Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 5172–5176, May 1997
Evolution

A natural classification of the basic helix–loop–helix class of transcription factors

WILLIAM R. ATCHLEY*† AND WALTER M. FITCH‡

*Center for Quantitative Genetics, Department of Genetics, North Carolina State University, Raleigh, NC 27695-7614; and ‡Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92717

Contributed by Walter M. Fitch, January 21, 1997

ABSTRACT **A natural (evolutionary) classification is provided for 242 basic helix–loop–helix (bHLH) motif-containing proteins. Phylogenetic analyses of amino acid sequences describe the patterns of evolutionary change within the motif and delimit evolutionary lineages. These evolutionary lineages represent well known functional groups of proteins and can be further arranged into five groups based on binding to DNA at the hexanucleotide E-box, the amino acid patterns in other components of the motif, and the presence/absence of a leucine zipper. The hypothesized ancestral amino acid sequence for the bHLH transcription factor family is given together with the ancestral sequences of the subgroups. It is suggested that bHLH proteins containing a leucine zipper are not a natural, monophyletic group.**

Transcription factors belonging to the helix-loop-helix family are important regulatory components in transcriptional networks of many developmental pathways (1–3). They are involved in regulation of neurogenesis, myogenesis, cell proliferation and differentiation, cell lineage determination, sex determination, and other essential processes in organisms ranging from yeast to mammals.

Helix-loop-helix proteins are characterized by common possession of highly conserved bipartite domains for DNA binding and protein–protein interaction (1–2). A motif of mainly basic residues permits helix-loop-helix proteins to bind to a consensus hexanucleotide E-box (CANNTG) (4). A second motif of primarily hydrophobic residues referred to as the helix-loop-helix domain allows these proteins to interact and to form homo- and/or heterodimers (2). The dimerization motif contains about 50 amino acids and produces two amphipathic α-helices separated by a loop of variable length. Additionally, some basic helix–loop–helix (bHLH) proteins contain a leucine zipper (LZ) dimerization motif characterized by heptad repeats of leucines that occur immediately C-terminal to the bHLH motif.

The bHLH motif was first identified in murine transcription factors E12 and E47 (1). Subsequent descriptions of numerous new bHLH proteins have made it increasingly difficult to understand their interrelationships, and a natural classification scheme is badly needed to bring order to this large and important group of proteins. A "natural" classification is an evolutionary one based on rules of descent and uses sequence data and information about function and is predictive with regard to new information. Herein, we provide an evolutionary classification of the bHLH motif based on 242 distinct amino acid sequences. The final phylogenetic analyses described here are based on a subset of 122 divergent sequences. The selection of a subset of 122 sequences was based primarily on extent of sequence similarity (sequences that were very similar were culled). The bHLH sequences were aligned using the Clustal-W algorithm (5) and were improved by eye. Only the bHLH motif was used in these analyses because the flanking sequences for proteins from independent clades are either nonhomologous or are so diverged that the alignments are meaningless. Fig. 1 provides a representative group of aligned protein sequences showing the limits of the motif components and a numbering scheme for the amino acids suggested by Ferre–D'Amare *et al.* (6).

A neighbor-joining (NJ) tree (7, 8) was constructed for these 122 sequences based on the p-distance (fraction of sites that differ) (Fig. 2). Gapped sites were omitted from pair-wise distance estimations, and the resultant trees were boot strapped 500 times to provide information about statistical reliability of a branch length estimate. Boot strap values give the probability that a given branch length is greater than 0. To simplify interpretation of such a large and complex phylogenetic tree, nodes were collapsed to a single horizontal line when statistical support for a particular node was $<35\%$. In instances in which branches were collapsed, the lengths of the branches removed were added to the descending branches of the lower node that will disappear so that the distance from a clade to the common ancestor of all the clades descended from the same point remained unchanged. This process increases by the length of the removed branch, the distance between any two sequences, one each from the two sister taxa whose cenancestor was the node that was removed. Other thresholds for the boot strap value can be selected. Presentation of such a collapsed tree facilitates recognition of independent evolutionary lineages and inhibits speculation about statistically unsupported nodes. Using the NJ tree as a starting point, a maximum parsimony procedure estimated ancestral sequences.

Major Patterns of Divergence. The NJ tree (Fig. 2) describes the major patterns of bHLH motif sequence divergence. The NJ tree indicates at least 24 separate major paralogous lineages representing major functional groups of bHLH proteins. These 24 clades are then nested into four higher order groups that correspond to DNA binding patterns. If the tree is collapsed using a critical boot strap value of 50%, the results differ very little from those obtained with a critical value of 35% as used in Fig. 2. Table 1 gives additional information about each clade, some of the included proteins, a brief generalization about function, and groupings based on function, DNA binding, and presence/absence of LZ motifs.

The diverse array of functionally different proteins found in plants, yeast, and animals indicates that the bHLH motif is evolutionarily ancient. Indeed, gene duplications associated with these numerous, paralogous lineages apparently occurred very early in the history of the motif.

The bHLH motif has good resolving power to delimit families of proteins and describe their evolutionary relationships at the tips of the clades. The deep branching structure shows how these separate families of proteins are interrelated. Unfortunately, small boot strap values at the deep nodes indicate that the bHLH

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

0027-8424/97/945172-5$2.00/0
PNAS is available online at **http://www.pnas.org**.

Abbreviations: bHLH, basic helix–loop–helix; LZ, leucine zipper; NJ, neighbor-joining.

†To whom reprint requests should be addressed. e-mail: atchley@ncsu.edu.

Clade	Phenotype	Group	Aligned Sequence
			BBBBBBBBBBBBB HHHHHHHHHHHHHHH LLLLLLLLLLLLLLLLLLLL HHHHHHHHHHHHHHH
			0000000001111 111111222222222 23333333333444444444 555555555566666
			1234567890123 4567890123456789 90123456789012345678 9012345678901234
bHLH	Ancestor		KRKXHOXXEXXRR XXXNXXXXELXSLXP ----------SXXB KXAILKXXXXYXXAX
Group B	Ancestor		KRKTHNXMEKKRR BXXNXXXXELXSLXP ----------SKXB KXAILKXXVXYXRAL
Group A	Ancestor		RRXTANAXERRRM KBINEGFXKLXXLIP ----------XKLX KXXILQQAVEYIXSL
Group C	Ancestor		KGXTKNXXERNRR DXXNXEXDZLASLLP ----------SKLD KXSILRLXVXYLRAX
Group D	Ancestor		QRXPAXZXJXXRL NDINEXYXKLKJLVP ----------RKLX KXEILQHVIDYIXDL
LYL	LYL1	A	RRVFTNSRERWRQ QNVNGAFAELRKLLP T-HPPD----------RKLS KNEVLRLAMKYIGFL
TWIST	SCLERAX	A	QRHTANARERDRT NSVNTAFTALRTLIP TERPND----------KLS KIETLRLASSYISHL
DHAND	dHAND	A	RRGTANRKERRRT QSINSAFAELRECIP N-VPAD----------TKLS KIKTLRLATSYIAYL
HEN	HELHEL	A	YRTAHATREGIRV EAFNVSFADVRKLLP T-LPPD----------KKLS KIEILKLAICYIAYL
ACS	ASCT5	A	RR---NARERNRV KQVNNGFSQLRQHIP AAVIADLSNGRRGIGPNKKLS KVSTLKMAVEYIRRL
ATONAL	ATONAL	A	RRLAANARERRRM QNLNQAFDRLRQYLP C-L----------GNDRQLS KHETLQMAQTYISAL
MYOD	MYOGENIN	A	RRRAATLREKRRL KKVNEAFEALKRSTL L----------NPNQRLP KVEILRSAIQYIERL
E12	PAN2	A	RRVANNARERLRV RDINEAFKELGRMCQ LHLSTE----------KPQT KLLILHQAVAVILSL
E12	DANS	A	RRQANNARERIRI RDINEALKELGRMCM THLKSD----------KPQT KLGILNMAVEVIMTL
AP4	AP4	?	RREIANSNERRRM QSINAGFQSLKTLIP HTDGE----------KLS KAAILQQTAEYIFSL
ID	ID	D	PALLDEQQVNVLL YDMNGCYSRLKELVP T-LPQN----------RKVS KVEILQHVIDYIRDL
AH	SIM1	C	KEKSKNA-ARTRR EKENSEFYELAKLLP --LPSA---------ITSQLD KASIIRLTTSYLK-M
HAIRY	HAIRY	B	RKSSKPIMEKRRR ARINESLSQLKTLIL DALKKDSSR-------HSKLE KADILEMTVNHLRNL
SREBP	ADD1	B	KRTAHNAIEKRYR SSINDKIVELKDLVV G----------TEAKLN KSAVLRKAIDYIRFL
TFE	TFEB	B	KKDNHNLIERRRR FNINDRIKELGMLIP KAND----------LDVRWN KGTILKASVDYIRRM
NO	INO2	B	RKWKHVQMEKIRR INTKEAFERLIKSVR T----------PPKENGKRI PKHILLTCVMNDIKS
MAD	MAD	B	SRSTHNEMEKNRR AHLRLCLEKLKGLVP L-GPES----------SRHT TLSLLTKAKLHIKKL
MYC	MAX	B	KRAHHNALERKRR DHIKDSFHSLRDSVP S-LQGE----------KKAS RAQILDKATEYIQYM
MYC	MYC	B	KRRTHNVLERQRR NELKRSFFALRDQIP E-LENN----------EKAP KVVILKKATAYILSV
USF	USF2	B	RRAQHNEVERRRR DKINNWIVQLSKIIP DCH---------ADNSKTGAS KGGILSKACDYIREL
CBF	CBF1	B	RKDSHKEVERRRR ENINTAINVLSDLLP ----------VRESS KAAILARAAEYIQKL
ESC	ESC1	B	LRTSHKLAERKRR KEIKELFDDLKDALP LDKT----------TKSS KWGLLTRAIQYIEQL
GBOX	G-Box	B	EPLNHVEAERQRR EKLNQRFYALRAVVP N----------VSKMD KASLLGDAISYINEL
R	R	B	KN--HVMSERKRR EKLNEMFLVLKSLLP S-IH----------RVN KASILAETIAYLKEL

For hypothetical ancestor, X = unknown (ambigious) residue, J = (D/E), O = (N/K), B = (D/N), Z = (E,Q)

FIG. 1. Fig. 1. Aligned bHLH motif for representative sequences for major evolutionary lineages in NJ tree in Fig. 2. Designation of basic (B), helix (H), and loop (L) regions and the numbering sequence for the individual amino acids follows Ferre–D'Amare (6). Ancestor of various groups reflects sequence inferred to be hypothetical ancestral sequence from maximum parsimony analysis. Clades are in order, as in Fig. 2.

motif alone has low information content for understanding the very early evolutionary history of the motif. The order of the relevant duplications in the evolution of these proteins is hidden toward the root of the tree by collapsing of branches and nodes in this region of the tree. Lack of statistical support probably arises because (*i*) the bHLH motif comprises only a small portion

Table 1. Helix-loop-helix transcription factors: Protein families, functions, and motifs

Protein families	Included proteins	Groupings: E-box	Groupings: Murre *et al.*	Groupings: LZ	Function
AC-S	ac, sc, ase, l'sc, mash, ash	A	II		Neurogenesis; determination of neuronal precursors
ATONAL	atonal, lin-32, math1, neuroD	A			Neurogenesis
DELILAH	delilah	A			Differentiation of epidermal cells into muscle
dHAND	dhand, ehand, hxt, hed	A			Rardiac morphogenesis; trophoblast cell development
E12/Da	e12, e47, itf, pan, G12, me2, da	A	I		Neurogenesis, sex determination; regulation of myogenesis
HEN	hen, helhel	A			Neurogenesis
LYL	lyl, scl, nscl, tal	A			Hematopoietic proliferation and differentiation
MYOD	myod1, myogenin, myf5, myf6	A	II		Myogenesis
NEX	nex-1, rat4	A			Neurogenesis
TWIST	twist, ec2, paraxis, scleraxis, dermo	A			Specification of mesoderm lineages
ARNT	arnt	B			Regulation of aryl hydrocarbon receptor activity
CBF	cbf-1	B			Rentromeric binding and chromosomal segregation
ESC	esc1	B			Sexual differentiation in yeast
G-Box	G-box	B			
HAIRY	hlhm, hairy, hes, deadpan, e(spl)	B	VI		Neurogenesis; segmentation
MAD	mad, mxi1	B	IV	Yes	Regulation of cell proliferation
MYC	c-myc, n-myc, l-myc, max	B	III	Yes	Cell proliferation, differentiation; oncogenesis
NO	ino2, ino4	B			Phospholipid synthesis
PHO4	pho4, nuc1	B			Phosphate regulation in yeast
R	r, delila	B			Regulation of anthcynanin pigmentation
SREBP	srebp, add1, hlh106	B		Yes	Sterol synthesis; adipocyte determination
TFE	tfe3, tfeb, mi	B	III	Yes	Activates transcription in immunoglobulin heavy chain enhancer
USF	usf, spf1, namalwa	B		Yes	Upstream stimulation factor; insulin enhancer
SIM	sim, trh, ah	C			Central nervous system midline lineage regulation; tracheal cell induction
ID	id, heira, emc, hlh462	D	V		Negative inhibition of DNA binding; myogenesis, neurogenesis
CENPBR	cenpbr	?			Centromeric binding protein
AP-4	ap-4	?		Yes (2)	Enhance viral and cellular gene activation

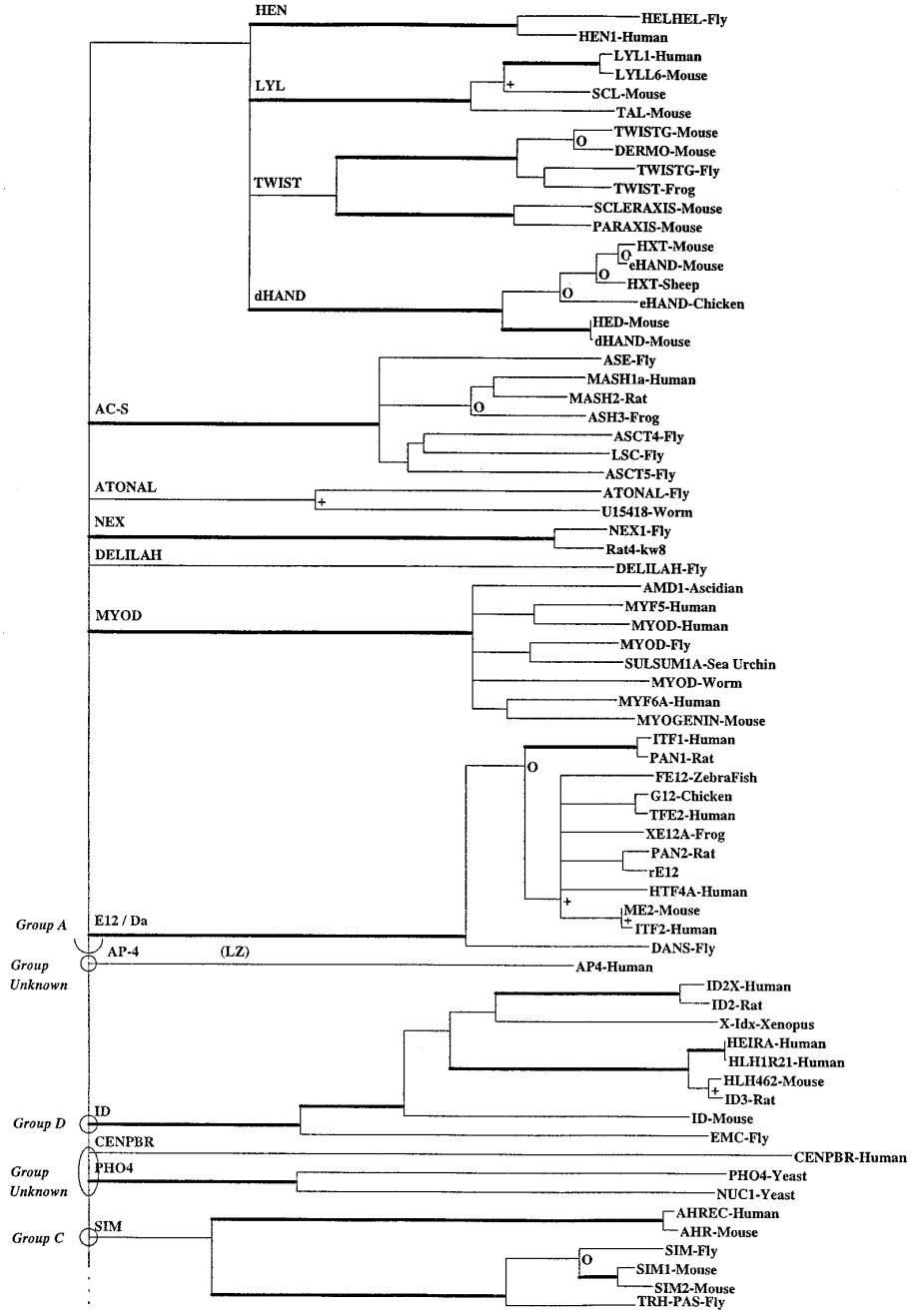
of the overall sequence, (*ii*) the bHLH motif is a cassette or sequence module; (*iii*) there are many paralogues; and (*iv*) the paralogues are very old.

Relationship of Phylogeny to Function. There have been several attempts to categorize bHLH proteins into higher order groups of protein families (e.g., refs. 2, 9–11). Currently, the most widely followed classification of bHLH proteins is one based on how the proteins bind to the core CANNTG E-box (Fig. 1). This classification is naturally depicted by the NJ tree in the present analyses.

Deng *et al.* (9) categorized most of the bHLH proteins into groups A and B, and Swanson *et al.* (11) suggested that Ah and Sim proteins comprise a distinct group C based on their half-site pairing behavior within the E-box. We propose a natural fourth group of proteins (group D) for proteins like Id that lack the typical basic DNA binding region and have a very low frequency of basic residues in the first 13 amino acid sites. Our analyses established patterns of amino acids at sites 5, 8, and 13 (Fig. 1) that discriminate these four groups of bHLH proteins with considerable accuracy.

Group A proteins bind to CA*GC*TG and have a distinctive pattern of amino acids at sites 5, 8, and 13, i.e., a basic amino acid at site 8 and a 5–8-13 configuration of xRx (where R = arginine at site 8 and x is another amino acid at 5 and 13). Furthermore, group A has only small aliphatic residues (A, G) at site 19. The only exceptions are dHand and AP-4, where lysine (K) is substituted at site 8. Accepting the low statistical support at the deep nodes, group A appears monophyletic and includes the protein families Lyl, Twist, Hen, Atonal, Delilah, dHand, AC-S, MyoD, E12, and Da. Validity of group A is strongly reflected in the NJ tree. AP-4 is an odd protein with an unusual E-box binding configuration and is the only group A sequence to contain an LZ (10). We consider AP-4 as a special case and did not include it as a group A protein for these discussions.

Group B binds to CA*CG*TG and has the 5–8-13 E-box configuration BxR with a basic amino acid (either K or H) at



(*Figure continues on the opposite page.*)

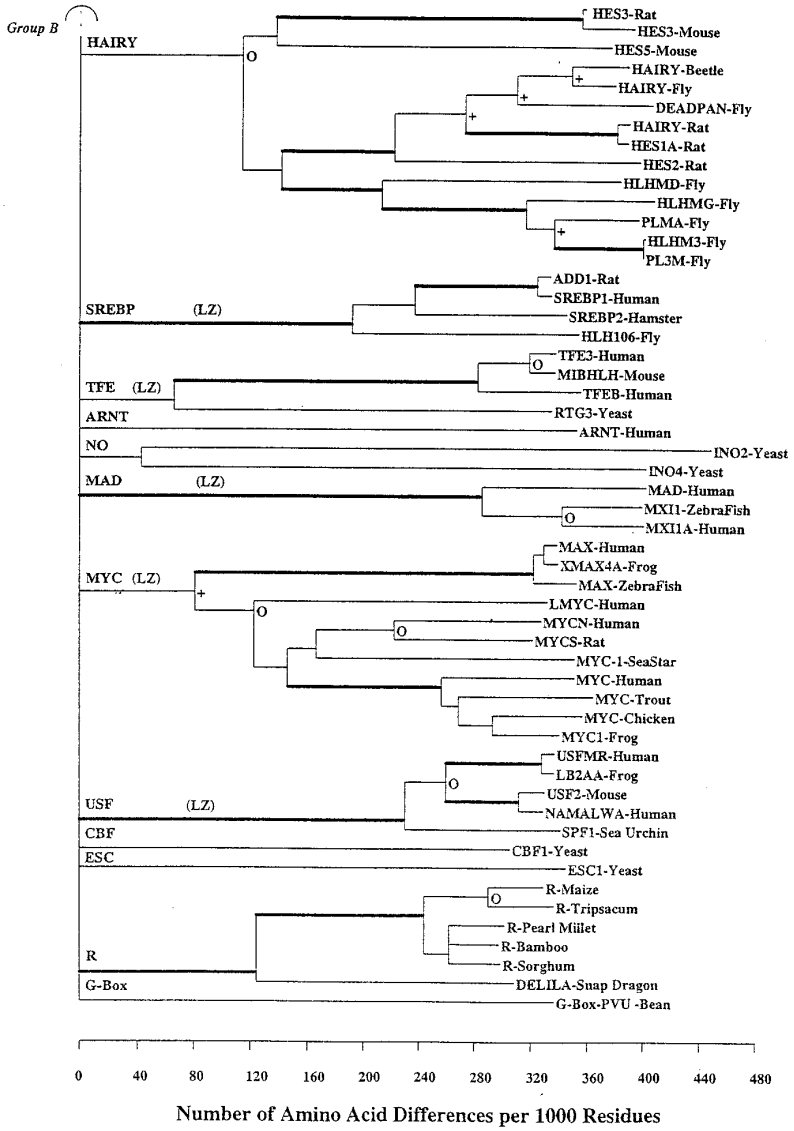


FIG. 2. An NJ phylogenetic tree of 122 bHLH proteins. To simplify the tree, nodes were collapsed to a single horizontal line when statistical support for a particular node was questionable, i.e., when boot strap support was <35%. Furthermore, the boot strap values havebeen coded. All branches subtending a clade supported by >95% of the 500 trials are shown by thickened branches. Branches with support between 80 and 94% are marked with an open circle (○), and those with support between 65 and 79% are delimited with a + sign. Branches with values <65% are unmarked. Clades reflecting proteins containing an LZ are indicated with "LZ."

site 5 and arginine at site 13. Group B includes Arnt, Cbf, Esc, Hairy, Mad, No, Myc, Pho4, R, Srebp, Tfe, Usf, and others (Table 1). Group B can be further partitioned into protein families with or without an LZ motif. For group B proteins with an LZ, the E-box configuration is HxR. Additionally, sequences with an LZ have a very high frequency of N residues (93%) at site 6, the residues at site 8 are almost all aliphatic (I, L, or V), and site 56 is K at 88%

Group C represents a statistically well supported, separate lineage derived from group B but has no consistent amino acid configuration at sites 5, 8, or 13. Furthermore, group C could be further distinguished in these analyses by the absence of basic residues at site 2, A or K at site 9, and E at site 19.

Group D proteins, which include Id, Emc, Heira, and Hhl462 (12), lack the basic DNA binding region, have a very low frequency of basic residues in the first 13 amino acid sites, and frequently have prolines at sites 4 and 9. Group D proteins do not bind DNA; rather, they form protein–protein dimers that function as negative regulators of DNA binding behavior (12). The Id lineage is a statistically well supported single lineage (boot strap value = 99%), and the included proteins probably were derived from a common ancestor that possessed a DNA binding region that was subsequently lost during evolution.

Group D proteins act as dominant negative regulators of MyoD proteins. The question arises of whether a classification of the bHLH proteins based on the helix-loop-helix component alone (basic region removed) would place the group D proteins in the same clade as MyoD. Such an analysis was carried out, and the result was that the Id proteins were still distinct and separate from MyoD. Furthermore, the major clades as seen in Fig. 2 persisted, indicating that evolutionary relationships among protein groups persist when components of the motif are removed. However, the way the major clades were linked together deep in the tree was altered in several instances compared with the results using the full motif. These alterations would be expected in view of the low boot strap values for the deep nodes described above.

These four higher order groups (A-D) are depicted in the NJ tree in Fig. 2, but the boot strap values this deep in the tree are low. Hence, we have used a simple procedure here to further explore the validity of these groups. At each site shown in Fig. 1, the most frequently occurring group is paired with its relevant amino acid at that site. This amino acid then can be used for

Table 2. Unique amino acid changes in bHLH motif for specific clades in NJ tree

Clade name	Unique amino acid changes	Boot strap
Lyl	5:A→T, 13:T→Q, 31L→H, 33:A→P, 51:I→N	99
Twist	1:R→Q	62
dHand	8:R→K, 18:J→S, 25:K→E, 26:L→C, 46:K→T, 52:E→K	100
Hen	1:R→Y, 5:A→H, 6:N→A, 27:I→L, 33:A→P, 53:T→I	100
Ac-s		99
Atonal	27:V→L, 55:R→Q, 56:L→M, 58:K→Q	71
Nex1	20:F→L, 22:K→N, 25:Q→K, 45:N→T, 46:K→Q, 59:K→N	100
Delilah	8 changes for 1 sequence	
MyoD	6:N→T, 13:M→L, 46:K→Q, 47:K→R	100
E12/Da*	(0:D→E), 5:A→N, 14:K→R, 15:B→D, 22:K→E, 26:L→M, 48:L→Q, 60:Y→V	100
Ap4	8 changes for 1 sequence	
Id	20:F→Y, 58:V→I	99
Cenpbr	16 changes for 1 sequence	
Pho4	65:T→S	97
Ah/Sim	52:A→S, 55:K→R	57
Hairy	6:N→P, 49:D→E, 52:A→D, 55:K→E	91
Srebp/Addl	12:R→Y, 15:N→S, 53:I→V, 58:V→I	99
Tfe3	none	100
Arnt	17 changes for 1 sequence	
Ino2	6:N→V	35
Mad/Mxil	3:T→S, 11:K→N, 14:J→A, 17:K→R, 47:K→R, 50:K→T, 53:I→L, 60:Y→H	100
Myc/Max	none (differences are clear at lower levels like Max vs. others)	78
USF	(0:Q→R), 14:E→D, 21:D→V, 25:D→K, 27:L→I, 47:E→G, 59:E→D	100
CBF1	10 changes for 1 sequence	
Escl	11 changes for 1 sequence	
R		99
G-box	6 changes for 1 sequence	

The amino acids are numbered according to ref. 6 as shown in Fig. 1. The boot strap values are given for the relevant node in the NJ tree.

classifying these sequences into the groups. Informative sites (those with probability values greater than 80%) are found in all four sequence components (i.e., basic, helices, and loop). Within the basic region, the sites and their respective probability values are 4 (81%), 5 (92%), 8 (98%), and 13 (95%). Within helix I, the sites are 14 (87%), 19 (92%), 21 (81%), and 25 (86%); within the loop, 29 (82%) and 46 (83%). And for helix II are 52 (85%), 55 (85%), and 56 (86%). Thus, it is clear that the major groups as elucidated by the NJ tree are consistent, and amino acid sites that are informative with regard to group classification occur in all components of the motif.

LZ-Containing Sequences. Six bHLH protein families contain an LZ dimerization motif that follows immediately 5′ to the bHLH motif, including Myc/Max, Mad, Srebp, Ap-4, Usf, and Tfe (Fig. 2). Indeed, some authors have suggested that these bHLH/LZ proteins comprise a natural evolutionary group. The LZ-containing proteins bind to the core CA*CG*TG hexanucleotide, and all have the HxR configuration at amino acid sites 5, 8, and 13. All bHLH proteins with an LZ are found within group B, and, with the exception AP-4 (which has two zippers), none falls into groups A, C, or D.

Based on the collapsed tree, we cannot determine if the LZ-containing proteins are a monophyletic group. However, several lines of evidence support a polyphyletic origin for the LZ-containing proteins. These include the presence of two LZs in AP-4, a protein that is more like group A based on bHLH sequence similarity. Furthermore, in at least one instance in bHLH proteins (E12), an LZ motif occurs that is not related to the function of the bHLH motif (13). In addition, the procedure used to discriminate the various groups of bHLH proteins described above was not very successful in discriminating those bHLH proteins that possess a LZ. Finally, one must consider the simple structure of the LZs and the fact that they also are found in many different transcription factor groups including helix-turn-helix, zinc fingers, steroid receptors, as well in nontranscription factors (14). Indeed, Brendel and Karlin (15) showed, on statistical grounds, that a simple motif represented by heptad repeats of leucines would be expected to occur by chance at a relatively high frequency.

Unique Changes at Specific Nodes. There are amino acid changes in each lineage unique to the clades in question for the 242 sequences used in these analyses (Table 2). Thus, in the Lyl lineage, amino acid site 5 is changed from an A in the ancestor of this lineage to a T in all sequences examined here. Similarly, at 13, T is changed to Q, and, at 31, L is changed to H, and so on. In other lineages, e.g., AC-S, R, or Tfe, there are no changes that would uniquely characterize that clade. Furthermore, in other lineages in which only a single protein is known, e.g., CBF1 or Esc, a number of changes is often noted, but there are no additional proteins in the clade to evaluate the conserved nature of these changes.

Evolution of the bHLH Motif. The ancestral amino acid sequences were estimated by a maximum parsimony analysis using the noncollapsed NJ tree in Fig. 2 as a starting point. Fig. 1 shows the estimated ancestral sequence for the entire bHLH family of proteins as well as the ancestral sequences for groups A-D. These various phylogenetic analyses suggest that group B is most like the ancestral bHLH sequence. We suggest that group B gave rise to group A as well as to group C. Group D may have arisen from group A. The currently known distribution of these proteins in various groups of organisms supports this suggestion. Group B proteins are currently found in plants, yeast, and animals, and group A proteins are found only in animals. Groups C and D are small groups of proteins so far found only in animals.

The authors are indebted to Robin Bush, Neil Abernethy, Jeff Thorne, and Werner Terhalle for their computational assistance. James Mahaffey, Leo Parks, and two anonymous reviewers provided critical comments on the manuscript. The research was supported by grants from the National Institutes of Health (5-46472) and the Sloan Foundation to W.R.A.

1. Murre, C., McCaw, S. S. & Baltimore, D. (1989) *Cell* **56,** 777–783.
2. Murre, C. Bain, G., van Dijk, M. A., Engle, I., Furnari, B. A., Massari, M. E., Matthews, J. R., Quong, M. W., Rivera, R. R. & Stuiver, M. H. (1994) *Biochim. Biophys. Acta* **1218,** 129–135.
3. Sun, X. & Baltimore, D. (1991) *Cell* **64,** 459–467.
4. Voronova, A. & Baltimore, D. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 4722–4726.
5. Thompson, J. D., Higgins D. G. & Gibson, T. J. (1994) *Nucleic Acids. Res.* **22,** 4673–4680.
6. Ferre-D'Amare, A. R., Prendergast, G. C., Ziff E. B. & Burley, S. K. (1993) *Nature (London)* **363,** 38–45.
7. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
8. Atchley, W. R. & Fitch, W. M. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 10217–10221.
9. Deng, C. V., Dolde, D., Gillison M. L. & Kato, G. J. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 599–602.
10. Hu, Y.-F., Luscher, B., Admon, A., Mermod N. & Tjian, R. (1990) *Genes Dev.* **4,** 1741–1752.
11. Swanson, H. I., Chan W. K. & Bradfield, C. A. (1995) *J. Biol. Chem.* **270,** 26292–26302.
12. Benerzra, R. Davis, R. L., Lockshon, D., Turner D. L. & Weintraub, H. (1990) *Cell* **61,** 49–59.
13. Kajimoto, Y. R., Kawamori, Y., Umayahara, H., Watada, N., Iwama, T., Morishima, Y., Yamasaki & Kamada, T. (1994) *Gene* **139,** 247–249.
14. Lewin, B. (1994) *Genes V* (Oxford Univ. Press, Oxford).
15. Brendel, V. & Karlin, S. (1989) *Nature (London)* **341,** 574–575.