J Mol Evol (1999) 48:501–516

JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York Inc. 1999

Positional Dependence, Cliques, and Predictive Motifs in the bHLH Protein Domain

William R. Atchley,[1] Werner Terhalle,[2] Andreas Dress[2]

[1] Department of Genetics, North Carolina State University, Raleigh, NC 27695-7614, USA
[2] Fakultät für Mathematik, Universität Bielefeld, Bielefeld, Germany

Received: 15 December 1997 / Accepted: 1 October 1998

Abstract. Quantitative analyses were carried out on a large number of proteins that contain the highly conserved basic helix–loop–helix domain. Measures derived from information theory were used to examine the extent of conservation at amino acid sites within the bHLH domain as well as the extent of mutual information among sites within the domain. Using the Boltzmann entropy measure, we described the extent of amino acid conservation throughout the bHLH domain. We used position association (*pa*) statistics that reflect the joint probability of occurrence of events to estimate the "mutual information content" among distinct amino acid sites. Further, we used *pa* statistics to estimate the extent of association in amino acid composition at each site in the domain and between amino acid composition and variables reflecting clade and group membership, loop length, and the presence of a leucine zipper. The *pa* values were also used to describe groups of amino acid sites called "cliques" that were highly associated with each other. Finally, a predictive motif was constructed that accurately identifies bHLH domain-containing proteins that belong to Groups A and B.

Key words: Basic helix–loop–helix proteins — bHLH — Information theory — Boltzmann entropy — Predictive motifs — Positional dependence — Cliques

Introduction

Many biological processes are spatially and temporally controlled at the level of transcription. To understand the transcriptional regulation of gene expression, one needs to decipher the molecular modes of differentiation and development of eukaryotic cells. Transcriptional control is mediated by complex interactions between regulatory transcription factors, with their various enhancer elements giving rise to sequence-specific multiprotein complexes that control gene expression at multiple control points (Novina and Roy 1996). Hence, it is crucial that we understand the structure of the various components of these transcriptional complexes, are able to classify their components into well-defined categories, and understand their origin and evolution.

Transcription factors are structurally complex proteins containing distinct functional components associated with DNA binding, protein oligomerization, phosphorylation, activation, and other activities. As a consequence, functionally heterogeneous proteins are often classified based upon small, highly conserved amino acid domains which are discrete connected parts of proteins that can be equated with a particular function. Thus, transcription factors are generally grouped into families like zinc fingers, helix–turn–helix, helix–loop–helix, or basic leucine zippers because the relevant proteins share a particular, short domain associated with DNA binding, oligomerization, or other activities (Lewin, 1997).

Several problems are inherent to evolutionary classifications based on domains. First, the domains are often short and highly conserved so that the amount of information contained within them that can be used for clas-

Correspondence to: W.R. Atchley; *e-mail:* atchley@ncsu.edu

sification may be small. Complicating the issue is the fact that outside the conserved domain, these proteins may exhibit considerable sequence dissimilarity to the point of being apparently unrelated. Second, these domains are associated with a limited number of functions like DNA binding and oligomerization. Mechanistically, there may be only a few ways to solve a particular problem. As a consequence, convergent evolution often cannot be excluded, particularly for structurally simple domains, e.g., the structurally equivalent E-box and G-box domains involved with DNA binding or the leucine zipper oligomerization domain. Third, the definition of the domains in terms of primary sequences are not well understood so that determining whether a particular protein should be included in one of these families is sometimes difficult [e.g., zinc finger proteins (Nakata 1995)].

Consequently, detailed analyses are needed to characterize rigorously the structure and function of these important domains and to deduce their origin and evolution. Such studies require large amounts of divergent data to elucidate better their structural and functional limits as well as to explore the constraints regarding their evolution.

In this paper, we examine some structural aspects of the basic helix–loop–helix domain (bHLH) which defines an important group of transcription factors. bHLH proteins are characterized by highly conserved bipartite domains for DNA binding and protein–protein interaction (Murre et al. 1989). Proteins containing the evolutionarily conserved helix–loop–helix domain are an important class of regulatory components in transcriptional networks of many developmental pathways (Murre et al. 1994). 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 plants to mammals. These various proteins can be grouped into clades and groups reflecting their evolutionary history (Atchley and Fitch 1997).

Since the bHLH domain was first described, a large number of helix–loop–helix proteins have been identified. Most are classified as bHLH transcription factors based on overall sequence similarity with existing bHLH proteins. Several important questions exist regarding the structure of the domain and sequence variability in bHLH proteins. (1) What primary sequence structure identifies a helix–loop–helix protein and how does this structure vary among related proteins? (2) How much sequence variability is permitted while still preserving the necessary helix–loop–helix configuration? (3) Which sites are most highly conserved? (4) What dependencies exist between the amino acid distribution observed at variable sites and clade membership, loop length, and the existence of a leucine zipper? (5) Are there significant associations between the function(s) of these residues and the extent of their evolutionary conservation and/or coevolution?

Consequently, the goal of the analyses reported here is to examine the extent of primary sequence variability in a large number of functionally diverse bHLH proteins, suggest a short hypothetical motif that will serve as a predictive model for identifying putative bHLH proteins, and explore the goodness of fit of this motif to a wide variety of known and of previously unrecognized bHLH proteins.

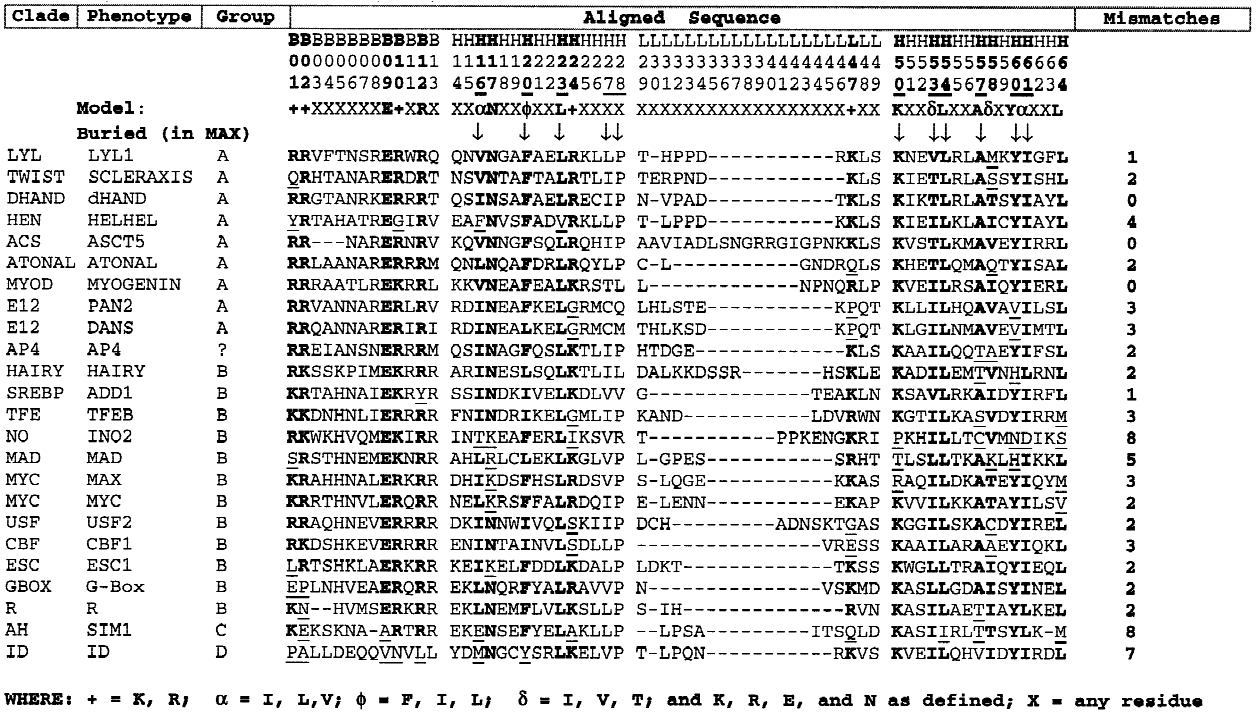
Definition and Structure of the bHLH Domain

The bHLH domain is comprised of approximately 60 amino acids (Fig. 1). A component of mainly basic residues (*b*) permits HLH proteins to bind to a consensus hexanucleotide *E-box* (CANNTG). A second component, referred to as the *HLH* domain, allows these proteins to interact and to form homo- or heterodimers. The dimerization component contains about 50 primarily hydrophobic residues and produces two amphipathic α-helices (H1, H2) separated by a loop (L) of variable length. Additionally, some bHLH proteins contain a leucine zipper (LZ) dimerization domain characterized by heptad repeats of leucines that occur immediately C-terminal to the bHLH domain.

Several authors including Ferre-D'Amare et al. (1993, 1994), Ma et al. (1994), and Ellenberger et al. (1994) 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, with hydrophobic residues from H1 and H2 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 structure appears to be similar to that of other bHLH proteins, such as E47 (Ellenberger et al. 1994), MyoD (Ma et al. 1994), and USF (Ferre-D'Amare et al. 1994).

For consistency, we are following the scheme proposed by Ferre-D'Amare et al. (1993) for delimiting the components of the domain. Numbering of the amino acids included within domain components and delimiting the major evolutionary groups, clades, and lineages follows Atchley and Fitch (1997).

Sites 5, 8, 9, and 13 determine the overall DNA binding configuration (Atchley and Fitch 1997). The presence of a glutamic acid residue (E) at site 9 is required for DNA binding to the E-box and has been shown to contact the CA element of the E-box sequence CANNTG (Ellenberger et al. 1994; Ma et al. 1994; Swanson et al. 1995). This critical glutamic acid residue is found at site 9 in all Group A and B proteins, but in none of the Group C and D proteins.



Clade	Phenotype	Group	Aligned Sequence	Mismatches
			BBBBBBBBBBBBB HHHHHHHHHHHHHH LLLLLLLLLLLLLLLLLLL HHHHHHHHHHHHHH	
			0000000001111 1111112222222222 2333333333344444444444 5555555556666666	
			1234567890123 4567890123456789 9012345678901234567890123456789 0123456789012345	
	Model:		++XXXXXXE+XRX XXφNXXφXXL+XXXX XXXXXXXXXXXXXXXXXXX+XX KXXδLXXAδXYφXXL	
	Buried (in MAX)		↓ ↓ ↓ ↓↓ ↓ ↓↓ ↓ ↓↓	
LYL	LYL1	A	RRVFTNSRERWRQ QNVNGAFAELRKLLP T-HPPD----------RKLS KNEVLRLAMKYIGFL	1
TWIST	SCLERAXIS	A	QRHTANARERDRT NSVNTAFTALRTLIP TERPND----------KLS KIETLRLASSYISHL	2
DHAND	dHAND	A	RRGTANRKERRRT QSINSAFAELRECIP N-VPAD----------TKLS KIKTLRLATSYIAYL	0
HEN	HELHEL	A	YRTAHATREGIRV EAFNVSFADVRKLLP T-LPPD----------KKLS KIEILKLAICYIAYL	4
ACS	ASCT5	A	RR---NARERNRV KQVNNGFSQLRQHIP AAVIADLSNGRRGIGPNKKLS KVSTLKMAVEYIRRL	0
ATONAL	ATONAL	A	RRLAANARERRRM QNLNQAFDRLRQYLP C-L----------GNDRQLS KHETLQMAQTYISAL	2
MYOD	MYOGENIN	A	RRRAATLRERRRL KKVNEAFEALKRSTL L-------------NPNQRLP KVEILRSAIQYIERL	0
E12	PAN2	A	RRVANNARERLRV RDINEAFKELGRMCQ LHLSTE----------KPQT KLLILHQAVAVILSL	3
E12	DANS	A	RRQANNARERIRI RDINEALKELGRMCM THLKSD----------KPQT KLGILNMAVEVIMTL	3
AP4	AP4	?	RREIANSNERRRM QSINAGFQSLKTLIP HTDGE----------KLS KAAILQQTAEYIFSL	2
HAIRY	HAIRY	B	RKSSKPIMERKRRR ARINESLSQLKTLIL DALKKDSSR-------HSKLE KADILEMTVNHLRNL	2
SREBP	ADD1	B	KRTAHNAIEKRYR SSINDKIVELKDLVV G--------------TEAKLN KSAVLRKAIDYIRFL	1
TFE	TFEB	B	KKDNHNLIERRRR FNINDRIKELGMLIP KAND----------LDVRWN KGTILKASVDYIRRM	3
NO	INO2	B	RKWKHVQMEKIRR INTKEAFERLIKSVR T----------PPKENGKRI PKHILLTCVMNDIKS	8
MAD	MAD	B	SRSTHNEMEKNRR AHLRLCLEKLKGLVP L-GPES----------SRHT TLSLLTKAKLHIKKL	5
MYC	MAX	B	KRAHHNALERKRR DHIKDSFHSLRDSVP S-LQGE----------KKAS RAQILDKATEYIQYM	3
MYC	MYC	B	KRRTHNVLERQRR NELKRSFFALRDQIP E-LENN----------EKAP KVVILKKATAYILSV	2
USF	USF2	B	RRAQHNEVERRRR DKINNWIVQLSKIIP DCH---------ADNSKTGAS KGGILSKACDYIREL	2
CBF	CBF1	B	RKDSHKEVERRRR ENINTAINVLSDLLP ----------VRESS KAAILARAAEYIQKL	3
ESC	ESC1	B	LRTSHKLAERKRR KEIKELFDDLKDALP LDKT----------TKSS KWGLLTRAIQYIEQL	2
GBOX	G-Box	B	EPLNHVEAERQRR EKLNQRFYALRAVVP N----------VSKMD KASLLGDAISYINEL	2
R	R	B	KN--HVMSERKRR EKLNEMFLVLKSLLP S-IH----------RVN KASILAETIAYLKEL	2
AH	SIM1	C	KEKSKNA-ARTRR EKENSEFYELAKLLP --LPSA---------ITSQLD KASIIRLTTSYLK-M	8
ID	ID	D	PALLDEQQVNVLL YDMNGCYSRLKELVP T-LPQN----------RKVS KVEILQHVIDYIRDL	7

WHERE: + = K, R; α = I, L,V; φ = F, I, L; δ = I, V, T; and K, R, E, and N as defined; X = any residue

Fig. 1. Representative bHLH proteins, amino acid number scheme, and components of the bHLH domain. Designation of basic (B), helix (H), and loop (L) regions and the numbering sequence for the individual amino acids follow Ferre-D'Amare et al. (1993). Predictive model and its relationship to the aligned bHLH domain for representative sequences of major evolutionary lineages according to Atchley and Fitch (1997). The elements of the predictive model are shown in *boldface* and mismatches are *underlined* and counted to provide the deviation from the predictive model among these representative sequences. *Arrows* and *underlined* site numbers denote amino acids that are buried in the interior of the four-helix bundle in Max according to Ferre-D'Amare et al. (1993). + = K, R; α = I, L, V; φ = F, I, L; δ = I, V, T; and K, R, E, and N are as defined; and X = any residue.

Clades and Groups. Atchley and Fitch (1997) provide an evolutionary analysis of 242 bHLH domain-containing proteins. A neighbor-joining tree describing the major evolutionary lineages (=clades) rooted using the Delia sequence (a bHLH protein found in plants) is given in Fig. 2. This tree has been "pruned" at the terminal nodes to summarize only information about interrelationships about major families of bHLH proteins. More detailed information is provided by Atchley and Fitch (1997).

These numerous clades can be assembled into four major monophyletic groups based upon how the proteins bind to the consensus E-box, the presence of leucine zippers and other attributes. Group A proteins bind to an CAGCTG E-box configuration, while Group B binds to CACGTG (Dang et al. 1992). Group C is a statistically well-supported separate lineage that lacks the critical glutamic acid residue at site 9. The latter predicts that Group C proteins do not bind to any known E-box (Swanson et al. 1995). Group C can be further discriminated by the possession of a unique "PAS" domain composed of two approximately 50-amino acid repeats spaced by approximately 150 residues that is critical for dimerization with other PAS-containing proteins (Zelner et al. 1997). Group D proteins lack the basic DNA binding region, have a very low frequency of basic residues in the first 13 sites, and frequently have proline residues 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 (Murre et al. 1994).

Materials and Methods

Database. A large database of over 400 aligned bHLH domain sequences has been assembled from GenBank, SwissProt, and other sources for the present analyses. They were aligned using the Clustal W alignment algorithm, and the resultant alignment was improved by eye. Two hundred forty-two of these sequences were employed in a previous phylogenetic analyses (Atchley and Fitch 1997).

Predictive Motif. From the 242 sequences used by Atchley and Fitch (1997), we derived a hypothetical search motif to identify putative bHLH proteins which is based upon the frequency of amino acids at individual sites within this large database. Using this search motif, we probed the GenBank and SwissProt databases using a modification of the *agrep* algorithm of Wu and Manber (1991). The agrep algorithm uses "fuzzy" logic to search files for a string and permits searches with (i) a defined level of mismatch including gaps and (ii) site-specific specification of acceptable variants. This fuzzy logic approach avoids problems with "typology" where sequences must conform to an idealized sequence type and therefore permits us to identify protein sequences that match the pattern of the query sequence in a biologically

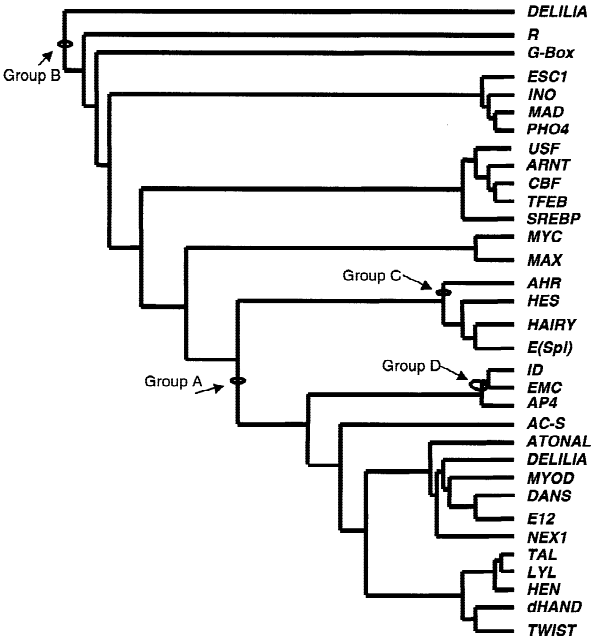


Fig. 2. Neighbor-joining tree summarizing the evolutionary relationships among the protein families containing a bHLH domain. The tree was computed using the PHYLIP software based on a PAM 001 matrix and the terminal nodes have been consolidated to show only protein families. Groups A, B, C, and D (as described by Atchley and Fitch 1997) are denoted. The tree was rooted on Delila, a bHLH sequence found in plants. More extensive information about a neighbor-joining analysis of these data is given by Atchley and Fitch (1997).

more realistic fashion. Using this algorithm, we identified a number of additional bHLH domain-containing proteins in GenBank and SwissProt.

Consequently, we were able to expand the database used by Atchley and Fitch (1997) to one containing over 400 proteins known to exhibit the bHLH domain (392 of these sequences are examined here). The additional protein sequences were added only in cases where the protein had been shown experimentally to contain a bHLH domain or appeared to be closely related (by sequence analysis) to proteins known to contain a bHLH domain. In addition, our search procedures identified a number of further sequences that probably are bHLH proteins, including cosmids, open reading frames from the various genome projects, and proteins whose function is yet unknown. The latter proteins are currently not part of the database but are being added as we determine whether they actually contain bHLH domains. They are not included currently because the purpose of the present analyses is to characterize that domain, and to this end we used only proteins known to contain it. Subsequent papers will deal with these other proteins and with methods for assessing their membership to the bHLH family.

Estimating Sequence Variability. This report focuses on patterns of primary amino acid sequence variability. Within highly conserved domains like bHLH, some sites exhibit very little variability. No doubt, this must be caused by functional and structural constraints exercising strong evolutionary pressure to preserve a particular pattern of residues. Such highly conserved sites generally reflect the structure and function of a given domain. Other sites can be much more variable, and the combination of residues at particular variable sites often contains a strong phylogenetic signal, distinguishing evolutionary lineages and providing information characterizing clade structure (Atchley and Fitch 1997). Because evolutionary variation at the molecular level can be neutral, one might anticipate some variation not related to a functional signal (i.e., random noise). However, because of the highly conserved nature of domains like bHLH, the amount of random sequence variability within them is probably quite small.

Comparing amino acid sequences presents several difficult problems, not the least of which is that such analyses involve variability in "symbols" for which there is no natural ordering or metric. Characterization of the relative information at each site for all 392 bHLH sequences is measured in terms of the Boltzmann entropy E as employed by Shannon (Shannon and Weaver 1949). It measures the degree of variation among categories of amino acids at each site in the domain and is defined as

$$E: = -\sum P_i \log_2 P_i$$

where P_i is the relative frequency of residues belonging to category i (with $P_i \log_2 P_i: = 0$ if $P_i = 0$; the colon before an equation sign indicates that the left-hand term is defined by the right-hand term). So E is zero when all elements are in the same category. It increases with both the number of categories and their equiprobability, its maximal possible value being $\log_2 n$, if n categories are being considered. Gaps are excluded from the computation of E in our analyses.

The Boltzman–Shannon statistics were computed in two ways. First, E was computed at each site with every single amino acid forming one category. Second, E was computed (denoted E_F) according to the following classification: acidic (D, E), basic (K, R, H), aromatic (F, Y. W), aliphatic (A, G, I, L, V, M), amidic (N, Q), hydroxylated (S, T, Y), cysteine (C), and proline (P).

Positional Interdependence in the bHLH Domain. The phenomenon of "covariability" among sequence elements, where the amino acid composition at one site can be estimated with some reliability by the amino acid composition at another site, is an important concept for understanding sequence evolution and function. However, estimating covariation is difficult with biological sequence data because they involve "symbols" (letters for nucleotides or amino acids) having no underlying natural ordering or metric thus preventing use of conventional statistical procedures. Using methodology derived from information theory, we can estimate the *mutual information content* between distinct amino acid sites. This is a measure derived from the probability of joint occurrence of events (Kullback 1959). If events are independent, then the mutual information is 0; if events are dependent, mutual information is positive (Farber et al. 1992; Herzel and Gross 1995; Clarke, 1995; Roman-Roldan et al. 1996). This approach permits one to estimate not only the association among various amino acid sites but also the extent of association between the amino acid composition at any given site with other variable properties. The latter might include estimating the phylogenetic information content of a given amino acid site by the extent of association of its amino acid composition with the phylogenetic structure of the group of proteins in question. The amino acid composition at various sites can also be related to variables reflecting the function of structural components of proteins, the length of the loop or turn in bHLH or HTH proteins, and the presence of another conserved domain like a leucine zipper or a PAS domain.

We refer to these estimates as *position association* (*pa*) *statistics.* With regard to the association between amino acids in an aligned family of sequences, for any given variable v (a position in the alignment) defined on a set S of aligned sequences and for each possible value A of v (a single amino acid or a functional group of amino acids), we estimate the unweighted probability $p(A/v)$ for the variable v to attain the value A by

$$p(A|v) = (\text{No. sequences } s \text{ in } S \text{ with } v(s) = A)/\text{No. S}$$

Similarly, for any pair of positions v, w and any pair of amino acids A and B or, more generally, possible values of v and w, respectively, we define $p(A,B|v,w)$ as the number of sequences s in S with $v(s) = A$ and $w(s) = B$ divided by No. *S.* There is no "association" between variable

v and variable w if $p(A,B|v,w)$ roughly coincides with $p(A|v) \times p(B|w)$ for all A, B, as is obviously the case when w has only a single value B, and one has $p(A,B,|v,w) = p(A|v) = p(A|v) \times p(B|w)$.

According to standard techniques, based on the convexity of the function $f(x) := x\ln x$, the association between variable v and variable w can now be measured by v, w:

$$pa(v,w) := \sum_{A,B} p(A,B|v,w) \times \ln(p(A,B|v,w)/p(A|v) \times p(B|w))$$

with $p(A,B|v,w)/p((A|v) \times p(B|w) := 1$ whenever $p(A|v) \times p(B|w) = 0$. It is worth noting that, more generally, given a strictly convex function $f(x)$ defined for all nonnegative numbers x, satisfying, in addition, the relation $f(1) = 0$, e.g., $f(x) := x(x-1)$, it can be shown that the number

$$pa_f(v,w) = \sum_{A,B} p(A|v) \times p(B|w) f(p(A,B|v,w)/(p(A|v) \times p(B|w))$$

is always nonnegative, while this number vanishes if and only if $p(A|v) \times p(B|w)$ equals $p(A,B|v,w)$ for all A,B that is, if and only if v and w are statistically independent for each other.

More generally, the identity

$$\sum_A p(A,B|v,w) = p(B|w)$$

and the inequality

$$0 \le p(A,B|v,w)/p(A|v) \le 1$$

implies

$$\begin{aligned} pa(v,w) &= \sum_{A,B} p(A,B|v,w)\ln(p(A,B|v,w)/p(A|v)p(B|w)) \\ &= \sum_B \left(\sum_A p(A,B|v,w)\ln(p(A,B|v,w)/p(A|v) \right) \\ &\quad - \sum_A p(A,B|v,w)\ln(p(B|w)) \\ &\le -\sum_B p(B|w)\ln(p(B|w)) \\ &=: E(w) \end{aligned}$$

if we define the Boltzmann–Shannon entropy $E(w)$ of the variable w in this way; so the value of the association of v and w can never exceed that of $E(w)$ or—by symmetry—that of $E(v)$.

We have used position association values to describe the extent of association between the amino acid composition at the 64 sites within the bHLH domain. Further, we use *pa* values to describe *cliques* of sites defined as groups of positions such that *any* two positions in that clique have *pa*-values among the highest 5% of all such values, which, as it turned out, are exactly those with values greater than 1.0. These cliques, of course, describe higher-order association and indicate mutual interdependence between a whole range of positions.

As mentioned above, this mutual information approach can be extended to measure association among other types of variables. For example, we estimate the phylogenetic signal exhibited by the various sites by computing *pa* values between each site and a variable representing membership of each protein in a particular evolutionary *group* or *clade*. The term "groups" refers to Groups A, B, C, and D described by Atchley and Fitch (1997). The term "clade" refers to monophyletic lineages contained in these groups which usually reflect functionally similar families of proteins. These clades are defined in Table 1 of Atchley and Fitch (1997).

We have also computed *pa* values between amino acid sites and (i) the number of amino acids in the loop (*loop length*) or (ii) the presence or absence of a leucine zipper (*zipper*). In each of the two latter instances, the *pa* values provide information about either the predictability of loop length or the presence of a leucine zipper from the amino acid composition at various sites.

Results

Conservation of Amino Acids Within the bHLH Domain

At each site, the extent of primary sequence variability and the most frequently occurring amino acids, E and E_F, together with the resulting rank order, are given in Table 1.

Referring to the structure of Max as a general model (Ferre-D'Amare et al. 1993), there are several specific sites within the bHLH domain worthy of notice. Using the numbering system in Fig. 1, the highly conserved basic residue at site 2 begins the first α-helix, which continues to site 27. Within the basic region, site 2 has an arginine (R) residue in 77%, site 9 has glutamic acid (E) in 93%, and sites 10 and 12 have arginine (R) in 81 and 91% of all the proteins, respectively. In Helix 1, site 16 has an aliphatic residue (I, L, or V) in 91%, site 17 has asparginine (N) in 74%, site 20 has F, I, or L in 95%, and at site 23 leucines (L) occur in 98% of all proteins. The end of Helix 1 (site 28) has a proline in 63% of the proteins, an amino acid well-known to break helices. The first residue in the second helix (site 50) is lysine (K) in 93% of bHLH proteins in our database, while 98% have leucine at site 54.

The amino acid sequence in most parts of the loop is quite variable; however, some sites exhibit consistent patterns of amino acid conservation. Properly aligned, site 47 has basic residues (K or R) occurring in 80% of proteins, while 45% of the proteins have a leucine at site 48.

The extent of amino acid diversity at each site is another important attribute when characterizing domains (Table 1, Fig. 3a). By definition, the (theoretical) maximum value for E is $\log_2(20) = 4.32$. Ranking E values in the H1 and H2 components (loop omitted) shows that the 10 sites with the greatest amino acid diversity (excluding the loop) are 21 ($E = 3.47$, H1), 62 ($E = 3.45$, H2), 3 (B), 63 (H2), 7 (B), 14 (H1), 18 (H1), 59 (H2), 26 (H1), and 56 ($E = 3.08$, H2). The 13 sites with the smallest E values are (in increasing order) 23 ($E = 0.15$, H1), 54 ($E = 0.20$, H2), 9 (B), 50 (H2), 12 (B), 10 (B), 2 (B), 17 (H1), 64 (H2), 57 (H2), 53 (H2), 60 (H2), and 20 ($E = 1.27$, H1). In these least diverse sites, the highly conserved residue is leucine (L) at sites 23, 54, and 64, glutamic acid (E) at site 9, lysine (K) at site 50, arginine (R) at sites 12, 10, and 2, asparginine (N) at site 17, alanine at site 57, isoleucine (I) at site 53, tyrosine (Y) at site 60, and phenylalanine (F) at site 20.

Very similar results were obtained for E_F. The Spear-

Table 1. Percentage occurrence of amino acids in the bHLH domain of 392 bHLH domain-containing proteins[a]

Position and component	Amino acid frequency within the bHLH domain	Shannon		Rank	
		E	E_F	E	E_F
1 basic	***K(27%), R(61%)***	1.6880	0.8730	15	14
2 basic	***K(16%), R(77%)***	1.1670	0.5223	7	9
3 basic	A(6%), K(18%), M(9%), R(21%), S(8%), T(5%), V(11%)	3.4199	1.8819	41	29
4 basic	A(35%), K(5%), M(5%), N(11%), S(6%), T(20%)	3.0254	2.1559	31	38
5 basic	A(27%), H(41%), K(8%), N(12%)	2.3358	1.8147	22	28
6 basic	N(59%), P(9%), T(16%), V(6%)	2.0407	1.8121	19	26
7 basic	A(22%), E(11%), I(8%), L(10%), M(16%), V(14%)	3.3456	1.5272	39	22
8 basic	I(5%), L(26%), R(44%), S(7%)	2.4660	1.5815	24	23
9 basic	***E(93%)***	0.5163	0.4509	3	8
10 basic	***K(14%), R(81%)***	0.9876	0.3357	6	6
11 basic	K(13%), L(8%), N(9%), Q(19%), R(35%)	2.9159	1.9269	29	30
12 basic	***R(91%)***	0.5907	0.5944	5	11
13 basic	L(18%), M(5%), R(49%), T(5%), V(17%)	2.1366	1.4621	20	21
14 helix 1	A(9%), D(9%), E(12%), K(14%), N(19%), Q(9%), R(12%), S(7%)	3.3118	2.3851	38	42
15 helix 1	D(20%), E(13%), H(5%), K(25%), N(7%), R(11%), S(9%)	3.0784	1.9793	33	33
16 Helix 1	***I(35%), L(33%),*** M(6%), ***V(23%)***	2.0320	0.2965	18	4
17 Helix 1	K(15%), ***N(74%),*** R(9%)	1.1814	0.9321	8	15
18 Helix 1	D(11%), E(30%), G(5%), L(12%), N(7%), R(5%), S(10%), T(6%)	3.2587	2.1470	37	36
19 Helix 1	A(41%), C(10%), G(6%), M(6%), R(5%), S(21%)	2.6321	1.9382	26	31
20 Helix 1	***F(72%), I(9%), L(14%),*** Y(5%)	1.2716	1.0034	13	18
21 Helix 1	A(11%), D(11%), E(19%), F(12%), K(11%), L(13%), S(6%)	3.4749	2.3636	43	41
22 Helix 1	A(23%), E(27%), Q(6%), R(8%), T(14%), V(5%)	3.0341	2.1558	32	37
23 Helix 1	***L(98%)***	0.1482	0.0254	1	1
24 Helix 1	G(12%), ***K(35%), R(44%)***	1.9306	0.9435	16	16
25 Helix 1	D(28%), E(9%), K(12%), Q(5%), R(25%), S(9%), T(6%)	2.9019	1.9480	28	32
26 Helix 1	C(9%), H(8%), I(6%), L(33%), M(11%), Q(10%), S(6%), V(5%)	3.0969	2.0980	35	34
27 Helix 1	C(8%), I(30%), L(13%), T(15%), V(32%)	2.2784	0.9730	21	17
28 Helix 1	L(7%), P(63%), Q(10%), S(5%), V(6%)	1.9776	1.6449	17	24
29 loop	A(7%), D(7%), E(20%), L(14%), S(14%), T(16%)	3.4376	2.2630	na	na
30 loop	A(17%), C(6%), E(9%), H(39%), S(7%), T(8%), Y(5%)	2.7402	2.3567	na	na
31 loop	I(9%), L(36%), N(17%), V(12%)	3.0844	1.8051	na	na
32 loop	A(11%), D(5%), E(7%), H(6%), K(12%), P(35%), Q(5%)	3.2011	2.4182	na	na
33 loop	A(11%), G(5%), K(8%), N(31%), P(8%), Q(7%), S(17%)	3.1333	2.2760	na	na
34 loop	D(28%), E(16%), N(23%), Q(21%)	2.6119	1.6820	na	na
35 loop	G(33%), L(13%), P(13%), S(13%)	2.9333	2.0874	na	na
36 loop	A(18%), E(30%), I(7%), P(5%), R(5%), S(23%), T(5%)	2.8213	2.0861	na	na
37 loop	A(12%), H(19%), K(12%), L(7%), N(10%), Q(7%), R(29%)	2.8269	1.5888	na	na
38 loop	G(67%), I(11%), Q(22%)	1.2244	0.7642	na	na
39 loop	G(83%), R(17%)	0.6500	0.6500	na	na
40 loop	A(33%), D(17%), G(17%), I(17%), R(17%)	2.2516	1.2516	na	na
41 loop	D(10%), G(90%)	0.4690	0.4690	na	na
42 loop	I(10%), L(10%), P(10%), R(60%), S(10%)	1.7710	1.5710	na	na
43 loop	G(45%), H(9%), N(32%), S(14%)	1.7492	1.7492	na	na
44 loop	A(8%), I(11%), L(5%), P(8%), S(24%), T(18%), V(18%)	2.9495	1.6100	na	na
45 loop	H(19%), I(14%), K(14%), N(11%), S(6%), T(11%), V(13%)	3.1495	2.1663	na	na
46 loop	E(24%), K(33%), R(9%), S(11%), T(14%)	2.6358	1.9215	na	na
47 loop	***K(58%),*** P(9%), ***R(24%)***	1.7630	0.9758	na	na
48 loop	A(20%), L(45%), M(5%), Q(12%), V(9%)	2.4830	1.0830	na	na
49 loop	A(7%), D(9%), E(8%), N(5%), P(27%), S(31%), T(14%)	2.5702	1.9968	na	na
50 Helix 2	***K(93%)***	0.5247	0.3202	4	5
51 Helix 2	A(20%), I(10%), L(15%), V(42%)	2.4597	0.5839	23	10
52 Helix 2	D(9%), E(32%), G(6%), L(10%), Q(5%), S(10%), V(17%)	3.0015	1.8139	30	27
53 Helix 2	***I(74%), T(15%), V(7%)***	1.2396	0.6592	11	12
54 Helix 2	***L(98%)***	0.1988	0.0711	2	2
55 Helix 2	A(6%), E(9%), H(8%), K(20%), Q(7%), R(36%)	2.8410	1.6650	27	25
56 Helix 2	E(6%), K(28%), L(19%), M(6%), N(9%), Q(11%), S(11%)	3.0844	2.1417	34	35
57 Helix 2	***A(76%),*** S(5%), T(14%)	1.2329	0.8213	10	13
58 Helix 2	***I(31%), T(23%), V(27%)***	2.4911	1.2916	25	19
59 Helix 2	A(17%), D(13%), E(24%), K(7%), Q(6%), R(8%), S(13%)	3.1594	2.2849	36	40
60 Helix 2	H(8%), ***Y(77%),*** V(10%)	1.2473	1.3090	12	20
61 Helix 2	***I(69%), L(16%), V(8%)***	1.4565	0.1071	14	3

Table 1. Continued

Position and component	Amino acid frequency within the bHLH domain	Shannon E	Shannon E_F	Rank E	Rank E_F
62 Helix 2	E(13%), H(6%), K(13%), L(21%), Q(8%), R(16%)	3.4489	2.2498	42	39
63 Helix 2	A(10%), D(6%), E(9%), F(6%), G(5%), K(7%), N(5%), R(6%), S(29%), Y(8%)	3.3618	2.4877	40	43
64 Helix 2	***L(80%),*** M(7%), V(5%)	1.2084	0.3714	9	7

[a] Amino acids are listed if they occur at least 5% of the time. The Shannon statistic (E) is a measure of variety and is zero when all elements (amino acids) are the same at a given site. E was computed based on 20 amino acids, while E_F was computed based on eight functional groups of amino acids. The theoretical maximum value of E for 20 amino acids is 4.32 and 3.0 for eight functional groups. Sites given in boldface italics are those sites included in the predictive model. E values were not ranked for the loop positions because of the high frequencies of gapped sites arising from difficulties in homologizing the loop positions.

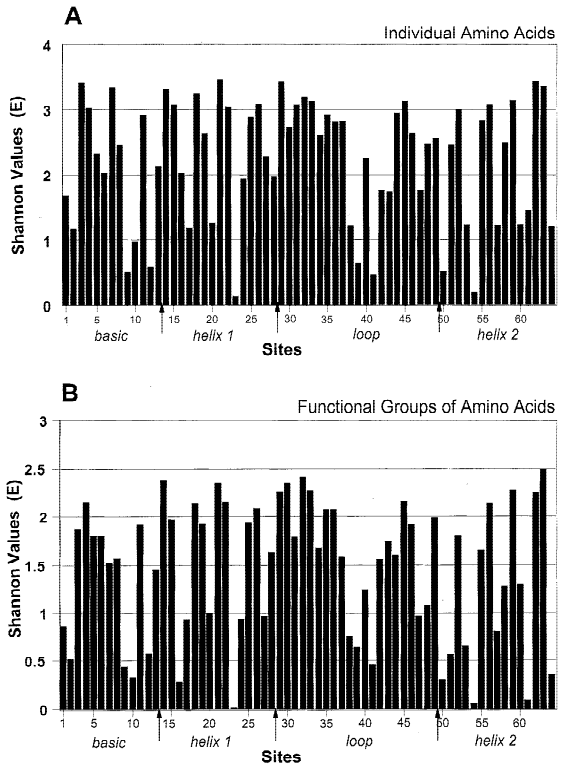


Fig. 3. Shannon uncertainty statistic scores for each site within the bHLH domain. A total of 392 sequences is included in the plot. **a** reflects the individual residues, while **b** reflects residues transformed to "functional groups" of amino acids. See text for further details.

man rank correlation (Sokal and Rohlf, 1995) between E and E_F was 0.89 ($P < 0.001$). Exceptions are sites like 16 and 61, whose diversity reflects variation in residues that belong to the same functional class.

Among the functionally least diverse positions, sites 23, 54, 61, 16, and 64 all are aliphatic, with L, I, and V among the most prevalent residues; sites 50, 10, and 2 are basic sites, with K or R as the most prevalent residues; and the exception is site 9, where E (an acidic residue) is the most highly conserved.

Ten sites are common among the lowest 13 sites with respect to both E and E_F. Sites 17, 60, and 20 (61, 16, and 51) are the sites which are among the lowest 13 with respect to E (or E_F) but not to E_F (or E, respectively).

Relations Among Variability, Structure, and Function

Relationship to DNA Binding. The basic component of the domain is characterized experimentally as being associated with DNA binding to a consensus hexanucleotide "E-box" (CANNTG) (Voronova and Baltimore 1990). In its primary sequence, highly conserved positively charged (basic) residues (K or R) occur at a 90% frequency or more at sites 1, 2, 10, and 12. These sites mark the beginning and end of the DNA binding region. Site 9 has a highly conserved glutamic acid (E) residue (93%) which is found in every Group A and B protein but none of the Group C and D proteins. Fisher and Goding (1992), Ellenberger et al. (1994), Ma et al. (1994), and others suggest that a glutamic acid residue is in contact with the CA element of the E-box and is required at site 9 for DNA binding to occur. Group D does not have a DNA binding component. The situation in Group C is not clear. Swanson et al. (1995) suggest that these Group C proteins may not bind to DNA, at least they do not bind any known E-box. Instead, they may be involved with other bHLH proteins like ARNT in a combinatorial mechanism of gene regulation.

Sites 3–8, 11, and 13 exhibit more diversity. At least three of these more variable sites are involved in enhancing DNA-binding specificity, and the patterns of amino acids at sites 5, 8, and 13 discriminate four phylogenetic lineages within the bHLH proteins (Atchley and Fitch 1997).

Fisher and Goding (1992) have shown that single amino acid substitutions at sites flanking the core CANNTG-binding motif can also change binding site specificity. Thus, while sites 5, 8, and 13 provide binding specificity for Groups A to D, flanking residues at sites 3, 4, 7, and 11 appear to provide more refined discrimination with regard to clades of bHLH proteins.

Buried vs. Exposed Helix Residues. There is often a relationship between the structural conformation of a protein and the extent of amino acid variability at relevant sites. For example, in folded structures, amino acid residues F, L, I, and M tend to be fully buried, while charged residues R, K, H, E, and D tend to be fully

exposed on the surface (Richards 1992). In addition, variation at exposed sites might regulate fine-tuning of function, while variation at buried sites might correspond to drastic changes of the folding architecture.

In their folded (native) state, proteins exhibit an intricate three-dimensional structure. The molecular arrangement of proteins, including the relationship between the three-dimensional structure and the site-specific amino acid specification and variability, is important from both functional and evolutionary perspectives.

By the series of arrows in Fig. 1, we have indicated those residues in Max that are buried in the dimer according to Ferre-D'Amare et al. (1993). Extrapolating from Max to other bHLH proteins, we can relate variability in primary sequence to structural conformation. The buried sites of Max in Helix 1 are 16, 20, 23, 24, 27, and 28. And those in Helix 2 include 50, 53, 54, 57, 60, and 61. As expected, buried sites have a high preponderance of hydrophobic residues. The exceptions include site 50 (which is 93% K) and site 60 (77% Y).

An important null hypothesis to evaluate in a large group of proteins is whether buried and exposed sites in Helix 1 and Helix 2 have the same level of primary sequence variability. To test this hypothesis, Boltzmann–Shannon values (E) for each site were ranked and a Mann–Whitney nonparametric test (Sokal and Rohlf 1995) carried out to evaluate the null hypothesis that the two samples (buried versus exposed sites) have the same median E value. There are 30 sites in the two helices and the median rank for the exposed sites was 21, while the median for the buried sites was 8. The null hypothesis of equal medians in the two samples was rejected at $P < 0.001$, indicating that, per site, buried sites have significantly less sequence variability than exposed sites, over these 397 sequences. This test was repeated using the functional groups of amino acid (E_F) as data. With these data, the null hypothesis was again rejected at $P < 0.001$.

Characteristics of the Loop

The loop is variable in length and difficult to align, suggesting little, if any, sequence homology. These and other attributes raise interesting questions about loop structure. Table 2 gives the frequencies of distinct loop sequences of differing lengths. The shortest loop has five residues (CBF-1), suggesting that at least that many residues are needed to maintain a parallel four-helix bundle structure.

Observations about variability of primary structure and length might suggest that the loop exists simply to provide spacing required for dimerization and its amino acid content is not very important from a functional or evolutionary perspective. Several lines of evidence bring this conclusion into question.

First, while it is difficult to align portions of the loop, sites 47 and 48 exhibit rather high levels of conservation and there are no gaps in the aligned sequences for these two sites. Clearly, this argues against the idea that composition of the loop is irrelevant.

Table 2. Distribution and frequency of more common loop lengths in the database of 392 sequences

Loop length	No. unique sequences
5	2
6	9
7	3
8	14
9	49
10	26
11	2
12	11
13	3
14	10
15	0
16	2
17	0
18	1
19	0
20	2
21	3

Second, the length of the loop varies *systematically* among bHLH sequences. While loop length varies from 5 residues in CBF to 39 in RTG1 in yeast, its length within any given protein family is quite stable. Table 3 provides the average loop length and its standard deviation for the various bHLH clades and groups. In many of the clades, there is little or no variation in loop length. Indeed, the *pa* value regarding the variables for "clade" and loop length is quite high (>1.4) (Table 4), while the *pa* value regarding "group" and loop length is low (0.35).

Within the achaete–scute protein family, however, there is considerable heterogeneity and the loop length varies from 8 to 21 residues. Achaete–scute proteins in *Drosophila* such as Ast5 and Ast8 have 21 residues in the loop; however, homologues of achaete–scute (Mash) in mammals and chickens have only 8 residues.

Third, there is a reasonably high association between loop length and amino acid composition at certain otherwise quite variable sites in the basic and helix components (Fig. 3): *pa* values >1.0 are found for sites 52 (1.08) and 21 (1.04). Values >0.9 are found for 14, 29, 15, and 56. All of these sites are highly variable (E rank ≥ 30).

Fourth, swap experiments show that these loops are not always functionally interchangeable among bHLH proteins (Pesce and Benezra 1993).

An important question is which residues at highly correlated sites are associated with loops of differing lengths? Table 5 details the relationship between specific amino acid residues at individual sites and loop length. The 10 amino acid sites with the highest *pa* values are given, together with the amino acid with the largest observed/expected ratio. Thus, at site 52 (which has the

Table 3. Means and standard deviations of loop length for various bHLH protein families

Family	*n*	Average	SD[a]
CBF	5	5.00	0.00
R	23	6.00	0.00
SREBP	9	7.00	0.00
AP4	2	8.00	0.00
ESC	1	8.00	N/A
MyoD	59	8.02	0.13
Twist	11	8.27	0.47
Myc	83	8.98	0.56
Atonal	6	9.00	0.00
Dhand	7	9.00	0.00
Hen	6	9.00	0.00
LYL	14	9.00	0.00
Nex	3	9.00	0.00
Mad	7	9.00	0.00
ID	16	9.00	0.00
Arnt	2	10.00	0.00
NO2	1	10.00	N/A
NO4	1	10.00	N/A
AH/Sim	10	10.00	0.00
E12	47	10.04	0.29
TFE	6	11.00	2.45
Pho4	1	12.00	N/A
USF	10	12.00	0.00
ACS	21	13.38	5.47
Hairy	32	13.63	0.71
Delila	2	16.00	0.00
Nuc1	1	20.00	N/A

[a] Standard deviations of zero indicate no variation in loop length within that family.

largest *pa* value between amino acid content and loop length), alanine occurs altogether in only 15 of the 392 sequences, yet it occurs in all 5 sequences with a loop length of 5. Similarly, at site 21 (which has the second highest *pa* value), a glutamic acid residue occurs in altogether 70 sequences, yet 56 of the 79 sequences have a loop length of 8.

Clique Structure

Understanding the structure of evolutionarily conserved functional domains is facilitated by elucidating the extent of association among pairs of amino acid sites. The "clique" structure of the bHLH domain is an important functional and evolutionary concept. Cliques are defined as groups of positions, all of which are more highly associated with each other than any are to a nonmember of the clique. Maximum cliques are those not contained in larger cliques.

Position association values [$pa_{(i,j)}$] computed for sites *i* and *j* in the bHLH domain describe the interdependence among amino acid sites. The highest value found was just above 1.25 and the highest 5% (= 101) of values found were all just above 1.00. To evaluate the significance of this finding, the amino acids in any column of the alignment were rearranged at random, while none were exchanged between columns. This procedure kept constant the amino acid distribution and, hence, the *E* and E_F values at each position, as well as the quality of the overall alignment. It turned out that more than 40% of the pairs in the original data had a higher *pa* value than the highest *pa* value of 0.375 found for a shuffled alignment, while more than 15% had a *pa* value at least twice that large.

Table 4. Positional association statistics showing association between variables reflecting clade, group, presence or absence of a leucine zipper, and length of the loop

	Clade	Group	Zip	Loop length
Clade	—			
Group	0.953	—		
Zip	0.655	0.377	—	
Loop length	1.415	0.352	0.185	—

The maximal cliques we found among the pairs of positions with a *pa* value above 1.00 are listed in Fig. 4. Clearly, the positions 3, 4, 7, 14, 21, 52, 56, and 62 are most highly involved in these cliques and would form a clique themselves if only the *pa* value of the pair (7, 52) were above 1. Note that these positions exhibit also a very high association with loop length.

Phylogenetic Information Content

Variability in amino acids at various sites reflects functional, structural, and phylogenetic information together with a random noise component. Success in estimating evolutionary histories of proteins based on sequence information is possible because variability at the various amino acid sites includes a strong phylogenetic signal and the distribution of specific amino acid residues is often highly associated with specific nodes in a phylogenetic tree. One could argue that the random noise component in evolutionarily highly conserved domains might be smaller than in other portions of the overall sequence due to natural selection placing functional constraints on random sequence variability. We can gain insight about protein evolution by examining position association values between the distribution of amino acids at various sites or loop length, on the one hand, and clade membership and related features, on the other.

Association Values. Position association (*pa*) statistics between the amino acid composition and the designations for clade or group measure the amount of phylogenetic signal contained in the various amino acid sites. Figure 5 provides graphical summaries of the position association values describing the extent of association between each amino acid site and either clade or group.

Table 5. The *pa* values describing the association between specific amino acid residues at individual sites and loops of various lengths[a]

			Association with loop length and individual sites and amino acids							
Site	*pa*	Rank	5	6	7	8	9	10	12	14
Number of sequences			5	25	9	79	150	65	15	24
Basic-3	0.869	6	D-29 (3)	Z-7 (23)	T-19 (8)	K-4 (44)	R-2 (56)	M-5 (28)	A-11 (9)	V-5 (12) S-6 (10)
Basic-5	0.848	9	H-3 (5)	Z-4 (23)	H-3 (9)	A-4 (58)		N-6 (39)	H-2 (10)	K-14 (24)
Helix 1-14	0.976	3	E-5 (3)	E-9 (25)	S-12 (S)	K-3 (34)	N-2 (70)	R-5 (38)	D-7 (10)	A-10 (23)
Helix 1-15	0.942	4	N-15 (5)	K-4 (25)	S-12 (9)	K-3 (56)	E-2 (47)	D-3 (42)	K-3 (10)	R-9 (24)
Helix 1-21	1.042	2	N-24 (3)	L-7 (22)	I-37 (6)	E-4 (56)	F-3 (46)	K-5 (39)	V-19 (10)	D-4 (11)
Helix 1-26	0.845	10	L-3 (5)	L-2 (17) V-7 (8)	L-3 (9)	C-4 (28)	Q-3 (41)	M-6 (41)	I-12 (10)	L-3 (21)
Helix 2-52	1.078	1	A-26 (5)	S-10 (25)	A-15 (5) G-7 (4)	E-2 (55)	V-2 (65)	L-6 (38)	G-10 (10)	D-10 (23)
Helix 2-55	0.864	7	A-10 (3)	A-12 (19)	R-3 (9)	R-2 (58)	K-2 (62)	H-6 (33)	S-13 (9)	E-12 (24)
Helix 2-56	0.912	5	R-22 (4)	E-14 (21)	K-4 (9)	N-5 (32) S-3 (26)	K-2 (76)	Q-5 (38)	K-2 (10)	L-2 (11) M-6 (9)
Helix 2-62	0.860	8	Q-12 (5)	K-7 (21)	E-5 (5)	E-5 (51)		L-3 (38)	R-4 (9)	R-3 (10)

[a] The 10 amino acid sites with the highest *pa* values between site and loop length are given, plus their rank. For loops with lengths varying between 5 and 14, the amino acid residue (a) with the highest n/N value is given, where n is the number of $s_{(i)} =$ a and N is the theoretical value under the assumption of independence. These proportions are rounded to the nearest whole number. The number in parentheses is the number of sequences containing the residue in question. Not enough sequences with loops of length 11, 13, and 15–21 are available to make meaningful analyses.

There are several large *pa* values between clade designation and amino acid composition. Of the 64 values, the largest is 1.7 and there are 19 sites with values >1.4. The largest values occur for sites 21, 14, 3, 52, 15, and 56, all with values >1.5, which, except for either site 56 or site 15, also form a (nonmaximal) clique.

Some sites in the basic region are involved with enhancing DNA binding specificity in particular protein families. Sites 5, 8, and 13 are important residues in Groups A–D that define DNA-binding patterns. Other sites in the basic region may be involved with an enhanced level of binding specificity within these major groups of protein families. For example, Fisher and Goding (1992) have shown experimentally that a single amino acid substitution converted the binding specificity of the bHLH protein Pho4 to that of Cpf1 (= Cbf1). Both proteins are Group B but in separate clades. Hence, one expects high *pa* values between individual amino acids in the basic region and the clade variable since these individual amino acids (or combinations of amino acids) are specifying protein family specific binding patterns. And indeed, almost two-thirds of the sites in the basic components (namely, the eight sites 3–8, 11, and 13) have values >1.4, while only one-third of the sites in the H1 and the H2 components (namely, the five sites 14, 15, 21, 25, and 26 and the five sites 52, 55, 56, 58, and 62) exhibit that high a value.

The *pa* values between amino acid sites and group designation (Groups A, B, C, and D) are considerably lower than those seen with the clade variable. The largest value is 0.9, seen for site 8, and the next four highest values are found at sites 19, 13, 3, and 24. Lower *pa* values with the group variable indicate that there is less information about phylogenetic structure at this level of organization compared to that seen at the level of the protein clades. This result is also evident from the bootstrap values in the neighbor-joining tree presented by Atchley and Fitch (1997).

Ranked values for E and the *pa* values for clade (loop sites excluded) are shown in Fig. 6. It is apparent from Table 1 and Fig. 6 that an association occurs between *pa* values and the E statistic. In correspondence with the inequality $pa(v,w) \leq \min(E(v),E(w))$ derived above, amino acid sites with the largest *pa* value for the clade

Basic Component						Helix 1							Loop							Helix 2					
3	4	5	7	8	11	14	15	18	19	21	25	26	29	31	32	33	34	48	49	51	52	55	56	62	63
X	X		X			X				X			X										X		
X	X		X			X				X						X								X	
X	X		X			X				X													X	X	
X	X					X				X			X								X		X		
X	X					X				X											X		X	X	
X			X			X	X			X						X								X	
X						X	X			X		X									X			X	
X						X				X		X									X		X	X	
X			X							X			X	X									X		
X			X							X				X		X								X	
X			X							X				X									X	X	
	X					X				X	X		X										X		
			X		X		X			X														X	
						X				X											X	X	X		
										X								X			X	X	X		
	X			X		X					X														
	X			X		X															X				
						X				X									X		X				
										X		X			X						X				
										X			X		X						X				
	X							X		X															
		X				X															X				
				X		X			X																
					X					X												X			
						X			X	X															
						X				X															X
						X														X					
										X							X								

Fig. 4. Maximum cliques found within the top 5% of pairs of positions (= 101), which by chance coincide with just those pairs of positions with a *pa* value above 1.0. The cliques are defined by each row, with an X designating each site in the clique.

variable also have the largest *E* values, while those sites with little phylogenetic signal have low *E* values. The Spearman correlation is 0.88 ($P < 0.01$). This high level of statistical association is easily understood: Those sites with low *E* values are providing information about highly conserved sites, but by definition, those sites where a small number of residues are highly conserved are not very informative phylogenetically. The correlation is less ($r = 0.66$) between the ranks for the *E* statistics and the *pa* values for group.

Is there information within the bHLH domain that would predict the presence of a leucine zipper in some proteins? Within Group B, a number of protein families possess a leucine zipper, which is an additional dimerization device to facilitate protein–protein interactions. The *pa* values in this instance are not very large, with the largest values being found for sites 56 (*pa* = 0.53), 52 (0.49), 48 (0.44), 8 (0.44), and 19 (0.43).

Predictive Motif

Based upon the pattern of sequence conservation, a hypothetical motif was generated that includes 19 elements, i.e., 18 elements from the basic and helix components and 1 from the loop, and of course also specifies (lower and upper bounds for) the distance between these elements. Because these represent the most conserved elements in the bHLH motif, it is reasoned that this pattern of residues will accurately discriminate proteins containing the bHLH domain. The predictive motif (shown in Fig. 1) is

$$\mathbf{++X_{(3–6)}E+XRX_{(3)}\alpha NX_{(2)}\phi X_{(2)}L+X_{(5–22)}}$$
$$\mathbf{+X_{(2)}KX_{(2)}\delta LX_{(2)}A\delta XY\alpha X_{(2)}L}$$

where + = K, R; α = I, L, V; ϕ = F, I, L; δ = I, V, T; E, R, K, A, and Y are as defined; X = any residue; $X_{(i)}$ = any *i* residues; and $X_{(i–j)}$ = *i* to *j* of any residues. Note that the altogether 30 H1 and H2 sites encompass 11 buried sites, of which 9 are used in this predictive motif (of 14 H1 and H2 sites used in the motif).

Interactive Aspects of the Predictive Motif. An initial version of the predictive model was generated using the 242 known bHLH sequences analyzed by Atchley and Fitch (1997). This initial motif was then used to probe GenBank and SwissProt for proteins containing zero to eight mismatches. With this first version, no sequences were found that fit the predictive motif exactly (with no mismatches), 105 were found that had one mismatch, 213 with up to two, 697 with up to three, and 6492 with up to four mismatches. The motif then was redefined at certain sites to improve its goodness of fit to the known bHLH sequences. The revised motif (as described here)

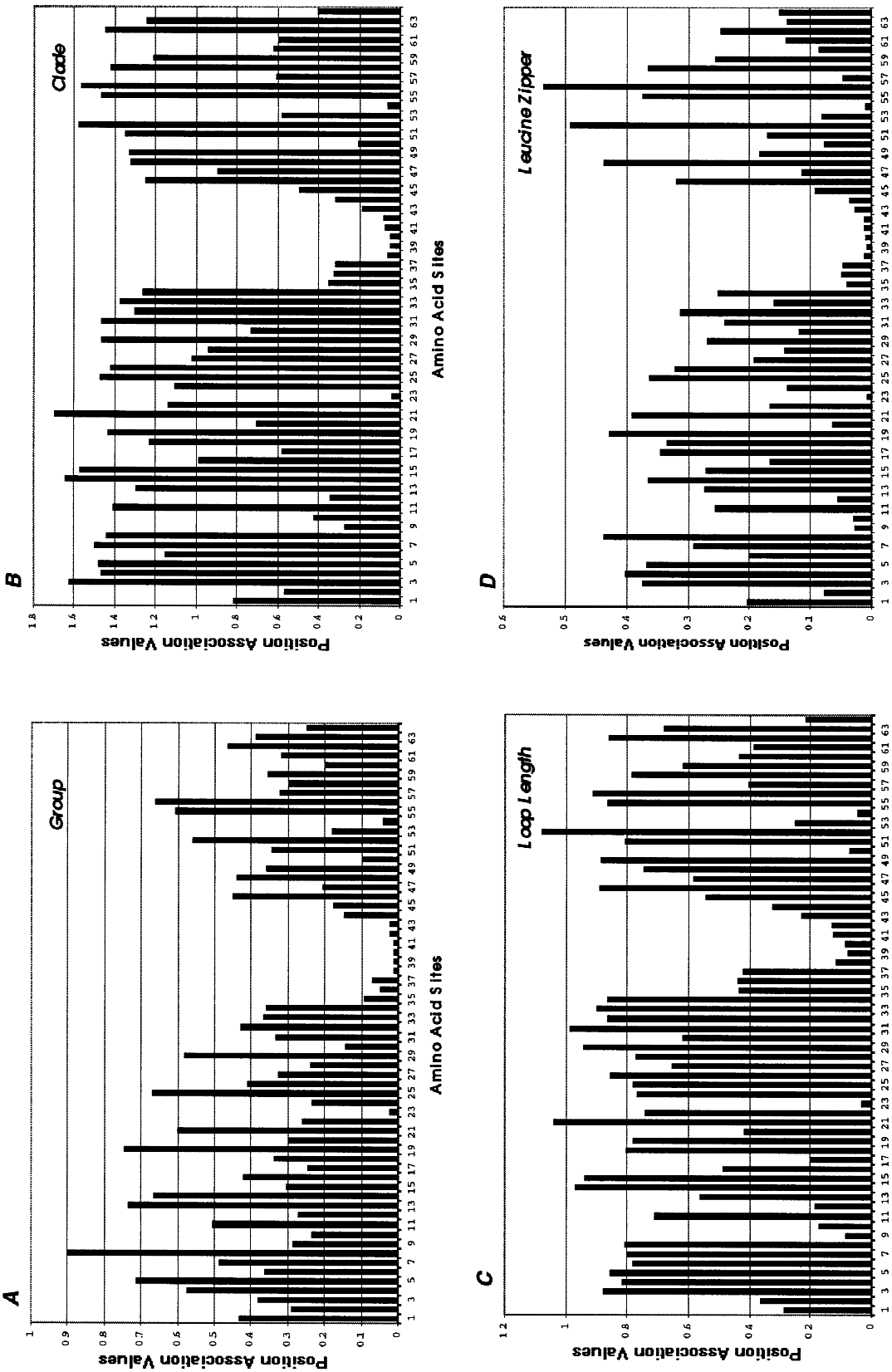


Fig. 5. Bar diagrams of *pa* values for amino acid sites and variables reflecting group **(A)**, clade **(B)**, loop length **(C)**, and the presence of a leucine zipper **(D)**. Amino acid sites 1–13 reflect the basic or DNA-binding region, 14–28 are Helix 1, sites 29–49 are the loop, and sites 50–64 reflect Helix 2.

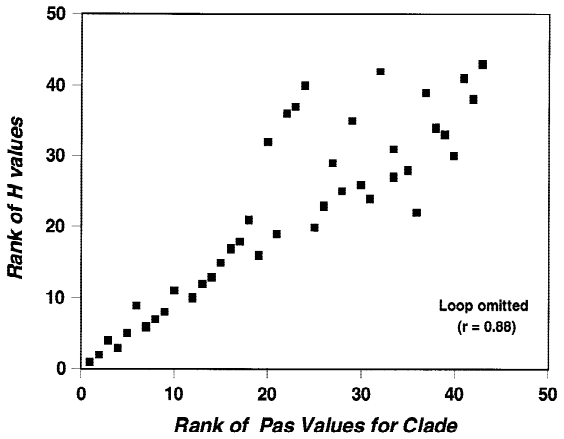


Fig. 6. Bivariate scatter plot of the rank of the Shannon *H* values versus the rank of the *pa* values for clade.

is considerably more accurate: there are 111 sequences that fit the motif exactly (no mismatches), 228 sequences fit with up to one mismatch, 333 with up to two, and 450 with up to three mismatches, while 562 sequences were found with up to four mismatches.

To verify this approach further, a set of comparable predictive motifs was constructed from other sequence patterns. Ten searches were carried out on GenBank using predictive motifs with the same degree of specification as used here but constructed from three sequences randomly chosen from GenBank. For each motif, the number of "hits" in the database was computed for an ordered array of up to zero to seven mismatches. The following results (average hits ± SD) were obtained: 0 mismatches (7.5 ± 9.7), 1 (8.2 ± 11.1), 2 (9.3 ± 11.5), 3 (9.6 ± 11.7), 4 (28.4 ± 23.2), 5 (399 ± 442.7), 6 (4736 ± 4357), and 7 (31,721 ± 16,927).

Goodness of Fit to the Motif. Figure 1 provides a comparison of this predictive motif with a series of bHLH domain sequences representing the various evolutionary lineages (= clades) from the phylogenetic analyses of Atchley and Fitch (1997). Figure 1 gives the numbering system and the relevant elements of the predictive motif; the extent of agreement within the various representative sequences is given in boldface. From Fig. 1, it can be seen that the degree of fit to the predictive motif ranges from zero mismatches (perfect fit) for dHand, ASCT5, and MyoD (Group A) to seven mismatches for ID (Group D) and eight mismatches for Sim1 (Group C) and Ino2 so sequences from these two groups cannot be expected to be singled out by the present Group A/B adapted predictive motif. The highest levels of correct matches are found in proteins representative of Groups A and B, and with the exception of the Mad sequence, the extent of mismatches in this group of representative proteins ranges from zero to three.

The greatest lack of fit to this predictive model is found in proteins representative of Group D (ID) and Group C (Sim1) (Atchley and Fitch 1997) or in sequences whose E-box binding affinities are uncertain (e.g., Ino2). For Group D (ID and related proteins), there is no basic DNA binding region and these proteins function as dominant negative regulators. Indeed, five of the seven mismatches in ID are in the basic region, while those in Ah (Group C) are more equally distributed over the various components of the domain. Ino2, on the other hand, has all of its mismatches in the two helix regions. Indeed, the first residue in the H2 region is a proline, an amino acid known to break helices.

Table 6. Goodness of fit of predictive model to observed basic helix–loop–helix data for the four groups that reflect E-box binding affinities: Average number of mismatches (with SD in parentheses) for each component of the bHLH domain

Total	*n*	Basic	Helix 1	Loop	Helix 2	Total
A	178	0.13	0.34	0.26	0.52	1.26
		(0.10)	(0.15)	(0.08)	(0.18)	(0.12)
B	182	0.30	0.75	0.09	1.47	2.62
		(0.15)	(0.22)	(0.05)	(0.30)	(0.17)
C	10	3.40	1.30	0.30	2.30	7.30
		(0.44)	(0.28)	(0.08)	(0.36)	(0.28)
D	16	4.38	2.06	0.31	1.44	8.19
		(0.47)	(0.35)	(0.09)	(0.29)	(0.30)

Table 6 summarizes the goodness of fit of the sequences to the predictive motif summarized over the four evolutionary groups (as described by Atchley and Fitch 1997). Most of the sequences in the data set are Group A and B proteins, and only 27 of 392 fall into Groups C and D. Over the entire database, Table 7 shows that sites that best conform to the predictive motif (those with mismatches <9%) over all sequences are basic region sites 9, 10, and 12; Helix 1 sites 20 and 23, both buried; and Helix 2 sites 50, 53, 54, and 61, also buried.

However, when concordance to the predictive motif is examined within the four groups, quite different results occur. Clearly, Groups C and D fit much less than Groups A and B. Groups A and B constitute the vast majority of the sequences in the database. The fit to the motif by Group A proteins is excellent for sites 2, 9, 10, 12, 16, 17, 20, 23, 50, 53, 54, 57, 61, and 64, where the percentage of mismatches is <4% at each site. The fit by Group B sequences is of that order only for sites 9, 10, 20, 23, 47, and 54, with up to 10% mismatches for sites 2, 12, 16, 50, and 53.

Table 8 describes the goodness of fit of the predictive motif to the database in terms of the average number of mismatches per component of the motif. Thus, for the 178 sequences classified as Group A an average of 0.13 mismatches is observed in the basic component, 0.34 in Helix 1, and 0.52 in Helix 2. Overall, there is an average of 1.26 mismatches over the entire motif among Group A proteins. Group B has, on average, about twice as many mismatches as Group A, with an average of 2.62 mismatches over all of the motif. Groups C and D have

Table 7. Extent of variation in amino acids at each site from the predictive model for the four major bHLH groups of proteins as defined by Atchley and Fitch (1997)

		Percentage mismatches per site in group				
Position	Model	A (157)	B (177)	C (11)	D (16)	Total
1 (b)	KR	13	12	64	81	18
2 (b)	KR	1	10	55	56	10
9 (b)	E	—	—	100	100	7
10 (b)	KR	1	1	—	100	5
12 (b)	R	—	6	64	100	9
16 (h1)	ILV	4	7	36	100	10
17 (h1)	N	1	53	—	6	26
20 (h1)	FIL	1	—	—	100	5
23 (h1)	L	3	1	—	—	2
24 (h1)	KR	18	15	100	—	18
47 (L)	KR	18	—	36	31	18
50 (h2)	K	1	10	—	38	7
53 (h2)	ITV	—	7	—	—	4
54 (h2)	L	—	2	36	6	2
57 (h2)	A	4	32	100	100	25
58 (h2)	ITV	25	14	—	—	19
60 (h2)	Y	18	27	—	—	24
61 (h2)	ILV	1	12	—	—	6
64 (h2)	L	2	40	91	—	21

much greater mismatch frequencies. Group C has 3.4 mismatches in the basic region alone and the remaining 3.9 in the remainder of the motif. The identifying characteristic of Group D is the absence of a DNA-binding component, which is reflected by an average of 4.4 mismatches in the basic component. Regarding the remaining part of the motif, the average of 3.8 mismatches per protein is slightly smaller than in Group C.

Table 8 also gives mismatch frequencies by component for the major protein clades. The fit is quite good in some evolutionary lineages (e.g., achaete–scute, Dhand, and MyoD, with mismatches below 0.25 over the whole motif) but poor in others. For example, Mad has a considerable lack of fit in the Helix 2 component and a mismatch rate of 6.0 overall. Considered over all clades the goodness of fit in the basic and Helix 1 components is high for most clades and considerably lower in Helix 2.

To ascertain the relative efficacy of the functional components of the bHLH domain to the predictive model, we probed GenBank and SwissProt using elements of the predictive motif corresponding to the functional components of the domain. Using only those elements corresponding to the basic region ($++X_{3-6}E+XR$), the search reported 2819 sequences with no mismatches. Searching with the Helix 1 component only ($\alpha NX_2\phi X_2L+$) gave 2091 sequences with no mismatches. Thus, there are many other proteins in addition to those with a bHLH domain that contain these two small motifs. Searching with the Helix 2 component only ($KX_2\delta LX_2A\delta XYAX_2L$) gave 172 sequences with no mismatches, 475 with one mismatch, and 4930 with up to two mismatches. Using the Helix 1 + Helix 2 components (no basic component) without specifying order or loop length gave 120 sequences with zero mismatches, 236 sequences with one mismatch, and 461 with two mismatches. The latter search is important for identifying Group C and D proteins, which do not have a defined basic region. No attempt was made as yet to ascertain how many of these sequences actually are bHLH proteins.

Table 8. Goodness of fit of predictive model to observed basic helix–loop–helix data in selected individual clades (protein (families): Average number of mismatches (with SD in parentheses) for each component and the total bHLH domain

Taxon	*N*	Basic	Helix 1	Loop	Helix 2	Total
ACS	21	0.10 (0.09)	0.00	0.00	0.05 (0.06)	0.14 (0.04)
Atonal	6	0.00	0.17 (0.11)	0.33 (0.09)	1.00 (0.25)	1.50 (0.13)
Dhand	7	0.14 (0.10)	0.00	0.00	0.00	0.14 (0.04)
E12	47	0.00	0.96 (0.24)	0.96 (0.15)	1.00 (0.25)	2.91 (0.18)
Hen	6	1.17 (0.29)	1.17 (0.27)	0.00	0.00	2.33 (0.16)
LYL	14	0.07 (0.07)	0.07 (0.07)	0.00	0.93 (0.24)	1.07 (0.11)
MyoD	59	0.02 (0.04)	0.08 (0.07)	0.00	0.14 (0.09)	0.24 (0.05)
Twist	11	1.00 (0.27)	0.18 (0.11)	0.00	0.82 (0.23)	2.00 (0.15)
Hairy	32	0.25 (0.14)	0.16 (0.10)	0.00	2.50 (0.37)	2.91 (0.18)
Mad	7	1.00 (0.27)	1.00 (0.25)	0.00	4.00 (0.44)	6.00 (0.26)
Myc	83	0.01 (0.03)	1.04 (0.25)	0.01 (0.02)	0.99 (0.25)	2.05 (0.15)
R	23	1.00 (0.27)	0.04 (0.05)	0.00	1.26 (0.28)	2.30 (0.16)
SREBP	9	1.00 (0.27)	0.00	0.00	0.22 (0.12)	1.22 (0.12)
USF	10	0.10 (0.09)	0.90 (0.24)	0.90 (0.14)	1.10 (0.26)	3.00 (0.18)
AH/Sim	10	3.40 (0.44)	1.30 (0.28)	0.30 (0.08)	2.30 (0.36)	7.30 (0.28)
ID	16	4.38 (0.47)	2.06 (0.35)	0.31 (0.09)	1.44 (0.29)	8.19 (0.30)

Discussion

At the outset, we asked several questions about the bHLH domain. First, we inquired what primary sequence structure identifies a bHLH protein and how this sequence structure varies among related proteins. To resolve this question, we deduced a 19-element *predictive motif* based on relative variability at sites from the basic and helix components. This motif shows considerable efficacy for identifying putative bHLH proteins. It is quite accurate, in particular, in identifying those that belong to Groups A and B and considerably less accurate

with proteins belonging to Groups C and D. The primary reason for the loss of accuracy in the latter two groups is the absence of a well-defined basic or DNA-binding region in Groups C and D. The predictive motif shows considerable promise regarding the identification and evolutionary classification of putative bHLH proteins. Work is currently under way using the predictive motif to identify open reading frames, cosmids, and other similar data in various databases that could be bHLH proteins so that they can be verified with more detailed experimental analyses.

Second, we inquired about the relative distribution of highly conserved and variable sites within the motif and how these relate to function and phylogeny. Computation of entropy values for each site in the domain indicates patterns of highly conserved and highly variable sites that tend to relate to sites constrained by function or by evolution, respectively.

Third, we asked about whether dependencies exist between specific amino acid sites and various extrinsic variables such as loop length, group, and clade assignments. There are various sites that show high levels of association with the length of the loop and clade membership. Further, we have described maximum cliques among sites where significant levels of association occur among amino acids at various sites. These results offer strong impetus for further analyses regarding the functional and phylogenetic bases for these high association values.

Fourth, we asked whether significant associations occur between functions of residues and their evolutionary conservation. Another way of stating this is to ask for the amount of "information" contained in the various sites in the bHLH domain. The relative information content is reflected in the extent of diversity and/or conservation in amino acid residues at each site in the domain. Sites that are highly conserved and that show very little amino acid diversity over almost all bHLH proteins tell us about preservation of function in the entire group of bHLH proteins. High levels of overall conservation probably reflect functions shared by all the proteins in the bHLH family. Thus, the presence of a glutamic acid residue at site 9 in the basic region is well-known to be associated with DNA binding, and this amino acid is found in all of the Group A and B but in none of the Group C and D proteins. Similar instances of amino acid conservation probably related to function include the sites starting both α-helices (sites 2 and 50). In both instances, these sites contain basic residues (K or R) in 93% of all the proteins. Likewise, site 47 has been shown in Max to stabilize the path of the loop, and in 82% of the proteins this site has a basic residue (K or R). In other instances, single residues are highly conserved at specific sites but the function of these sites is less clear, i.e., leucine residues at sites 23 and 54 occur in 98% of all the proteins. Clearly, it is important to determine the function of such sites. What functional role do these positions play that has necessitated such rigorous constraints.

In other instances, variability is clearly restricted within highly conserved functional groups of amino acids. For example, sites 61 and 16 exhibit an aliphatic residue (I, L, or V) for 92% or more of the proteins. Hopefully, additional crystal structure studies on a diverse set of bHLH proteins will provide the necessary information regarding the role of functional amino acid classes.

Correspondingly, the high levels of diversity in amino acid composition at a number of interesting sites within the bHLH domain tell us about patterns of evolutionary divergence. Sites with high *E* values are often the sites that best distinguish various clades in the phylogenetic tree and often show a high degree of association with phylogenetic structure as indicated by the "group" and "clade" variables. Clade structure clearly reflects the functionality component in that it reflects families of transcription factors that have well-specified roles in specific developmental processes, e.g., neurogenesis, myogenesis, and cell proliferation.

The results given here point to the value of formal mathematical modeling for understanding the structure and function of large families of related proteins. These results provide considerable quantitative insight into the structure of the bHLH domain, its evolution, and its function. Similar analyses of other transcription factor families would probably also be highly beneficial.

Acknowledgments. We are indebted to Walter Fitch, Jeff Thorne, Hanah Margalit, Kurt Wollenberg, and James Rosinski for their helpful comments on early drafts of the manuscript. Neil Abernethy and Marianne Barrier provided computational assistance. The research was generously supported by the National Institutions of Health (5-46472), the National Science Foundation (INT-9603452), the Deutscher Akademischer Austauschdienst (315-/PPP/ru-ab), and the Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (413-4001-01 IB 306 B).

References

Atchley WR, Fitch WM (1997) A natural classification of the basic helix-loop-helix class of transcription factors. Proc Natl Acad Sci USA 94:5172–5176

Clarke ND (1995) Covariation of residues in the homeodomain sequence family. Prot Sci 4:2269–2278

Crews S (1998) Control of cell lineage-specific development and transcription by bHLH-PAS proteins. Genes Dev 12:607–620

Deng CV, Dolde D, Gillison ML, Kato GJ (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

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

Farber R, Lapedes A, Sirotkin K (1992) Determination of eukaryotic protein coding regions using neural networks and information theory. J Mol Biol 226:471–479

Ferre-D'Amare AR, Prendergast GC, Ziff EB, Burley SK (1993) Recognition by Max of it cognate DNA through a dimeric b/HLH/Z domain. Nature 363:38–45

Ferre-D'Amare AR, Pognonec P, Roeder RG, Burley SK (1994) Structure and function of the b/HLH/Z domain of USF. EMBO J 13: 180–189

Fisher F, Goding CR (1992) Single amino acid substitutions alter helix-loop-helix protein specificity for bases flanking the core CANNTG motif. EMBO J 11:410–4109

Herzel H, Gross I (1995) Measuring correlations in symbol sequences. Physica A 216:518

Kullback S (1959) Information theory and statistics. Wiley, New York

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

Li W, Marr TG, Kaneko K (1994). Understanding long-range correlations in DNA sequences. Physica D 75:392–416

Ma PCM, Rould RA, Wentraub H, Pabo CO (1994) Crystal structure of MyoD bHLH domain DNA complex: Perspectives on DNA recognition and implications for transcriptional activation. Cell 77:451–459

Murre C, McCaw PS, Baltimore D (1989) A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD and myc proteins. Cell 56:777–783

Murre C, Bain G, van Dijk G, et al. (1994) Structure and function of helix-loop-helix proteins. Biochim Biophys Acta 1218:129–135

Nakata K (1995) Prediction of zinc finger DNA binding protein. Comput Appl Biosci 11:125–131

Novina CD, Roy AL (1996) Core promoters and transcriptional control. Trends Genet 12:351–355

Pesce S, Benezra R (1993) The loop region of the helix-loop-helix Id1 is critical for its cominant negative activity. Mol Cell Biol 13:7874–7880

Richards FM (1992) Folded and unfolded proteins: An introduction. In: Creighton TE (ed) Protein folding. W.H. Freeman, San Francisco, pp 1–58

Roman-Roldan R, Bernaola-Gavan P, Oliver JL (1996) Application of information theory to DNA sequence analysis: A review. Patt Recog 29:1187–1194

Shannon C, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Urbana

Sokal RR, Rohlf FJ (1995) Biometry. W.H. Freeman and Sons, San Francisco

Swanson HI, Chan WK, Bradfield CA (1995) DNA binding specificities and pairing rules of the Ah receptor, ARNT and SIM proteins. J Biol Chem 270:26292–26302

Voronova A, Baltimore D (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

Wu S, Manber U (1992) Fast text searching allowing errors. Commun ACM 35:83–91

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