nature*, 1990, vol. 347, pp. 572–574.
5. Weiss, M.A., Ellenberger, T., Wobbe, C.R., Lee, J.P., Harrison, S.C., and Struhl, K., *Nature*, 1990, vol. 347, pp. 575–578.
6. Ferre-D'Amare, A.E., Prendergast, G.C., Ziff, E.B., and Burley, S.K., *Nature*, 1993, vol. 363, pp. 38–45.
7. Ellenberger, T., Fass, D., Arnaud, M., and Harrison, S.C., *Genes Dev.*, 1994, vol. 8, pp. 970–980.
8. Anthony-Cahill, S.J., Benfield, P.A., Fairman, R., Wasserman, Z.R., Brenner, S.L., Stafford III, W.F., Altenbach, C., Hubbell, W.L., and DeGrado, W.F., *Science*, 1992, vol. 255, pp. 979–983.
9. Angel, P., Imagawa, M., Chiu, R., Stein, B., Imbra, R.J., Rahmsdorf, H.J., Jonat, C., Herrlich, P., and Karin, M., *Cell*, 1987, vol. 49, pp. 729–739.
10. Montminy, M.R. and Bilezikjian, L.M., *Nature*, 1987, vol. 328, p. 175.
11. Mellon, P.L., Clegg, C.H., Correll, L.A., and McKnight, G.S., *Proc. Natl. Acad. Sci. USA*, 1989, vol. 86, pp. 4887–4891.
12. Chiu, R., Angel, P., and Karin, M., *Cell*, 1989, vol. 59, pp. 979–986.
13. Igarashi, K., Kataoka, K., Itoh, K., Hayashi, N., Nishizawa, M., and Yamamoto M., *Nature*, 1994, vol. 367, pp. 568–572.
14. Andrews, N.C., Erdjument-Bromage, H., Davidson, M.B., Tempst, P., and Orkin, S.H., *Nature*, 1993, vol. 362, pp. 722–728.
15. Andrews, N.C., Kotkow, K.J., Ney, P.A., Erdjument-Bromage, H., Tempst, P., and Orkin, S.H., *Proc. Natl. Acad. Sci. USA*, 1993, vol. 90, pp. 11488–11492.
16. Hai, T. and Curran, T., *Proc. Natl. Acad. Sci. USA*, 1991, vol. 88, pp. 3720–3724.
17. Yamamoto, K.K., Gonzalez, G.A., Biggs III, E.H., and Montminy, M.R., *Nature*, 1988, vol. 334, p. 494.
18. Rehfuss, R.P., Walton, K.M., Loriaux, M.M., and Goodman, R.H., *J. Biol. Chem.*, 1991, vol. 266, pp. 18431–18434.
19. De Groot, R.P., Den Hertog, J., Vandenheede, J.R., Goris, J., and Sassone-Corsi, P., *EMBO J.*, 1993 vol. 12, pp. 3903–3911.
20. Foulkes, N.S., Borrelli, E., and Sassone-Corsi, P., *Cell*, 1991, vol. 64, pp. 739–749.
21. Foulkes, N.S., Mellström, B., Benusiglio, E., and Sassone-Corsi, P., *Nature*, 1992, vol. 355, pp. 80–84.
22. Johnson, P.F., Landschulz, W.H., Graves, B.J., and McKnight, S.L., *Genes Dev.*, 1987, vol. 1, pp. 133–146.
23. Landschulz, W.H., Johnson, P.F., Adashi, E.Y., Graves, B.J., and McKnight, S.L., *Genes Dev.*, 1988, vol. 2, pp. 786–800.
24. Vinson, C.R., Sigler, P.B., and McKnight, S.L., *Science*, 1989, vol. 246, pp. 911–916.
25. Vinson, C.R., Hai, T., and Boyd, S.M., *Genes Dev.*, 1993, vol. 7, pp. 1047–1058.
26. Vallejo, M., Ron, D., Miller, C.P., and Habener, J.F., *Proc. Natl. Acad. Sci. USA*, 1993, vol. 90, pp. 4679–4683.
27. Mueller, C.R., Maire, P., and Schibler, U., *Cell*, 1990, vol. 61, pp. 279–291.
28. Donald, R.G.K., Schindler, U., Batschauer, A., and Cashmore, A.R., *EMBO J.*, 1990, vol. 9, pp. 1727–1735.
29. Huang, Z.J., Edery, I., and Rosbash, M., *Nature*, 1993, vol. 364, pp. 259–262.
30. Mermod, N., O'Neill, E.A., Kelly, T.J., and Tjian, R., *Cell*, 1989, vol. 58, pp. 741–753.
31. Santoro, C., Mermod, N., Andrews, P.C., and Tjian, R., *Nature*, 1988, vol. 334, p. 218.
32. Rupp, R.A.W., Kruse, U., Multhaup, G., Göbel, U., Beyreuther, K., and Sippel, A.E., *Nucleic Acids Res.*, 1990, vol. 18, pp. 2607–2616.
33. Kruse, U., Qian, F., and Sippel, A.E., *Nucleic Acids Res.*, 1991, vol. 19, p. 6641.

Kruse, U. and Sippel, A.E., *J. Mol. Biol.*, 1995, vol. 278, pp. 860–865.

Reith, W., Ucla, C., Barras, E., Gaud, A., Durand, B., Herrero-Sanchez, C., Kobr, M., and Mach, B., *Mol. Cell. Biol.*, 1994, vol. 14, pp. 1230–1244.

Reith, W., Herrero-Sanchez, C., Kobr, M., Silacci, P., Berte, C., Barras, E., Fey, S., and Mach, B., *Genes Dev.*, 1990, vol. 4, pp. 1528–1540.

Chodosh, L.A., Baldwin, A.S., Carthew, R.W., and Sharp, P.A., *Cell*, 1988, vol. 53, pp. 11–24.

Hatamochi, A., Golumbek, P.T., van Schaftingen, E., and de Crombrugghe, B., *J. Biol. Chem.*, 1988, vol. 263, p. 5940.

Dorn, A., Durand, B., Marfing, C., Le Meur, M., Benoist, C., and Mathis, D., *Proc. Natl. Acad. Sci. USA*, 1987, vol. 84, p. 6249.

Lim, L.C., Swendeman, S.L., and Sheffery, M., *Mol. Cell. Biol.*, 1992, vol. 12, pp. 828–835.

Jane, S.M., Nienhuis, A.W., and Cunningham, J.M., *EMBO J.*, 1995, vol. 14, pp. 97–105.

Tafuri, S.R. and Wolffe, A.P., *Proc. Natl. Acad. Sci. USA*, 1990, vol. 87, pp. 9028–9032.

Wolffe, A.P., Tafuri, S.R., Ranjan, M., and Familari, M., *New Biologist*, 1992, vol. 4, pp. 250–298.

Hanas, J.S., Hazuda, D.J., Bogenhagen, D.F., Wu, F.Y.H., and Wu, C.W., *J. Biol. Chem.*, 1983, vol. 258, p. 14120.

Miller, J., McLachlan, A.D., and Klug, A., *EMBO J.*, 1985, vol. 4, p. 1609.

Krust, A., Green, S., Argos, P., Kumar, V., Walter, P., Bornert, J.M., and Chambon, P., *EMBO J.*, 1986, vol. 5, pp. 891–897.

Kumar, V., Green, S., Staub, A., and Chambon, P., *EMBO J.*, 1986, vol. 5, pp. 2231–2236.

Severne, Y., Wieland, S., Schaffner, W., and Rusconi, S., *EMBO J.*, 1988, vol. 7, pp. 2503–2508.

Luisi, B.F., Xu, W.X., Otwinowski, Z., Freedman, L.P., Yamamoto, K.R., and Sigler, P.B., *Nature*, 1991, vol. 352, pp. 497–505.

Bresnick, E.H., Dalman, F.C., Sanchez, E.R., and Pratt, W.B., *J. Biol. Chem.*, 1989, vol. 264, pp. 4992–4997.

Pratt, W.B., *J. Biol. Chem.*, 1993, vol. 268, pp. 21455–21458.

Dalman, F.C., Koenig, R.J., Perdew, G.H., Massa, E., and Pratt, W.B., *J. Biol. Chem.*, 1990, vol. 265, pp. 3615–3618.

Evans, T., Reitman, M., and Felsenfeld, G., *Proc. Natl. Acad. Sci. USA*, 1988, vol. 85, pp. 5976–5980.

Evans, T. and Felsenfeld, G., *Cell*, 1989, vol. 58, pp. 877–885.

Omichinski, J.G., Clore, G.M., Schaad, O., Felsenfeld, G., Trainor, C., Appella, E., Stahl, S.J., and Gronenborn, A.M., *Science*, 1993, vol. 261, pp. 438–445.

Pan, T. and Coleman, J.E., *Proc. Natl. Acad. Sci. USA*, 1990, vol. 87, pp. 2077–2081.

Pan, T. and Coleman, J.E., *Biochemistry*, 1991, vol. 30, pp. 4212–4222.

58. Gardner, K.H., Pan, T., Narula, S., Rivera, E., and Coleman, J.E., *Biochemistry*, 1991, vol. 30, pp. 11292–11302.
59. Kraulis, P.J., Raine, A.R.C., Gadhavi, P.L., and Laue, E.D., *Nature*, 1992, vol. 356, pp. 448–450.
60. Marmorstein, R., Carey, M., Ptashne, M., and Harrison, S.C., *Nature*, 1992, vol. 356, pp. 408–414.
61. Pabo, C.O. and Sauer, R.T., *Annu. Rev. Biochem.*, 1984, vol. 53, pp. 293–321.
62. Laughon, A. and Scott, M.P., *Nature*, 1984, vol. 310, pp. 25–31.
63. Shepherd, J.C.W., McGinnis, W., Carrasco, A.E., de Robertis, E.M., and Gehring, W.J., *Nature*, 1984, vol. 310, pp. 70–71.
64. TenHarmsel, A., Austin, R.J., Savenelli, N., and Biggin, M.D., *Mol. Cell. Biol.*, 1993, vol. 13, pp. 2742–2752.
65. Walter, J., Dever, C.A., and Biggin, M.D., *Genes Dev.*, 1994, vol. 8, pp. 1678–1692.
66. Assa-Munt, N., Mostishire-Smith, R.J., Aurora, R., and Herr, W., *Cell*, 1993, vol. 73, pp. 193–205.
67. Klemm, J.D., Rould, M.A., Aurora, R., Herr, W., and Pabo, C.O., *Cell*, 1994, vol. 77, pp. 21–32.
68. Feuerstein, R., Wang, X., Song, D., Cooke, N.E., and Liebhaber, S.A., *Proc. Natl. Acad. Sci. USA*, 1994, vol. 91, pp. 10655–10659.
69. Schmeichel, K.L. and Beckerle, M.C., *Cell*, 1994, vol. 79, pp. 211–219.
70. Bürglin, T.R., *Guidebook to the Homeobox Genes*, Duboule, D., Ed., London: Oxford University Press, 1994, pp. 25–71.
71. He, X., Treacy, M.N., Simmons, D.M., Ingraham, H.A., Swanson, L.W., and Rosenfeld, M.G., *Nature*, 1989, vol. 340, pp. 35–42.
72. Jürgens, G. and Weigel, D., *Roux's Arch. Dev. Biol.*, 1988, vol. 197, pp. 335–337.
73. Weigel, D. and Jaeckle, H., *Cell*, 1990, vol. 63, pp. 455–456.
74. Clark, K.L., Halay, E.D., Lai, E., and Burley, S.K., *Nature*, 1993, vol. 364, pp. 412–420.
75. Sorger, P.K. and Nelson, H.C.M., *Cell*, 1989, vol. 59, pp. 807–813.
76. Harrison, C.J., Bohm, A.A., and Nelson, H.C.M., *Science*, 1994, vol. 263, pp. 224–227.
77. Ogata, K., Hojo, H., Aimoto, S., Nakai, T., Nakamura, H., Sarai, A., Ishii, S., and Nishimura, Y., *Proc. Natl. Acad. Sci. USA*, 1992, vol. 89, pp. 6428–6432.
78. Jamin, N., Gabrielsen, O.S., Gilles, N., Lirsac, P.-N., and Toma, F., *Eur. J. Biochem.*, 1993, vol. 216, pp. 147–154.
79. Liang, H., Olejniczak, E.T., Mao, X., Nettesheim, D.G., Yu, L., Thompson, C.B., and Fesik, S.W., *Proc. Natl. Acad. Sci. USA*, 1994, vol. 91, pp. 11655–11659.
80. Bürglin, T.R., *Cell*, 1991, vol. 66, pp. 11–12.
81. Ghosh, G., van Duyne, G., Ghosh, S., and Sigler, P.B., *Nature*, 1995, vol. 373, pp. 303–310.
82. Müller, C.W., Rey, F.A., Sodeoka, M., Verdine, G.L., and Harrison, S.C., *Nature*, 1995, vol. 373, pp. 311–317.
83. Cho, Y., Gorina, S., Jeffrey, P.D., and Pavletich, N.P., *Science*, 1994, vol. 265, pp. 346–355.

84. Nikolov, D.B., Hu, S.-H., Lin, J., Gasch, A., Hoffmann, A., Horikoshi, M., Chua, N.-H., Roeder, R.G., and Burley, S.K., *Nature*, 1992, vol. 360, pp. 40–46.
85. Kim, Y., Geiger, J.H., and Hahn, S., *Nature*, 1993, vol. 365, pp. 512–520.
86. Kim, J.L., Nikolov, D.B., and Burley, S.K., *Nature*, 1993, vol. 365, pp. 520–527.
87. Werner, M.H., Huth, J.R., Gronenborn, A.M., and Clore, G.M., *Cell*, 1995, vol. 81, pp. 705–714.
88. Love, J.J., Li, X., Case, D.A., Giese, K., Grosschedl, R., and Wright, P.E., *Nature*, 1995, vol. 376, pp. 791–795.
89. Pellegrini, L., Tan, S., and Richmond, T.J., *Nature*, 1995, vol. 376, pp. 490–498.
90. Hegde, R.S., Grossman, S.R., Laimins, L.A., and Sigler, P.B., *Nature*, 1992, vol. 359, pp. 505–512.
91. Verma, I.M., Stevenson, J.K., Schwarz, E.M., van Antwerp, D., and Miyamoto, S., *Genes Dev.*, 1995, vol. 9, pp. 2723–2735.
92. Inoue, J.-i., Kerr, L.-D., Kakizuka, A., and Verma, I.M., *Cell*, 1992, vol. 68, pp. 1109–1120.
93. Kel, O.V., Romaschenko, A.G., Kel, A.E., Wingender, E., and Kolchanov, N.A., *Nucleic Acids Res.*, 1995, vol. 23, pp. 4097–4103.
94. Jain, J., McCaffrey, P.G., Miner, Z., Kerppola, T.K., Lambert, J.N., Verdine, G.L., Curran, T., and Rao, A., *Nature*, 1993, vol. 365, pp. 352–355.
95. Brunvand, M.W., Schmidt, A., and Siebenlist, U., *J. Biol. Chem.*, 1988, vol. 263, pp. 18904–18910.
96. Shaw, J.-P., Utz, P.J., Durand, D.B., Toole, J.J., Emmel, E.A., and Crabtree, G.R., *Science*, 1988, vol. 241, pp. 202–205.
97. Boise, L., Petryniak, B., Mao, X., June, C.H., Wang, C.-Y., Lindsten, T., Bravo, R., Kovary, K., Leiden, J., and Thompson, C.B., *Mol. Cell. Biol.*, 1993, vol. 13, pp. 1911–1919.
98. Bochkarev, A., Barwell, J.A., Pfuetzner, R.A., Furey Jr., W., Edwards, A.M., and Frappier, L., *Cell*, 1995, vol. 83, pp. 39–46.
99. Ferrari, S., Harley, V. R., Pontiggia, A., Goodfellow, P.N., Lovell-Badge, R., and Bianchi, M.E., *EMBO J.*, 1992, vol. 11, pp. 4497–4506.
100. Carlsson, P., Waterman, M.L., and Jones, K.A., *Genes Dev.*, 1993, vol. 7, pp. 2418–2430.
101. Thanos, D. and Maniatis, T., *Cell*, 1992, vol. 71, pp. 777–789.