Evolution of bHLH Transcription Factors: Modular Evolution by Domain Shuffling?

Burkhard Morgenstern[*1] *and William R. Atchley*[†]

[*]GSF, National Research Center for Environment and Health, Institute of Biomathematics and Biometry, Neuherberg, Germany; and [†]Department of Genetics, North Carolina State University

Multidomain proteins usually contain several conserved and apparently independently evolved domains. As a result, classifications based on only a single small domain may obscure the true evolutionary relationships of the proteins. The current classification of basic helix-loop-helix (bHLH) domain-containing proteins is based on the conserved bHLH domain alone. Herein, we explore whether sequence homology and, therefore, evolutionary relationships can be detected among the flanking or non-bHLH components of the amino acid sequences of 122 bHLH proteins. These 122 proteins were the same proteins previously used to construct the existing classification of the bHLH-domain-containing proteins. Several possible scenarios are examined in order to explain the observed patterns of sequence divergence, including (1) monophyly, (2) convergent evolution, (3) addition of functional components to the bHLH domain, and (4) modular evolution with domain shuffling. Drawing on several lines of evidence, we suggest that modular evolution by domain shuffling may have played an important role in the evolution of this large group of transcriptional regulators.

Introduction

It is well established that proteins are often complex entities containing multiple, independent, and separately evolved domains (e.g., Campbell and Baron 1991; Doolittle and Bork 1993; Campbell and Downing 1994; Doolittle 1995; Hawkins and Lamb 1995). Indeed, multiple-domain proteins can be viewed as combinatorial arrangements of autonomously structured modules (Sonnhammer and Kahn 1994; Hegyi and Bork 1997).

Functionally heterogeneous proteins are often classified together because they share one or more small conserved domains, such as those involved with DNA binding or oligomerization (Lewin 1997). Lumping together heterogeneous groups of proteins based on small shared conserved domains presents special problems, because the resultant mosaic distribution of conserved domains makes it difficult to accurately estimate evolutionary history. A group of proteins might be considered a monophyletic group and derived from common ancestry if they share a small homologous domain. However, the issue of evolutionary relationships among the proteins may become clouded when these proteins share several other highly conserved domains. In the latter case, each separate conserved domain may (1) have an independent evolutionary history and (2) occur in other seemingly unrelated groups of proteins that do not contain the first conserved domain.

A good case in point is the basic helix-loop-helix (bHLH) family of proteins, which is a structurally complex and functionally heterogeneous group. Currently, over 400 bHLH-domain-containing proteins are known. These proteins act as transcriptional regulators and are involved with neurogenesis, myogenesis, cell proliferation, tissue differentiation, and other essential developmental processes. These proteins are grouped together based on the common possession of a small conserved bipartite domain containing approximately 60 amino acids involved with DNA binding and protein oligomerization. The basic (b) component of the bHLH domain includes a short component of mainly basic residues that bind to a consensus hexanucleotide ‘‘E-box’’ (CANNTG) (Voronova and Baltimore 1990). The helix-loop-helix (HLH) component is a highly hydrophobic oligomerization region of approximately 50 residues producing two amphipathic alpha-helices separated by a variable-length loop.

The bHLH-domain-containing proteins are structurally heterogeneous in that they contain several highly conserved domains (fig. 1 and table 1). In addition to the basic DNA-binding and helix-loop-helix dimerization components, various groups of these proteins may also contain leucine zippers (Atchley and Fitch 1997) or PAS domains (Zelnar, Wappner, and Shilo 1997). Leucine zippers are found in many other groups of unrelated proteins, and some PAS proteins are known that do not contain the bHLH domain, e.g., Period in *Drosophila* (Huang, Edery, and Rosbash 1993).

In addition to the mosaic distribution of these various domains among proteins, the relative placement of the bHLH domain can vary significantly (fig. 1). The bHLH domain can be located at the COOH end of the protein (as in the Myc proteins), at the NH end (as in Sim), or in an intermediate position (MyoD). As can be seen in figure 1, this relative variability in position of the bHLH domain in the protein causes serious problems in attempts to align the remainder of the protein sequence.

Atchley and Fitch (1997) recently produced an evolutionary classification of the bHLH domain involving almost 400 different proteins. These authors found 27 evolutionary lineages or clades that represented groups of functionally similar proteins. A clade here refers to a monophyletic collection of proteins that is statistically

[1] Present address: Rhone-Poulenc Rorer, Limited, Dagenham, Essex, England.

Key words: modular evolution, molecular evolution, basic helix-loop-helix domain, transcription factors, sequence alignment, phylogeny.

Address for correspondence and reprints: William R. Atchley, Department of Genetics, North Carolina State University Raleigh, North Carolina 27695-7614. E-mail: atchley@ncsu.edu.

Mol. Biol. Evol. 16(12):1654–1663. 1999

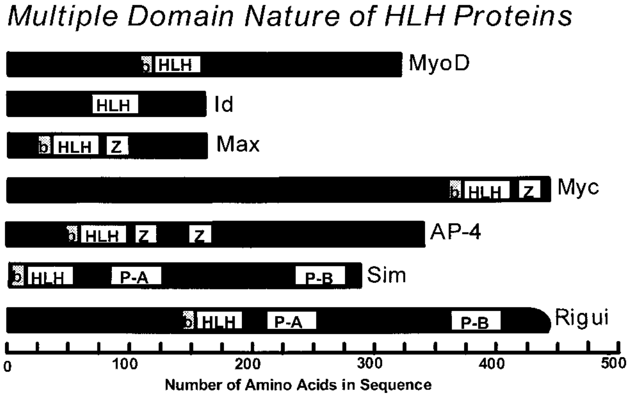

FIG. 1.—A simplified figure showing the multiple-domain nature of HLH proteins. The domain compositions and spatial positions of these various components can vary considerably. The *b* block reflects the basic DNA-binding component, *HLH* is the helix-loop-helix component, *Z* is a leucine zipper domain, and *P-A* and *P-B* refer to PAS-A and PAS-B domains, respectively. The Rigui protein is actually over 1,200 residues long.

strongly supported in phylogenetic analyses. These various clades could be classified into four major groups based on the basic DNA-binding patterns. Group A proteins bind the hexanucleotide CAGCTG E-box and include such proteins as Lyl, Twist, dHand, Achaete-scute, Atonal, MyoD, and E12. Group B proteins bind the CACGTG E-box and include Hairy, Srebp, Tfe, Myc, Mad, Mxi1, Usf, Cbf1, and Esc. Some of the group B proteins (e.g., Myc, Max, USF, Srebp, Mad, and Mxi1) also contain a leucine zipper domain between the bHLH domain and the COOH end of the protein. Group C proteins, including Sim, Trh, and Ahr, have an atypical basic region and contain a pair of PAS repeats that are involved with dimerization with other PAS-containing proteins (Swanson, Chan, and Bradfield 1995). Finally, group D proteins do not have the basic DNA-binding region and act as dominant negative regulators of other bHLH proteins. Included in group D are Id and Emc. Groups A, B, C, and D are well defined in the phylogenetic trees reported by Atchley and Fitch (1997) and have other distinguishing sequence attributes (Atchley, Terhalle, and Dress 1999).

In their phylogenetic analyses of the bHLH domain, Atchley and Fitch (1997, fig. 2), collapsed to a single line those nodes of the neighbor-joining (NJ) tree (Saitou and Nei 1987) where bootstrap values were <50% based on 500 trials. Their results could be depicted as a small ''forest'' where 27 small trees or clades arose from a horizontal surface. The 27 bHLH clades and their included proteins are described in table 1 of the present paper. Each of these individual lineages or clades had strong statistical support (usually a bootstrap value of >95%), implying that this particular clade was present in almost all of the 500 bootstrap trials for these data. These clades usually indicated groups of bHLH-

Table 1
Protein Families or Clades of bHLH-Domain-Containing Proteins

Clades	Included Proteins	E-Box Groups	Domains Present	Function
AC-S	[*ase*]; [*ac, sc, l'sc*]; [*mash, ash*]	A	Basic	Neurogenesis; determination of neuronal precursors
Atonal	[*atonal*] [*lin-32*] *math1, neuroD*	A	Basic	Neurogenesis
Delilah	*delilah*	A	Basic	Differentiation of epidermal cells into muscle
dHand	*dhand, ehand, hxt, hed*	A	Basic	Cardiac morphogenesis; trophoblast cell development
E12/Da	[*e12, e47, tfe2, pan2, me2*]; [*itf, pan1*]; [*da*]	A	Basic	Neurogenesis; sex determination; regulation of myogenesis
Hen	*hen, helhel*	A	Basic	Neurogenesis
Lyl	[*lyl*]; [*scl, nscl*]; [*tal*]	A	Basic	Haematopoietic proliferation and differentiation
Myod	*myod1, myogenin, myf5, myf6*	A	Basic	Myogenesis
Nex	*nex-1, rat4,*	A	Basic	Neurogenesis
Twist	[*twist, dermo*]; [*paraxis, scleraxis*]	A	Basic	Specification of mesoderm lineages
Arnt	*arnt*	B	Basic	Regulation of aryl hydrocarbon receptor activity
Cbf	*cbf-1*	B	Basic	Centromeric binding and chromosomal segregation
Esc	*escl*	B	Basic	Sexual differentiation in yeast
G-Box	*G-Box*	B	Basic	
Hairy	[*hes*]; [*hlhm, hairy, deadpan, e(spl)*]	B	Basic	Neurogenesis; segmentation
Mad	*mad, mxi1*	B	Basic, Z	Regulation of cell proliferation
Myc	[*c-myc, n-myc, l-myc*]; [*max*]	B	Basic, Z	Cell proliferation, differentiation; oncogenesis
No	*ino2, ino4*	B	Basic	Phospholipid synthesis
PHO4	*pho4, nuc1*	B	Basic	Phosphate regulation in yeast
R	[*r*]; [*delila*]	B	Basic	Regulation of anthcynanin pigmentation
SREBP	*srebp, add1, hlh106*	B	Basic, Z	Sterol synthesis; adipocyte determination
Tfe	*tfe3, tfeb, mi*	B	Basic, Z	Activates transcription in immunoglobulin heavy chain enhancer
Usf	[*usf, namalwa*]; [*spf1*]	B	Basic, Z	Upstream stimulation factor; insulin enhancer
Sim	[*sim, trh*]; [*ahr*]	C	PAS	CNS midline lineage regulation; tracheal cell induction
Id	[*id, heira, hlh462*]; [*emc*]	D	—[a]	Negative inhibition of DNA binding; myogenesis; neurogenesis
CENBPR . . .	*cenbpr*	?	Basic	Centromeric binding protein
Ap-4	*ap-4*	?	Basic, Z	Enhancement of viral and cellular gene activation

NOTE.—The elements of the table include proteins known to be included in these clades, groupings based on E-box binding patterns, presence of other conserved domains, and the primary functions of the proteins. Groupings within a given clade (=subclades) from the neighbor-joining tree reported by Atchley and Fitch (1997) are denoted with brackets.

[a] The Id proteins have no basic region.

domain-containing proteins involved in a particular developmental process or function.

The analyses of Atchley and Fitch (1997) based on the bHLH domain provided good resolution of the evolutionary relationships among proteins at the terminal nodes (tips of the branches of the tree). However, these analyses did not resolve the deep node structure of the tree. The deep nodes did not have strong statistical support, and the hypothesis of a single evolutionary origin for the bHLH domain could not be rigorously supported. Consequently, several important questions about the evolution of the entire protein need to be resolved.

Herein, we inquire if differentiation of the bHLH proteins has followed a model of modular evolution by domain shuffling. Specifically, we inquire whether a classification based on the non-bHLH components of the proteins (hereinafter referred to as the flanking regions) is concordant with one based on the bHLH domain alone. Resolution of this question depends on whether the flanking regions of the complete proteins contain significant regions of sequence homology. We have shown that the entire protein sequence can be accurately aligned within families (=clades *sensu* Atchley and Fitch 1997) of bHLH proteins like MyoD (Atchley, Fitch, and Bronner-Fraser 1994) and Myc (Atchley and Fitch 1995). However, the extent of sequence similarity between clades for the flanking regions has been difficult to resolve.

Thus, differential similarity among clades of proteins with regard to their bHLH domains might be due to one of the following reason: (1) monophyly where differential change has occurred in the conserved bHLH and nonconserved flanking regions; (2) functional or mechanistic convergence (Doolittle 1994); (3) a domain-only hypothesis, whereby the ancestral proteins contain little besides the bHLH domain and the additional sequence components were added later; and (4) modular evolution where the observed patterns have occurred as the result of domain insertion and rearrangement (Doolittle and Bork 1993; Hawkins and Lamb 1995; Patthy 1996; Li 1997). Herein, we discriminate among these alternative hypotheses by detailed analyses of sequence similarity patterns in the flanking or non-bHLH elements of the proteins.

Methods and Materials

We used the same set of 122 bHLH proteins employed by Atchley and Fitch (1997) in their analyses of the bHLH domain. Definitions of the various evolutionary lineages or clades within the bHLH domain, their higher group classifications, and their possession of other conserved domains are given in table 1. More complete definitions of the various proteins and their functions can be found in Atchley and Fitch (1997) and Atchley, Terhalle, and Dress (1999).

To determine the extent of sequence similarity and, correspondingly, the amount of phylogenetic signal in the flanking regions of the sequences, we excised the conserved basic HLH domains from the complete amino acid sequences and concatenated the remaining two components of the sequence for further analysis. The low degree of similarity among clades in the flanking regions of the sequences prevented use of the usual global alignment procedures like CLUSTAL W (Thompson, Higgins, and Gibson 1994). Consequently, we used DIALIGN, which is a novel algorithm for local multiple sequence alignment (Morgenstern 1999). This alignment program is available on the World Wide Web at http://bibiserv.techfak.uni-bielefeld.de/dialign/.

Most alignment methods rely on an algorithm proposed by Needleman and Wunsch (1970) which compares single residues and imposes gap penalties. In this case, the score of an alignment is defined as the sum of substitution values of aligned residue pairs minus a penalty for every gap introduced into the sequences. Standard alignment methods then try to find alignments with optimal scores in the sense of this definition. These methods produce meaningful alignments provided that the sequences are globally related.

However, if sequences share only isolated regions of local similarity and are otherwise unrelated, alignment methods relying on the Needleman-Wunsch scoring scheme often fail to produce biologically correct alignments. DIALIGN, on the other hand, relies on comparison of whole segments of the sequences. Alignments are composed of gap-free segment pairs (called diagonals). Every diagonal D is assigned a weight score $w(D)$ reflecting the similarity among the two respective segments. These weights are calculated based on probabilistic considerations (see Morgenstern et al. [1998] and Morgenstern [1999] for details). The program then tries to construct an alignment as a consistent collection of diagonals with a maximum sum of weights.

In short, we call a set of diagonals consistent if an alignment exists in which all segment pairs are matched (see fig. 2). Gaps are not penalized in this approach. They correspond to parts of the sequences not belonging to any of the selected diagonals. This strategy enables DIALIGN to find local similarities in otherwise unrelated sequences that cannot be detected by standard methods for global alignment (Morgenstern, Dress, and Werner 1996; Morgenstern et al. 1998).

Hierarchical Clustering from Flanking Regions

As expected, there are groups of sequences within our data set such that the sequences within these groups are more closely related to each other than to sequences from outside these groups. In order to identify such groups, we used two hierarchical clustering algorithms, called ‘‘minimum-linkage clustering’’ and ‘‘maximum-linkage clustering.’’

For each pair of sequences s_i and s_j, we determined a pairwise similarity score $S_{i,j}$, defined as the sum of weights of the diagonals connecting these sequences in the DIALIGN multiple alignment. For example, in the situation shown in figure 2, one would have $S_{1,2} = w(D_1) + w(D_2)$, since sequences s_1 and s_2 are connected by diagonals D_1 and D_2. Correspondingly, we have $S_{1,3} = w(D_3) + w(D_4)$ and $S_{2,3} = w(D_5)$.

Given these similarity scores, we formed clusters of sequences by first joining the two sequences s_i and s_j with the highest similarity score $S_{\mathrm{i,j}}$ and then succes-

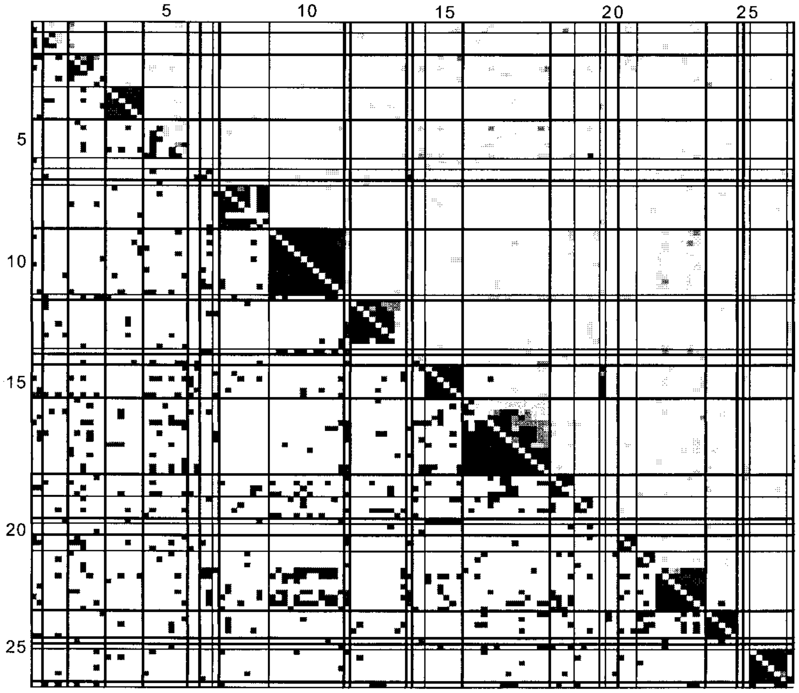

FIG. 2.—Results of the sequence-shuffling procedures. The upper right triangular half of the matrix contains coded values for the numbers of standard deviations by which the respective pairwise alignment scores exceed the mean DIALIGN alignment scores of the random sequences generated from these sequences. The results are coded as light gray (>4 SD), medium gray (>7 SD), and dark gray (>10 SD) values. The lower triangular half reports sequence pairs for which the alignment scores were higher than all of the scores of the random sequences generated from the original sequences. The clades are coded numerically along the edges of the matrix. The codes are as follows: 1 = Hen; 2 = Lyl; 3 = Twist; 4 = dHand; 5 = AC-S; 6 = Atonal; 7 = Nex; 8 = Delilah; 9 = MyoD; 10 = E12-Da; 11 = Ap-4; 12 = Id; 13 = CENBPR; 14 = Pho4; 15 = Sim; 16 = Hairy; 17 = SREBP; 18 = Tfe; 19 = Arnt; 20 = No; 21 = Mad; 22 = Myc; 23 = Usf; 24 = Cbf; 25 = Esc; 26 = R; 27 = G-Box.

sively joining the remaining sequences and clusters of sequences in the order of decreasing similarity scores. Here, for the minimum-linkage method, the similarity among two clusters of sequences, C_1 and C_2, is defined as the minimum similarity score of all pairs of sequences s_i and s_j, one each from clusters C_1 and C_2, respectively. With the maximum-linkage method, the maximum of these similarity scores is used. We refer to the resulting clustering schemes as DIALIGN similarity (D/S) clustering. Note that these clustering methods are naturally quite rough. Yet, in the present context, they have the advantage of being invariant with respect to even the most nonlinear transformations of the table of similarities, provided the transformations are performed using a strictly monotonously increasing function and, hence, preserve the order. Moreover, since the minimum linkage and maximum linkage are using opposite criteria for the hierarchical clustering of the sequences, one may have some confidence in those groups of sequences that appear as clusters in both approaches. It is clear that branch lengths in the treelike structures obtained by the D/S method do not directly reflect time in evolution. They roughly represent the degree of (dis)similarity among groups. The clusters produced by minimum linkage and maximum linkage were compared with the clades of the previously published NJ tree based on the conserved bHLH domain (Atchley and Fitch 1997).

The D/S trees were rooted using human centromere protein B (CENBPR) as the outgroup. This latter protein was initially thought to show homology with the HLH family of proteins (Sullivan 1991). However, more recently it has been suggested that it contains a helix-turn-helix domain (Iwahara et al. 1998). Consequently, inclusion of CENBPR as the outgroup may help to resolve questions about whether the bHLH-domain- containing proteins are a monophyletic group. The maximum-linkage tree for the flanking regions of these proteins from the DIALIGN procedure is included in this paper.

Sequence Shuffling

To quantify the degree of similarity among the flanking regions, we applied the following sequence-shuffling procedure: For each of the (122 × 121)/2 = 7,381 pairs of sequences, we generated 50 pairs of random sequences exhibiting the same amino acid composition as the non-bHLH parts of the original sequences. This was done by randomly shuffling sites within sequences. Each of the 50 pairs of random sequences was aligned by DIALIGN, and the resulting DIALIGN alignment scores were compared with the alignment scores of the original sequences. (Note that, unlike in standard approaches, in the segment-to-segment approach to sequence alignment, the ‘‘score’’ of an alignment is defined as the sum of weights of aligned segment pairs, or diagonals [see Morgenstern et al. 1998]. For example, the score of the alignment in fig. 3 would be the sum of weights of the five diagonals of which it is composed.) Only 50 pairs of sequences were gener-

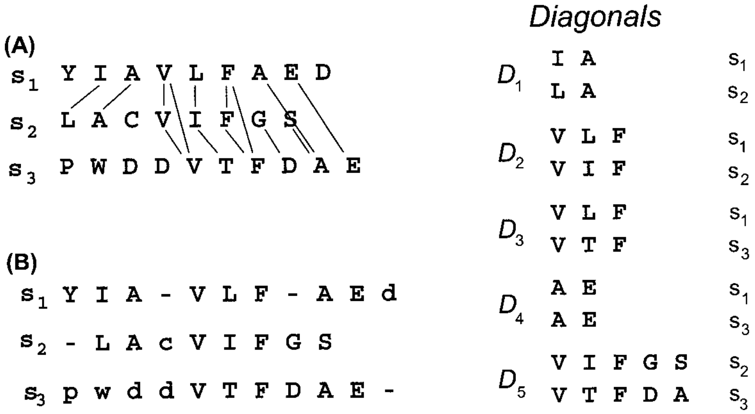

FIG. 3.—Construction of alignments with the segment-to-segment approach (DIALIGN). An alignment is constructed as a consistent collection of gap-free segment pairs (so-called ''diagonals'') (*A*). Here, consistency means that it is possible to introduce gaps into the sequences such that all segment pairs are matched (*B*). Every possible diagonal *D* is given a weight *w*(*D*) reflecting the degree of similarity between the two segments. The program then tries to find a consistent collection of diagonals with the maximum sum of weights. See Morgenstern (1999) for details.

ated, because this procedure is highly computer-intensive and very time-consuming (analyses of these 7,381 × 50 = 369,050 pairwise alignments took about a month to perform on a SPARC Ultra II computer).

For each pair of original sequences, quantitative information was derived from the shuffled alignments in two distinct ways: First, we determined the number of standard deviations by which the original alignment score exceeded the average alignment score of the corresponding pairs of random sequences. Second, we determined the number of pairs of generated random sequences that had higher alignment scores than the original sequences. For a number $p > 0$, we defined a set of sequences C to be a p-clique if for any two sequences contained in C, the original similarity score is at least p standard deviations greater than the average score of the shuffled random sequences. A p-clique is called maximal if it is not contained in any larger p-clique. The structure of these cliques is complicated. For example, we cannot expect the maximal p-cliques to be disjoint. We will investigate the structure of the p-cliques from the flanking regions of the bHLH sequences in a forthcoming paper. In this study, we compared the maximal 2-cliques to the clades of the NJ tree constructed from the conserved domain. The relatively low p value of 2 was chosen since we applied the restrictive condition that in a clique, all pairs of sequences must have scores that are at least p standard deviations greater than the average score of the random sequences.

Results

Concordance in Clade Composition

When applied to the flanking regions, the D/S clustering and shuffling procedure gave results similar to those obtained using an NJ tree of the bHLH domain alone (figs. 4 and 5). In figure 4, instances of change in clade composition are indicated in bold type. Approximately the same groups of sequences and the same group compositions were found for the bHLH domains and the non-bHLH regions for these 122 proteins. The most conspicuous difference between the flanking regions and conserved bHLH was that some clades in the bHLH analyses were split into two groups for the flanking regions. These groupings are evident in the NJ analyses of Atchley and Fitch (1997), but these authors chose to formally recognize them as separate lineages within a clade rather than as distinct clades.

The three primary examples of clades that are split apart in D/S clusters are the Twist, AC-S (Achaete-scute), and Myc clades (fig. 4 and 5). The Twist clade in the NJ analyses includes Paraxis and Scleraxis, but these two proteins separated out into a group of their own in the D/S procedures. Similarly, with AC-S, the vertebrate homologs of the Achaete-scute proteins (i.e., Mash and Ash) are placed in a separate lineage. In Myc, the Max proteins are broken out from Myc, as seen in the NJ tree, and the Mad and Mxi1 proteins are now shown to be more similar to Myc. The HES3 proteins are split out from the other Hairy group proteins, and MyoD in worms is quite distinct from other MyoD proteins.

Furthermore, the D/S clusters in figure 4 show that (1) the Lyl clade was altered with regard to the placement of Tal; (2) Ase is excluded from the AC-S clade; (3) the two elements within the Atonal, Pho4, and No clusters were separated out into separate and single lineages; (4) the protein Arnt is added to the Sim cluster; (5) G-Box was added to the R cluster; and (6) Emc has been excluded from Id. In many of these instances, the inclusion of some of these proteins into a given clade was rather tenuous in the original NJ analyses of the bHLH domain, e.g., the original inclusion of Emc with Id.

In all other instances, clades from bHLH NJ coincide with clades from D/S clustering and cliques from the shuffling procedure. The shuffling procedure gives results for the flanking regions that are very similar to those described for the NJ tree of the bHLH domain. For each sequence pair, the upper triangular half of figure 2 contains the number of standard deviations by

bHLH Domain		Flanking Regions					
Clade	NJ Tree	Clade	Minimum Linkage	Clade	Maximum Linkage	Clade	Maximal 2-Cliques
HEN	[2 – hen, helhel]	HEN	**[hen]**	HEN	**[hen]**	HEN	[2 – hen, helhel]
		HELHEL	**[helhel]**	HELHEL	**[helhel]**		
LYL	[2 – lyl]; [scl]; [tal]	LYL	**[2 – lyl];[scl]**	LYL	**[2 – lyl]**	LYL	**[2 – lyl];[scl]**
		TAL	**[tal]**	TAL	**[tal]**	TAL	**[tal]**
				SCL	**[scl]**		
TWIST	[4 – twist, dermo]; [2 – paraxis, scleraxis]	TWIST	**[4 – twist, dermo]**	TWIST	**[3 – twist, dermo]**[1]	TWIST	**[4 – twist, dermo]**
		PARAXIS	**[2 –paraxis, scleraxis]**	PARAXIS	**[2 – paraxis, scleraxis]**	PARAXIS	**[2 – paraxis, scleraxis]**
dHAND	[6 – dhand, ehand, hxt, hed]	dHAND	[6 – dhand, ehand, hxt, hed]	dHAND	[6 – dhand, ehand, hxt, hed]	dHAND	[6 – dhand, ehand, hxt, hed]
AC-S	[ase]; [3 – ac, sc, lsc]; [3 – mash, ash]	AC-S	**[ase]; [2 – ac,lsc]**	AC-S	**[3 – ac,sc,lsc]**	AC-S	**[3 – ac,sc,lsc]** **[3 – mash,ash]**
		MASH	**[3 – mash,ash]**	MASH	**[2 – mash]**		
						ASE	**[ase]**
ATONAL	[2 – atonal, lin-32]	ATONAL	**[atonal]**	ATONAL	**[atonal]**	ATONAL	**[atonal]**
		LIN	**[lin-32]**	LIN	**[lin-32]**	LIN	**[lin-32]**
NEX	[2 – nex-1, rat4]	NEX	[2 – nex-1, rat4]	NEX	[2 – nex-1, rat4]	NEX	[2 – nex-1, rat4]
MYOD	[8 – myod1, myogenin, myf5, myf6]	MYOD	**[7 – myod1, myogenin, myf5, myf6]**[2]	MYOD	**[7 – myod1, myogenin, myf5, myf6]**[2]	MYOD	**[7 – myod1, myogenin, myf5, myf6]**[2]
E12/Da	[9 – e12, e47, tfe2, pan2, me2]; [2 – itf, pan1]; [da]	E12/Da	[9 – e12, e47, tfe2, pan2, me2]; [2 – itf, pan1]; [da]	E12/Da	[9 – e12, e47, tfe2, pan2, me2]; [2 – itf, pan1]; [da]	E12/Da	[9 – e12, e47, tfe2, pan2, me2]; [2 – itf, pan1]; [da]
ID	[8 – id, heira, hlh462]; [emc]	ID	**[8 – id, heira, hlh462]**	ID	**[8 – id, heira, hlh462]**	ID	**[8 – id, heira, hlh462]**
		EMC	**[emc]**	EMC	**[emc]**	EMC	**[emc]**
PH04	[2 – ph04, nuc1]	PH04	[2 – ph04, nuc1]	PH04	[2 – ph04, nuc1]	PH04	[2 – ph04, nuc1]
SIM	[4 – sim, trh]; [2 – ah]	SIM	[4 – sim,trh]; [2 – ah];	SIM	**[4 – sim,trh]; [2 – ah]; [arnt]**	SIM	**[4 – sim,trh]; [2 – ah]; [arnt]**
HAIRY	[14 – hes, hlhm, hairy, deadpan, e(spl)]	HAIRY	[14 – hes, hlhm, hairy, deadpan, e(spl)]	HAIRY	**[12 – hes**[3]**, hlhm, hairy, deadpan, e(spl)]**	HAIRY	**[13 – hes, hlhm, hairy**[4]**, deadpan, e(spl)]**
				HES3	[2 – hes3]		
SREBP	[4 – srebp1, srebp2, add1, hlh106]	SREBP	**[3 – srebp1, add1, hlh106]**	SREBP	[4 – srebp1, srebp2, add1, hlh106]	SREBP	[4 – srebp1, srebp2, add1, hlh106]
TFE	[3 – tfe3, tfeb, mi]	TFE	[3 – tfe3, tfeb, mi]	TFE	[3 – tfe3, tfeb, mi]	TFE	[3 – tfe3, tfeb, mi]
NO	[ino2]; [ino4]	NO2	**[ino2]**	NO2	**[ino2]**	NO2	**[ino2]**
		NO4	**[ino4]**	NO4	**[ino4]**	NO4	**[ino4]**
MAD	[3 – mad, mxi1]	MAD	[3 – mad, mxi1]	MAD	[3 – mad, mxi1]	MAD	[3 – mad, mxi1]
MYC	[8 – c-myc, n-myc, l-myc]; [3 – max]	MYC	**[8 – c-myc, n-myc, l-myc]**	MYC	**[8 – c-myc, n-myc, l-myc]**	MYC	**[8 – c-myc, n-myc, l-myc]**
		MAX	**[3 – max]**	MAX	**[3 – max]**	MAX	**[3 – max]**
USF	[5 – usf, namalwa, spf1]	USF	[5 – usf, namalwa, spf1]	USF	[5 – usf, namalwa, spf1	USF	[5 – usf, namalwa, spf1]
R	[6 – R, delila]	R	**[6 – R, delila]; [G-Box]**	R	[6 – R, delila]	R	**[6 – R, delila]; [G-Box]**
G-BOX	[G-Box]			G-BOX	[G-Box]		

[1] Without TWIST-Fly
[2] Without MYOD-Worm
[3] Without hes3
[4] Without HAIRY-beetle

FIG. 4.—Results of (i) the neighbor- joining (NJ) phylogenetic analyses of the bHLH domain, (ii) the minimum- and maximum-linkage approaches for the flanking regions, and (iii) the maximum 2-cliques method. In each category, the clade names are given in the left column, while the right column indicates the subclades (with brackets) and the number of sequences within each subclade. Deviations of analyses of the flanking regions from the NJ analyses of the bHLH domain are indicated in bold type.

which the respective pairwise alignment score exceeds the mean alignment score of the random sequences generated from these sequences. The results are color-coded: light gray represents values of >4 SD, medium gray represents values of >7 SD, and dark gray represents values of >10 SD. The lower triangular half reports sequence pairs for which the alignment scores were higher than all of the scores of the random sequences generated from the original sequences. With regard to the upper portion of figure 2, there were only six instances in which large associations between clades (>10 SD) occurred. These involved the following sequence pairs: (1) CENBPR with Rat4-kw8; (2) Asct5-Fly with Tfeb-Human; (3) Hairy-Fly with Sim-Fly; (4) Arnt-Human with Ahr-Human, Sim-Fly, Sim1-Mouse, Sim2-Mouse, and Trh-Fly; (5) Usf2-Mouse with Myc-Chicken; and (6) R-Maize with G-Box PVU-Bean.

The most obvious associations that deviate from the results based on the bHLH domain alone involve the Arnt-Human sequence with Ahr, Sim, and Trh. On the basis of the bHLH domain, Atchley and Fitch (1997) placed the Arnt proteins in group B based on the sequence in the DNA-binding region, while Sim, Ahr, and Trh were included in group C. Furthermore, Antosonnsson et al. (1995) have shown experimentally that Arnt constitutively and specifically recognizes group B E-box target sequences. However, all of these proteins contain the PAS-A and PAS-B domains, which are highly conserved components containing approximately 200–300 residues. Consequently, this is an instance in which the inclusion of other conserved domains significantly influences the classification based on the bHLH domain.

We removed the two PAS domains, as well as the intervening sequences, and repeated the randomization experiments with the remaining parts of the sequences. Sim and Arnt still formed a cluster, i.e., significant similarity among these particular sequences was not confined to the conserved bHLH and PAS domains.

Loss of Information about Group Structure

Based on the phylogenetic analyses of the bHLH domain, almost all sequences in a database of over 400 proteins could be classified into four groups (Atchley and Fitch 1997). This grouping reflects how the proteins interacted with the consensus hexanucleotide E-box. A conspicuous difference between these NJ results for the bHLH domain and the D/S and shuffling results for the flanking regions is that this grouping into groups A, B, C, and D is lost. Instead, in the D/S clusters, the various clades previously classified into the groups are now intermixed. Thus, group C proteins like Sim and Ahr now cluster with the group B protein Arnt. All of these proteins contain PAS domains, and the group C proteins interact with Arnt (Crews 1998). They are quite distinct with regard to the basic component of the bHLH domain but show significant similarity in other parts of the protein.

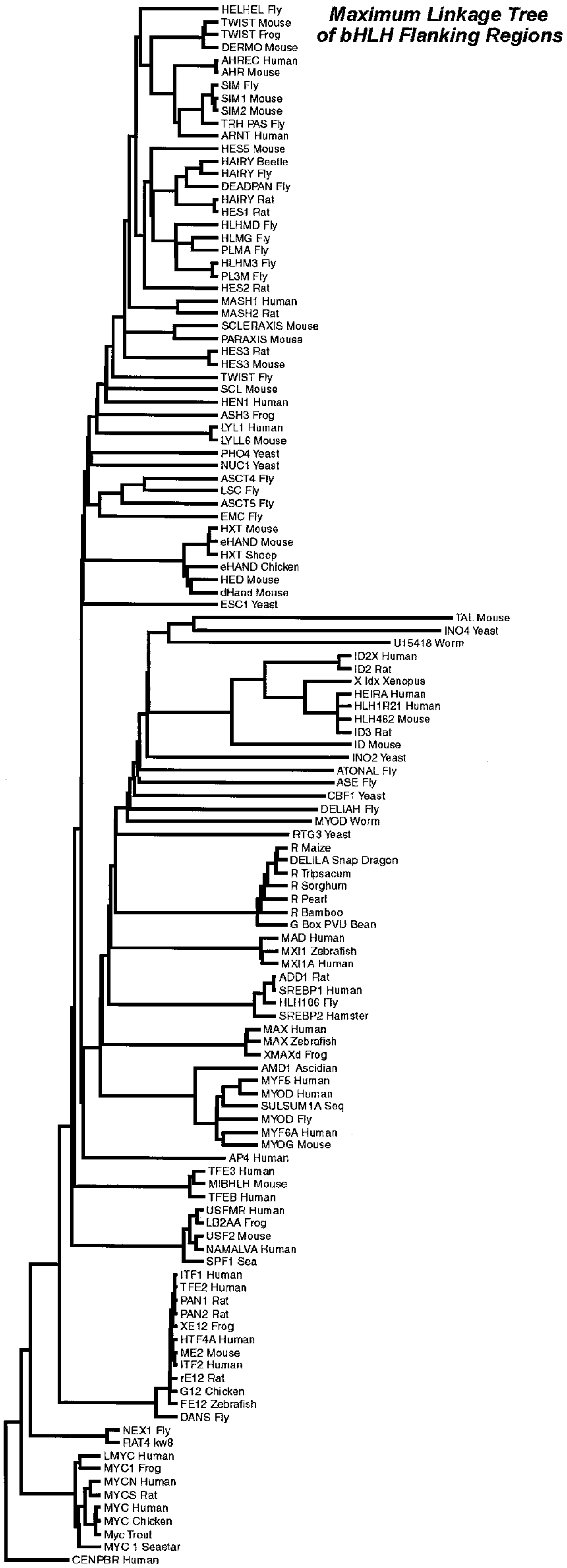

FIG. 5.—Maximum-linkage tree of the flanking regions for 122 bHLH-domain-containing proteins.

Discussion

Two obvious conclusions can be drawn from these analyses. First, clades based on the bHLH domain flanking regions usually contain the same proteins as clades derived from the bHLH motif alone. Thus, within clades, not only the bHLH domains, but also the flanking sequences seem to be evolutionarily related. This suggests that proteins contained within the same clade, as defined by bHLH domain similarity, are derived from a common ancestor. This conclusion is supported by more detailed phylogenetic analyses of sequences within clades. For example, Atchley, Fitch, and Bronner-Fraser (1994) examined the evolution of entire sequences in 29 proteins of the MyoD family, while Atchley and Fitch (1997) described the evolution of 45 Myc and Max proteins. In each instance, an ancestral protein in primitive organisms gave rise to the present-day distribution of proteins through a series of paralogous and orthologous events. Unpublished data from phylogenetic analyses of other clades of bHLH-domain-containing proteins such as Hairy exhibit divergence patterns similar to that of MyoD and Myc.

The second conclusion is that rather extensive regions of homology can be identified within the bHLH domain among the various clades. However, the same cannot be said for the flanking regions. In most instances, the flanking region sequences of different clades are so different as to be virtually random with respect to each other.

Scenarios for bHLH Protein Evolution

These findings suggest at least four possible explanations for the observed diversity among bHLH proteins. The first is the ‘‘monophyly’’ scenario, in which all proteins containing the bHLH domain were derived from a common ancestor. Over geological time, strong stabilizing selection has maintained the bHLH domain in all of these proteins so as to preserve the DNA-binding and dimerization capabilities. This facilitates transcriptional regulation in a functionally divergent collection of proteins. During the same time, the flanking regions of the relevant protein diverged as a result of diversifying selection acting on the other functional aspects of the various protein groups. A limited version of this scenario was demonstrated by Atchley and Fitch (1997) when they showed significantly different rates of sequence evolution between the bHLH domain and its flanking regions in Myc and Max proteins.

The second possible scenario is the ‘‘convergent-evolution’’ scenario, in which parts of sequences exhibit sequence similarity because functionally similar domains arose independently. Such a convergent evolution scenario seems a feasible hypothesis for certain simple motifs. For example, leucine zippers characterized by heptad repeats of leucine residues may have arisen independently in different protein lineages as a result of convergence (Brendel and Karlin 1989). Indeed, in the analyses of the flanking regions reported here, there is no systematic clustering of proteins based on the presence of the leucine zipper domain, suggesting that the leucine zipper may have arisen more than once in the bHLH-domain-containing proteins.

Crystal structure studies on six divergent members of the bHLH-domain-containing proteins argue for a common origin rather than convergent evolution for the bHLH domain. Several authors, including Ferre-D'Amare et al. (1993, 1994), Ellenberger et al. (1994), Ma et al. (1994), Shimizu et al. (1997), and Parraga et al. (1998), have examined the higher-order structure of representative bHLH proteins. The crystal structure of the Max protein homodimer, for example, is a parallel, left-handed, four-helix bundle, and hydrophobic residues from helix-1 and helix-2 are at the core of this globular domain, where they pack together and exhibit strong van der Waals interactions that stabilize the structure of the homodimer (Ferre-D'Amare et al. 1993). This crystal structure in Max appears to be quite similar to that found in other divergent bHLH proteins, such as USF (Ferre-D'Amare et al. 1994), MyoD (Ma et al. 1994), E47 (Ellenberger et al. 1994), PHO4 (Shimizu et al. 1997), and SREBP (Parraga et al. 1998). The considerable similarity in the complex structures of this domain in divergent lineages suggests that the structurally complex bHLH domain had a common origin and did not arise through convergence.

The third possible scenario is the ''domain-only'' scenario. The ancestral bHLH protein could have been a small protein composed of little more than the bHLH or the HLH domain alone. Subsequent evolution involved addition of other protein elements to this core. There are several small bHLH proteins that presently consist of this domain and little else. The proteins Id and Emc consist of little more than the HLH domain (e.g., 128 residues for Id in zebrafish). Max, Mad, and Mxi1 are very short proteins (e.g., 160 residues for Max in humans) containing little more than the bHLH and leucine zipper domains. These proteins function primarily as negative regulators of MyoD (as in the case of Id) or as dimerization partners of Myc (as in the case of Max).

Alternatively, these short proteins could have arisen by loss of all but the binding and dimerization function. For example, the Max protein seems to have heterodimerization (and regulation of expression) with Myc as a major function. Hence, one might envision a scenario in which much of an ancestral protein was lost to provide a protein whose primary function is only DNA binding and dimerization (like Max) or negative regulation (like Id and Emc).

The fourth possible scenario is a ''modular-evolution'' scenario in which the mosaic pattern of evolutionary lineages has occurred by domain shuffling. Domain shuffling is generally thought to be of two types, i.e., domain duplication and domain insertion (or loss) (Li 1997). A domain has been defined as a structurally defined sequence that folds spontaneously into a characteristic shape under a defined set of circumstances (Doolittle and Bork 1993). Many such domains have been shown to be able to move within and between proteins during evolution (de Chateau and Bjorck 1995; Doolittle 1995; Ikeo, Takahashi, and Gojobori 1995; Patthy 1996). The domain-only and modular-evolution scenarios are not necessarily separate and independent mechanisms. We list them separately so as to contrast different domain evolution mechanisms.

Arguments in Favor of a Modular Evolution Explanation

An unequivocal test does not exist to determine which of these four scenarios is most appropriate to explain the observed diversity of proteins containing the bHLH domain. However, several lines of evidence do support a modular evolution by domain shuffling. These arguments include:

1. *Spatial variation among domains.* There is significant variation in the spatial positions of the bHLH, leucine zipper, and PAS domains (but not different sequential ordering) in different clades (fig. 1). This is particularly evident in the Myc and Sim families, in which the bHLH domain occurs at the amino and carboxyl ends of the proteins, respectively. This spatial variation in the various domains brings into serious question the idea that the bHLH-domain- containing proteins have a common origin. To produce this level of spatial variation in position of the bHLH domain, any derivatives of a common ancestral protein would have had to move the bHLH domain from the amino end to the carboxyl end (or vice versa). Consequently, this suggests that the variation in placement of the bHLH domain within these diverse proteins results from domain shuffling.
2. *Lack of sequence similarity.* While there is considerable sequence similarity in the bHLH domains of different clades, there is very little, if any, similarity in the flanking regions or the non-bHLH domains. This strongly argues that the flanking regions and the bHLH domains have had separate and possibly independent evolutionary histories. That is, the bHLH domains are related, but flanking regions either are unrelated or diverged so long ago that any evidence of common ancestry in the flanking portions of the sequences has been lost.
3. *Component loss.* At least two major groups (i.e., the Id and Emc proteins) have lost (or never had) the basic DNA-binding component of the bHLH domain. Similarly, the Period (Per) proteins in *Drosophila* and mammalian species (Citri et al. 1987; Huang, Edery, and Rosbash 1993) contain the PAS domain (as found in bHLH proteins Ahr, Arnt, Sim. and Trh). However, the Per proteins do not have the bHLH domain. In this regard, the Per proteins are similar to the group D proteins Id and Emc, which have the HLH domain but not the basic region (Atchley and

Fitch 1997). This suggests that components of the bHLH domain and other domains have been gained or lost during evolution.

4. *Mosaic patterns of various domains in bHLH proteins.* Some bHLH proteins possess other highly conserved domains, like leucine zipper and PAS domains. The leucine zipper domain occurs in a number of bHLH-domain-containing proteins, but certainly not in all of them. Its presence in some lineages but not others (Atchley and Fitch 1997), together with its widespread distribution in other unrelated transcription factors, argues for independent evolution of this domain. The two PAS domains (PAS-A and PAS-B) found in the Sim, Ahr, and Arnt protein families may reflect possible degenerate direct repeats arising from domain duplication (Crews, Thomas, and Goodman 1988; Hawkins and Lamb 1995). This suggests that highly conserved domains in some of these proteins have been duplicated or added.

Mosaic Evolutionary Patterns in Modular Evolution

The impact of a mosaic distribution of independently evolved domains on the classification of proteins can be seen in those bHLH proteins that contain PAS domains. The bHLH/PAS proteins are placed in the same clade based on the flanking sequences, but in different clades based on the bHLH domain. This is best understood by consideration of two well-known PAS-domain-containing proteins, i.e., the dioxin receptor Ahr (aryl hydrocarbon receptor) and its dimerization partner Arnt (aryl hydrocarbon receptor nuclear translocator). The basic DNA-binding region in Arnt resembles those of group B bHLH proteins, which bind to the palindromic sequence 5′-CACGTG-3′/3′-GTGCAC-5′). In contrast, the equivalent region in Ahr is atypical and bears relatively little resemblance to the basic DNA-binding region of other bHLH proteins (Swanson, Chan, and Bradfield 1995; Dong, Ma, and Whitlock 1996). Ahr has been placed in a separate group (group C) along with the proteins trachealess and single-minded (Atchley and Fitch 1997).

Conclusions

The phylogenetic analyses reported here find little or no sequence similarity in the flanking (non-bHLH) regions among clades that were defined by the bHLH domain. Generally, the same clade composition is found in both the bHLH and the non-bHLH components of these proteins, suggesting that the clades themselves represent both functional and evolutionary lineages. Our interpretation is that the proteins composing each of the various clades had a common ancestor. However, our analyses found no statistically well supported deep node structure for the non-bHLH domain components of the proteins; indeed, these analyses show little, if any, similarity among clades in the non-bHLH components. This finding strongly suggests either that the various clades are not related or that they diverged so long ago that any evidence of common ancestry has been lost. Thus, aside from sequence similarity exhibited by the bHLH domain, we find little evidence of evolutionary relationship among these various clades.

The absence of sequence similarity in non-bHLH components, the extensive variation in the location of the bHLH domain, the mosaic pattern of inclusion of various other functional domains, and other considerations suggest that modular evolution of these various domains may have been an important component in evolution of these proteins. Under a modular-evolution hypothesis, the bHLH domain and other domains were shuffled during evolution so as to generate the polyphyletic assemblage of proteins that we now recognize as the HLH family of transcriptional regulators.

Acknowledgments

This work was done while B.M. was a visiting scientist at North Carolina State University. We would like to thank the GSF Research Center for their generous support for his stay at NCSU. We are indebted to our colleagues Walter Fitch, Andreas Dress, Jeffrey Thorne, Ronald Sederoff, and Michael Purugganan for constructive comments on the manuscript. We thank James Rosinski and Kurt Wollenberg for many helpful comments and discussions on this project. The research was supported by grants to W.R.A. from the National Institutes of Health (GM45344) and the National Science Foundation (INT-9603452).

LITERATURE CITED

ANTOSONNSSON, C., V. ARULAMPALAM, M. L. WHITLAW, S. PETTERSSON, and K. POELLINGER. 1995. Constitutive function of the basic helix-loop-helix/PAS factor Arnt. Regulation of target promoters via the E box motif. J. Biol. Chem. **270**:13968–13972.

ATCHLEY, W. R., and W. M. FITCH. 1995. Myc and Max: molecular evolution of a family of proto-oncogene products and their dimerization partner. Proc. Natl. Acad. Sci. USA **92**:10217–10221.

ATCHLEY, W. R., and W. M. FITCH. 1997. A natural classification of the basic helix-loop-helix class of transcription factors. Proc. Natl. Acad. Sci. USA **94**:5172–5176.

ATCHLEY, W. R., W. M. FITCH, and M. BRONNER-FRASER. 1994. Molecular evolution of the MyoD family of transcription factors. Proc. Natl. Acad. Sci. USA **91**:11522–11526.

ATCHLEY, W. R., W. TERHALLE, and A. DRESS. 1999. Positional dependence, cliques and predictive motifs in the bHLH protein domain. J. Mol. Evol. **48**:501–516.

BRENDEL, V., and S. KARLIN. 1989. Too many leucine zippers? Nature **341**:574–575.

CAMPBELL, I. D., and M. BARON. 1991. The structure and function of protein modules. Philos. Trans. R. Soc. Lond. B **332**:165–170.

CAMPBELL, I. D., and A. K. DOWNING. 1994. Building protein structure and function from modular units. Trends Biotechnol. **12**:168–172.

CITRI, Y., H. V. COLOT, A. C. JACQUIER, Q. YU, J. C. HALL, D. BALTIMORE, and M. ROSBASH. 1987. A family of unusually spliced biologically active transcripts encoded by a Drosophila clock gene. Nature **326**:42–47.

CREWS, S. T. 1998. Control of cell lineage-specific development and transcription by bHLH-PAS proteins. Genes Dev. **12**:607–620.

Crews, S. T., J. B. Thomas, and C. S. Goodman. 1988. The Drosophila single-minded gene encodes a nuclear protein with sequence similarity to the per gene product. Cell **52**: 143–151.

de Chateau, M., and L. Bjorck. 1994. Protein PAB, a mosaic albumin-binding bacterial protein representing the first contemporary example of module shuffling. J. Biol. Chem. **269**:12147–12151.

Deng, V. C., C. Dolde, M. L. Gillison, and G. J. Kato. 1992. Discrimination between related DNA sites by a single amino acid residue of Myc-related basic-helix-loop-helix proteins. Proc. Natl. Acad. Sci. USA **89**:599–602.

Dong, L., Q. Ma, and J. P. Whitlock Jr. 1966. DNA binding by the heterodimeric Ah receptor. J. Biol. Chem. **271**:7942–7948.

Doolittle, R. F. 1994. Convergent evolution: the need to be explicit. Trends Biochem. **19**:15–18.

———. 1995. The multiplicity of domains in proteins. Annu. Rev. Biochem. **64**:287–314.

Doolittle, R. F., and P. Bork. 1993. Evolutionarily mobile modules in proteins. Sci. Am. **269**:50–56.

Ellenberger, T., D. Fass, M. Arnaud, and S. C. Harrison. 1994. Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. Genes Dev. **8**:970–980.

Ferre-D'Amare, A. R., P. Pognonec, R. G. Roeder, and S. K. Burley. 1994. Structure and function of the b/HLH/Z domain of USF. EMBO J. **13**:180–189.

Ferre-D'Amare, A. R., G. C. Prendergast, E. B. Ziff, and S. K. Burley. 1993. Recognition by Max of it cognate DNA through a dimeric b/HLH/Z domain. Nature **363**:38–45.

Hawkins, A. R., and H. K. I. Lamb. 1995. The molecular biology of multidomain proteins: selected examples. Eur. J. Biochem. **232**:7–18.

Hegyi, H., and P. Bork. 1997. On the classification of evolution of protein modules. J. Protein Chem. **16**:545–551.

Huang, Z. J., I. Edery, and M. Rosbash. 1993. PAS is a dimerization domain common to Drosophila Period and several transcription factors. Nature **364**:259–262.

Ikeo, K., K. Takahashi, and T. Gojobori. 1995. Different evolutionary histories of kringle and protease domains in serine proteases: a typical example of domain evolution. J. Mol. Evol. **40**:331–336.

Iwahara, J., R. Kigawa, K. Kitagawa, H. Masumoto, T. Okazaki, and S. Yokoyama. 1998. A helix-turn-helix structure unit in human centromere protein B (CENP-B). EMBO J. **17**:827–837.

Lewin, B. 1997. Genes VI. Oxford University Press, New York.

Li, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

Ma, P. C. M., R. A. Rould, H. Weintraub, and C. O. Pabo. 1994. Crystal structure of MyoD bHLH domain_DNA complex: perspectives on DNA recognition and implications for transcriptional activation. Cell **77**:451–459.

Morgenstern, B. 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics **15**:211–218.

Morgenstern, B., W. R. Atchley, K. Hahn, and A. Dress. 1998. Segment-based scores for pairwise and multiple sequence alignments. Pp. 115–121 *in* J. Glasgow, R. Littlejohn, F. Major, R. Lathrop, D. Sankoff, and C. Sensen, eds. Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, Calif.

Morgenstern, B., A. Dress, and T. Werner. 1996. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc. Natl. Acad. Sci. USA **93**: 12098–12103.

Needleman, S., and C. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol Biol. **48**:443–453.

Parraga, A., L. Bellsolell, A. R. Ferre-D'Amare, and S. K. Burley. 1998. Co-crystal structure of sterol regulatory element binding protein 1a at 2.3 A resolution. Structure **6**: 661–672.

Patthy, L. 1996. Exon shuffling and other ways of module exchange. Matrix Biol. **15**:301–310.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Shimizu, T., A. Toumoto, K. Ihara, M. Shimizu, Y. Kyogoku, N. Ogawa, Y. Oshima, and T. Hakoshima. 1997. Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. EMBO J. **16**:4689–4697.

Sonnhammer, E. L. L., and D. Kahn. 1994. Modular arrangement of proteins as inferred from analysis of homology. Protein Sci. **3**:482–492.

Sullivan, K. F. 1991. CENP-B is a highly conserved protein with homology to the helix-loop-helix family of proteins. Chromosoma **100**:360–370.

Swanson, H. I., W. K. Chan, and C. A. Bradfield. 1995. DNA binding specificities and pairing rules of the Ah receptor, ARNT and SIM proteins. J. Biol. Chem. **270**:26292–26302.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Voronova, A., and D. Baltimore. 1990. Mutations that disrupt DNA binding and dimer formation in the E47 helix-loop-helix protein map to distinct domains. Proc. Natl. Acad. Sci. USA **87**:4722–4726.

Zelnar, E., P. Wappner, and B.-Z. Shilo. 1997. The PAS domain confers target gene specificity of Drosophila bHLH/PAS proteins. Genes Dev. **11**:2079–2089.

Marcy K. Uyenoyama, reviewing editor

Accepted August 5, 1999