Correlations Among Amino Acid Sites in bHLH Protein Domains: An Information Theoretic Analysis

William R. Atchley,[1] Kurt R. Wollenberg,* Walter M. Fitch,† Werner Terhalle,‡ and Andreas W. Dress‡*

*Department of Genetics, North Carolina State University; †Department of Ecology and Evolutionary Biology, University of California at Irvine; and ‡Fakultät für Mathematik, Universität Bielefeld, Bielefeld, Germany

An information theoretic approach is used to examine the magnitude and origin of associations among amino acid sites in the basic helix-loop-helix (bHLH) family of transcription factors. Entropy and mutual information values are used to summarize the variability and covariability of amino acids comprising the bHLH domain for 242 sequences. When these quantitative measures are integrated with crystal structure data and summarized using helical wheels, they provide important insights into the evolution of three-dimensional structure in these proteins. We show that amino acid sites in the bHLH domain known to pack against each other have very low entropy values, indicating little residue diversity at these contact sites. Noncontact sites, on the other hand, exhibit significantly larger entropy values, as well as statistically significant levels of mutual information or association among sites. High levels of mutual information indicate significant amounts of intercorrelation among amino acid residues at these various sites. Using computer simulations based on a parametric bootstrap procedure, we are able to partition the observed covariation among various amino acid sites into that arising from phylogenetic (common ancestry) and stochastic causes and those resulting from structural and functional constraints. These results show that a significant amount of the observed covariation among amino acid sites is due to structural/functional constraints, over and above the covariation arising from phylogenetic constraints. These quantitative analyses provide a highly integrated evolutionary picture of the multidimensional dynamics of sequence diversity and protein structure.

Introduction

Analyses integrating protein structure and evolution can proceed in different ways. One popular experimental approach, for example, is to carry out site-directed mutagenesis where specific amino acids within a protein are altered, and then to examine the effects of these changes on protein characteristics. Changes in amino acid attributes can then be correlated with changes in protein characteristics (e.g., Koshi and Goldstein 1997 and references therein).

Another approach is to model large families of naturally occurring proteins or protein domains and determine how nature has changed their characteristics over billions of years of evolutionary diversification. By examining patterns of sequence diversity, one can explore how naturally occurring sequence variability and amino acid properties (e.g., hydrophobicity, volume, and charge) are important in maintaining protein structure. The latter approach permits analyses of protein structure and function over extensive geological timescales where evolutionary processes have experimented in nature with amino acid changes with regard to protein stability, foldability, and functionality.

Experimental, as well as quantitative, analyses of proteins (including multiple-alignment procedures) often proceed by modeling frequencies of residues at individual amino acid sites. For computational expediency, these analyses assume that amino acid sites are independent, i.e., the presence of a residue at one site is assumed to be independent of residues at other sites (Swofford et al. 1996). However, it is well known that this assumption is naïve, since the activities and properties of proteins are the result of interactions among their constitutive amino acids. Interactions among amino acid sites include salt bridges between charged residues, hydrogen bonds between electron acceptors and donors, size constraints reflecting structural interactions between large and small side chains, electrostatic interactions, hydrophobic effects, Van der Waal's forces, and similar phenomena.

Detecting structural interactions and statistical covariance or associations among separate amino acid sites is fundamental for understanding protein structure and evolution. Consequently, it is important to determine the magnitude and direction of residue covariability, its origin, and its structural and functional significance. Because associations among separate amino acid sites may arise from several different sources, partitioning these associations into their component sources is fundamental to understanding protein structure, function, and evolution.

The observed covariation in residue composition between amino acid sites *i* and *j* (C_{ij}) arises from several separate underlying causes, which can be expressed by a linear model of the form:

$$C_{ij} = C_{\mathrm{phylogeny}} + C_{\mathrm{structure}} + C_{\mathrm{function}} + C_{\mathrm{interactions}} + C_{\mathrm{stochastic}}.$$

Indeed, the primary null hypothesis to be evaluated for such a model is that any component covariance is equal to zero and, as a consequence, makes no significant contribution to the observed association between amino acid sites *i* and *j*. Let us consider what these various sources of variation entail.

[1] Present address: Fakultät für Mathematik, Universität Bielefeld, Bielefeld, Germany.

Key words: mutual information, protein evolution, entropy, parametric bootstrap, helix-loop-helix.

Address for correspondence and reprints: William R. Atchley, Fakultät für Mathematik, Universität Bielefeld, Postfach 10 01 31, D-33501 Bielefeld, Germany.

Mol. Biol. Evol. 17(1):164–178. 2000

One obvious source of covariation among residues at different sites is common evolutionary history ($C_{\text{phylogeny}}$). Felsenstein (1985) discussed this problem with regard to evolution of complex polygenic traits among species. He pointed out quite elegantly that species are part of a hierarchically structured phylogeny and therefore cannot be regarded for statistical purposes as being drawn independently from the same distribution.

Felsenstein's (1985) argument holds as well for associations among amino acid sites in related proteins. For example, an ancient gene may have undergone early duplications followed by sequence diversification through mutation, natural selection, and genetic drift, which may act differentially in separate evolutionary lineages. The result will be collections of related proteins, e.g., families of bHLH proteins like MyoD and Myc. These families contain a number of functionally and structurally similar proteins that have arisen from a common ancestral protein followed by evolutionary diversification and hierarchical branching (Atchley, Fitch, and Bronner-Fraser 1994; Atchley and Fitch 1995). Within the individual members of such protein families, we would expect to find associations among residues at various amino acid sites that have persisted from the early duplication events.

Additionally, covariation among sites can arise for structural or functional reasons, i.e., $C_{\text{structure}}$ and C_{function}. In this instance, associations among amino acids arise independently of common ancestry and reflect a bias in amino acid replacements in order to satisfy structural demands. The folded nature of a native functioning protein requires that only certain amino acid replacements can occur at particular sites and still maintain the structural integrity of the folded protein. Furthermore, there are constraints on amino acid replacements that arise for functional reasons, such as amino acid bias at recognition sites related to DNA binding in transcriptional regulators. These functional changes may arise when selection operates to optimize adaptation and subsequently generate protein diversification.

Clearly, the main effects in the linear model (i.e., structure, function, and phylogeny) are confounded and therefore are not statistically independent. It is well known that structural and functional changes arise through evolutionary processes. Consequently, inclusion of a covariance term, $C_{\text{interactions}}$, in the model is necessary to account for such higher-order statistical nonindependence.

Finally, covariation among sites may occur that cannot be explained by the main effects in the model and their statistical interaction. This component, designated here $C_{\text{stochastic}}$, refers to the lack of fit of the data to the model and is analogous to the unexplained sum of squares in analysis of variance or regression. For the sake of simplicity, this stochastic effect can be assumed to represent background covariability.

While it is obvious that covariability among sites has a multidimensional basis, partitioning the observed covariability among sites into appropriate underlying components is not a simple matter. Rather, it is a process fraught with many statistical and computational difficulties. Not the least of these difficulties is that biological sequences are represented by symbols that have no natural ordering or underlying metric (Atchley, Terhalle, and Dress 1999). Consequently, conventional statistical analyses typically used to partition variability and covariability are difficult to apply with sequence data.

Herein, we use an entropy (information theoretic) approach coupled with simulation-based parametric bootstrap procedures to examine the magnitude and origin of associations among amino acid sites in the highly conserved basic helix-loop-helix (bHLH) domain. The bHLH domain is a DNA-binding and dimerization domain of approximately 50–60 amino acids found in a large and diverse family of transcription factors (Murre et al. 1994). A number of these proteins have been the focus of detailed structural and functional analyses. Furthermore, the bHLH domain has been the subject of several recent evolutionary analyses (Atchley and Fitch 1997; Atchley, Terhalle, and Dress 1999; Morgenstern and Atchley 1999).

The present paper explores a number of questions about amino acid associations and protein structure. First, we ascertain the magnitude of association or covariation among residues between amino acid sites within the highly conserved bHLH domain. Second, we carry out computer simulations to elucidate the underlying origins of the observed associations among amino acid sites. We inquire if the observed covariability arises simply from stochastic events or if it is due to evolutionary history or structural and functional constraints. Third, we integrate measures of variability and covariability derived from information theory with structural data from published crystal studies on the bHLH domain. In doing so, we explore the relationships between primary sequence diversity and protein structure/function. Fourth, we examine the evolution of the α-helical structure of the bHLH domain among a diverse collection of proteins.

Methods and Materials

Data

The analyses, and conclusions inherent to these analyses and discussions, are based on 242 bHLH domain sequences reported in Atchley and Fitch (1997) and Atchley, Terhalle, and Dress (1999). Multiple-sequence alignment was initially carried out using CLUSTAL W (Thompson, Higgins, and Gibson 1994), and the alignment was then improved by eye. The various phylogenetic analyses, definitions of various clades and evolutionary lineages, descriptions of the protein families, and the like discussed herein are reported in Atchley and Fitch (1997).

Structure of bHLH Proteins

Crystal structure studies have been carried out on the bHLH domains of six proteins, i.e., Max, E47, MyoD, USF, PHO4, and SREBP (Ferre-D'Amare et al. 1993, 1994; Ellenberger et al. 1994; Ma et al. 1994; Brownlie et al. 1997; Shimizu et al. 1997; Parraga et al. 1998). The Max protein, which is the dimerization part-

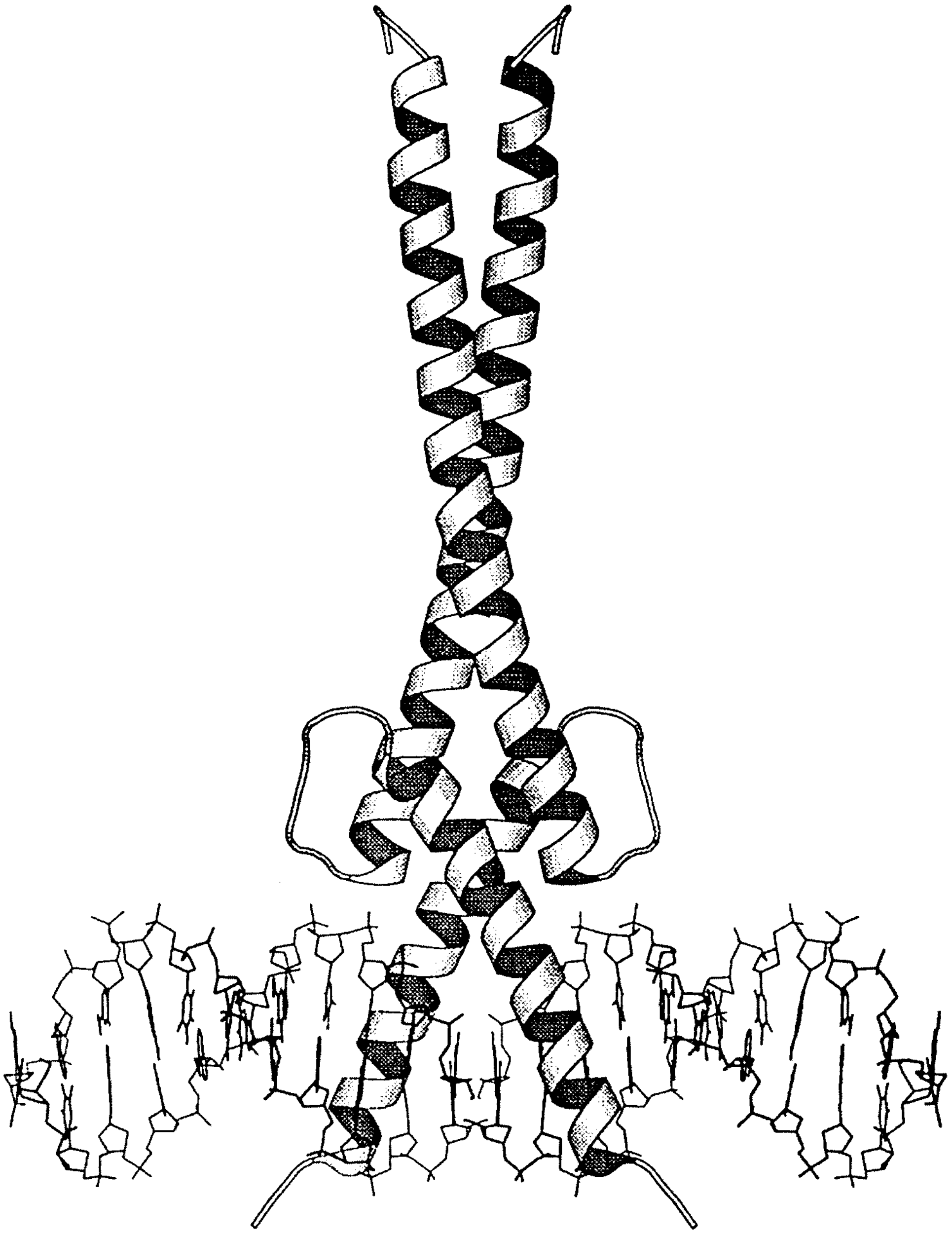


FIG. 1.—Illustration showing the protein homodimer–DNA interaction for Max. This figure (modified from Ferre-D'Amare et al. 1993) shows two α-helices separated by a loop. One α-helix comprises the basic DNA-binding region and helix 1, while the second component involves helix 1 and the leucine zipper. The basic region is shown interacting with DNA in this figure. Of particular relevance to these discussions is the interaction of the two helical components proximal to the loop.

ner of the protooncogene Myc, has been examined in considerable crystallographic detail and shown to have an amphipathic α-helical structure (fig. 1) in which the protein has opposing hydrophobic and hydrophilic faces. The crystal structure of the Max homodimer shows it to be a parallel, left-handed, four-helix bundle with a hydrophobic core (Ferre-D'Amare et al. 1993).

The conserved hydrophobic amino acids from helix 1 (H1) and helix 2 (H2) in Max are buried in the interior of the four-helix bundle, where they pack together and exhibit strong van der Waals interactions that stabilize the structure of the homodimer. The conserved, hydrophobic amino acids in H1 and H2 appear to be required for a stable protein–DNA complex formation in Max. The crystal structure of Max appears to be generally similar to that of other bHLH proteins, such as E47 (Ellenberger et al. 1994), MyoD (Ma et al. 1994), USF (Ferre-D'Amare et al. 1994), PHO4 (Shimizu et al. 1997, and SREBP (Parraga et al. 1998). Indeed, Parraga et al. (1998) point out the ''remarkable similarity'' in structure between SREBP and Max and that the hydrophobic cores are ''virtually identical'' in these two distantly related bHLH proteins.

The protein Max will be used as the structural model for our discussions, and it will be assumed that the 242 bHLH proteins involved in these analyses have the same general structural features as Max. This extrapolation is based on the studies of Ferre-D'Amare et al. (1993, 1994), Ma et al. (1994), Ellenberger et al. (1994), Shimizu et al. (1997), and Parraga et al. (1998).

Helical Wheel Projections

Helical wheel projections are used in these analyses to provide insight into residue interrelationships with a protein structure. Helical wheels graphically display the disposition of amino acid side chains about an assumed α-helix. The projection is along the central axis of the helix, from the N-terminus to the C-terminus, and it is a useful device for displaying the symmetry (or asymmetry) of hydrophobic/hydrophilic side chains. The helical wheel assumes a periodicity of 3.6 residues per helical turn.

Thirty-three exemplar sequences are used to explore the phylogenetic aspects of the helical wheel projections. Generally speaking, these 33 sequences are well-studied proteins that reflect the evolutionary diversity of the bHLH domain. The evolutionary relationships among the various clades and lineages for these sequences are represented by a neighbor-joining tree. Sequences are arranged phylogenetically and shown in a helical wheel configuration for helices 1 and 2.

Secondary Structure Prediction

In several instances, the secondary structure of a particular bHLH-domain-containing protein is examined using the Protein Sequence Analysis (PSA) Server from Boston University. The computer model analyzes amino acid sequences and calculates the probability of secondary structures and folding classes within regions of a sequence. The underlying theory for these predictions is described in White, Stultz, and Smith (1994), and the URL of the server is http://bmerc-www.bu.edu/psa/.

Variability and Covariability in Protein Sequences

As noted earlier, statistical analyses of biological sequences present difficulties because these sequences are represented by symbols that have no natural ordering or underlying metric (Atchley, Terhalle, and Dress 1999). Consequently, conventional statistical estimates of variability and covariability are difficult to apply. Recently, several authors have suggested the use of the concepts of entropy and mutual information (Korber et al. 1993; Clarke 1995; Herzel and Gross 1995; Schneider 1996; Roman-Roldan, Bernaola-Gavan, and Oliver 1996; Atchley, Terhalle, and Dress, 1999).

Entropy (E) is a measure of uncertainty derived from thermodynamics and statistical physics which has considerable utility for studies of protein structure. Assume X is a discrete random variable (the amino acid sites) for which we are uncertain which of its 20 values ($x_1, x_2, \ldots, x_{20}$) (amino acid residues) will occur at site X, but we do know their expected frequencies, $p_i, \ldots, p_n$. These expected frequencies can be used to calculate how much information $E(X)$ is present at site X. In this context, information is a measure of the uncertainty about which residue will occur at a specified site.

The Boltzmann-Shannon entropy $E(X)$ is defined (Applebaum 1996) by

$$E(X) := -\sum_{j=1}^{n} p_j \log_2(p_j), \tag{1}$$

where $n = 20$, p_j is the probability of an amino acid being of the jth kind, and $p_j \log_2 p_j := 0$ if $p_j = 0$. $E = 0$ when all elements are in the same category (the same amino acid residue at a particular site). E increases with both the number of categories (residues at a site) and their equiprobability. Entropy of a uniform distribution whose range has size n is

$$E_n = \log_2(n). \tag{2}$$

Thus, the minimum entropy or uncertainty value will be zero when only a single residue occurs at a particular site in all included proteins. The expected maximum entropy will be 4.32 when all 20 residues are present in equal frequencies at a given site.

The relative information content of Y contained in X is termed the mutual information, or MI(X, Y), where

$$MI(X, Y) = \sum_{j=1}^{n} \sum_{k=1}^{m} p_{jk} \log_2 \frac{p_{jk}}{p_j q_k}. \tag{3}$$

Note that MI(X, Y) = MI(Y, X), and if X and Y are independent, then MI(X, Y) = 0, corresponding to the fact that no information is obtained regarding Y by finding out about X. In biological sequences, MI describes the extent of ‘‘correlation’’ or association between residues at amino acid sites X and Y that might arise from evolutionary, functional, or structural constraints. More algebraic details are provided in Atchley, Terhalle, and Dress (1999).

Statistical Inference about Mutual Information Values

An important question in biological sequence analyses is whether one can distinguish signals due to various biological sources (phylogeny, structure, and function) from any background noise (stochastic variation) inherent in a set of sequences. This is analogous, in quantitative genetics, to partitioning phenotypic variability into genetic components (including additive, dominance, and epistatic variance components) and environmental components.

In these analyses, we use a parametric bootstrap approach (Efron and Tibshirani 1993; Goldman 1993; Huelsenbeck, Hillis, and Jones 1996) to generate a distribution of MI values reflecting only covariation involving stochastic and phylogenetic constraints. Additional details of this method are presented in Wollenberg and Atchley (2000). The parameters used in the parametric bootstrap simulations were the phylogenetic tree generated from the aligned protein sequences and a residue substitution matrix. Because the tree was derived from the data and the substitution matrix was not (it was chosen to reflect general amino acid substitution probabilities), the data sets generated in the parametric bootstrap simulations contained only stochastic and phylogenetic associations between sites.

A neighbor-joining tree (Saitou and Nei 1987) was computed for the 237 sequences using p-distances. The residue change matrix used was that used in the computer program PAML, version 1.3 (Yang 1997). This matrix was generated using the algorithm of Jones, Taylor, and Thornton (1992) and is hereinafter referred to as the JTT matrix. This matrix does not consider gaps as characters for the generation of replicate data sets. Therefore, for MI values calculated on the empirical data to be comparable with MI values calculated on the parametrically generated data sets, only ungapped sites could be used for this statistical analysis. (All sites were used for generating the phylogeny.) For this reason, the original 242 sequences of Atchley and Fitch (1997) were reduced to 237 to decrease the number of sites in the alignment having gaps. This resulted in 32 sites without gaps for analysis.

Like any numerical simulation of a physical process, the results depend on the assumptions of the underlying models for their validity. As in any phylogenetic analysis, results depend on the confidence one has that the tree is a realistic description of the history of the subjects being analyzed. The parametric bootstrap also depends on the tree as the source of information about the level and distribution of sequence variation. The residue substitution matrix used will control the changes that occur between sequences in the simulation. Biases in this matrix can affect the potential associations measured in the resulting simulated sequences. However, a matrix having no biases (i.e., a matrix of uniform substitution probabilities) would ignore the biology of the substitution process. Alternatively, one could use a substitution matrix derived from the empirical data, such as that calculated by the RIND program (Bruno 1996). However, a matrix of this type would reflect biases due to phylogeny, structure, and function that are inherent in the empirical data being analyzed (Wollenberg and Atchley 2000). For these reasons, we used a general protein substitution matrix derived using the JTT algorithm.

The statistical significance of the MI values was determined by comparing the frequency distributions of MI values for the 237 bHLH sequences and the results for the parametric bootstrap analyses. Any MI value above a specific threshold was considered to contain significant associations over and above those due to stochastic or phylogenetic constraints. This threshold MI value was that value in the frequency distribution of parametric bootstrap MI values greater than a specified percentage (i.e., 99%, 99.9%) of parametric bootstrap MI values. Thus, this procedure does not test whether a given MI value is different from random; rather, it tests whether an MI value reflects a significant association due to structural and functional constraints over and above covariation arising from evolutionary history and stochastic events.

Results

Entropy Values

Figure 2 provides a histogram of the entropy values (*E*) for the individual amino acid sites in the bHLH domain for 242 proteins. The basic region extends from site 1 to site 13, helix 1 extends from site 14 to site 28, and helix 2 extends from site 50 to site 64. The basic and helix 1 regions are modeled as a continuous α helix (Ferre-D'Amare et al. 1993) separated from the second helix by a variable-length loop. The loop is not considered in these discussions. Actual numerical values of *E* range from 0.15 (98% L residues) at site 23 to the maximum observed value of 3.45 for highly variable sites 21 and 62. As noted earlier, the maximum possible entropy value will be 4.32 when all 20 residues are present at a site in equal frequencies.

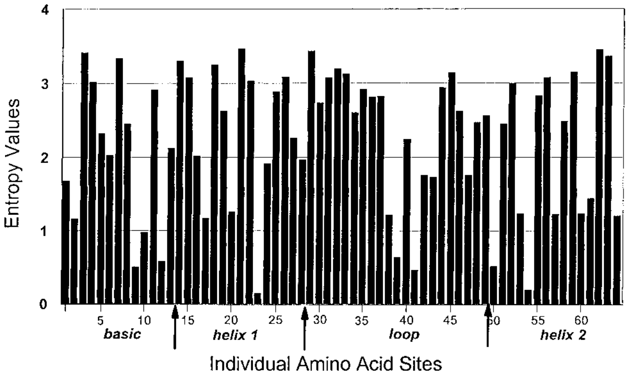


FIG. 2.—Histogram of entropy values for individual amino acid sites in the bHLH domain.

Crystal studies of the protein Max by Ferre-D'Amare et al. (1993) have shown that residues at various sites pack against each other. These sites are described herein as ‘‘contact sites,’’ using the numerical site identification described by Atchley and Fitch (1997) and Atchley, Terhalle, and Dress (1999). Thus, K1 refers to a K residue at site 1. With regard to the contact sites, I16 packs against F20, which contacts R50, I53, and L54. Site L23 abuts I53 and A57. V27 and P28 pack against Y60, and I61 packs against its symmetry mate (the site in the dimerization partner) I61′ and Y60′, while M64 interacts with its homodimer ‘‘symmetry mate’’ M64′. These packed residues, together with the relevant van der Waals forces, stabilize the structure of the homodimer and help conserve the hydrophobic core.

The analyses described here assume the crystal structure of Max, E47, MyoD, USF, PHO4, and SREBP, extrapolated to the other known bHLH proteins. Based on the multiple alignment of the bHLH domain sequences used in Atchley and Fitch (1997) and Atchley, Terhalle, and Dress (1999), the most prevalent residues at these contact or packing sites are 16 (I, L, V), 20 (F, I, L), 23 (L), 27 (I, L, V), and 28 (L) in helix 1, and 50 (K), 53 (I, T, V), 57 (A), 60 (Y), and 64 (L) in helix 2. Thus, the five relevant contact sites in helix 1 are highly hydrophobic while the five sites in helix 2 (except for site 50, which initiates the second helix) are predominantly hydrophobic. The packing interrelationships among sites in helix 1 and helix 2 are shown graphically in figure 3. Additionally, structural studies on SREBP (Parraga et al. 1998) also indicate interactions between site 12 in the basic region with site 17 in helix 1 and site 50 in helix 2 in the other monomer of the homodimer.

All known core contact positions given by Ferre-D'Amare et al. (1993) and Parraga et al. (1998) have

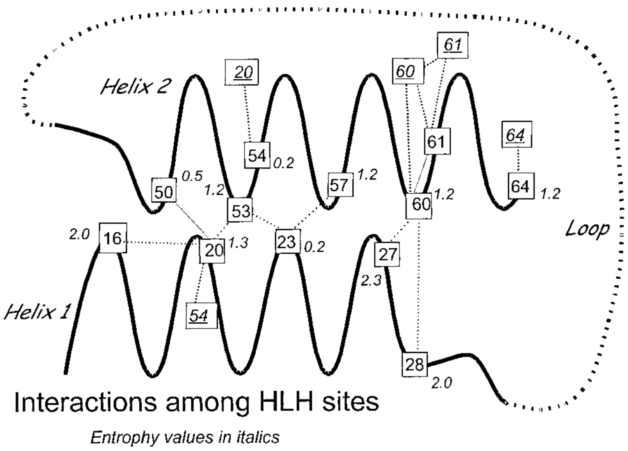


FIG. 3.—Interactions (packing) between helix 1 and helix 2 sites in bHLH proteins. The dotted lines between numbered boxes refer to those sites known to pack together. The entropy values for each site are given in italics. Sites that are symmetry mates with the dimerization partner are underlined.

Table 1
Entropy Values for Amino Acid Sites in Helix 1 and Helix 2 that Are in Contact or Not in Contact

Sites in Contact			Sites Not in Contact		
Site	Entropy	E_F	Site	Entropy	E_F
23	0.15	0.03	51	2.46	0.58
54	0.20	0.07	58	2.49	1.29
50	0.52	0.32	19	2.63	1.94
17*	1.18	0.93	55	2.84	1.66
64	1.21	0.37	25	2.90	1.95
57	1.23	0.82	52	3.00	1.81
53	1.24	0.66	22	3.03	2.16
60	1.25	1.31	56	3.08	2.14
20	1.27	1.00	15	3.08	1.98
61	1.46	0.11	26	3.09	2.10
24*	1.93	0.94	59	3.16	2.28
28	1.98	1.64	18	3.26	2.15
16	2.03	0.30	14	3.31	2.39
27	2.27	0.97	21	3.48	2.36
Mean . . .	1.28	0.68	Mean . . .	2.98	1.91
SD	0.65	0.49	SD	0.30	0.48

NOTE.—The "Entropy" columns give the entropy values for individual amino acids at each site. E_F is the entropy value when the amino acids at each site are converted to their functional groups (*sensu* Atchley, Terhalle, and Dress 1999). Sites 17 and 24 are putative contact sites based on their entropy values (see text for more details). A Mann-Whitney test of the null hypothesis that these two sets of sites have the same median entropy value is rejected at $P <$ 0.001 for both E and E_F.

entropy values which range from 0.15 to 2.27 (table 1). All of the remaining amino acid positions except two (described below) have entropy values ranging from 2.46 to 3.48. The distributions of E values for contact versus noncontact values are nonoverlapping, and their medians are statistically highly significantly different by a Mann-Whitney test (Sokal and Rohlf 1995) (table 1). One might conclude that contact residues can be identified by their entropy values, i.e., assigning contact positions to those with E values below a particular threshold value (approximately 2.3 in the case of these bHLH domains). While this works for the bHLH domain, it is unknown whether it can be generalized to other proteins.

If the various amino acid residues are classified into functional groups (e.g., D, E = acidic; K, R, H = basic) as described in Atchley, Terhalle, and Dress (1999), the contact sites have entropy values ranging from 0.03 to 1.64. All values are $<$1.0 except those of site 60 (1.31) and site 28 (1.64). For sites not in contact, the entropy values for functional groups range from 0.58 to 2.39. The noncontact site with the smallest entropy value (site 51, with $E = 0.58$) has a variety of residues, but they are all aliphatic/hydrophobic, which accounts for this seemingly anomalously low value.

A. R. Ferre-D'Amare generously provided information from his crystal studies about the structural interactions of site 51. The side chains of the alanine residue in Max and the leucine residue in E47 at site 51 interact with those of site 16 to facilitate stabilization of the protein dimers' hydrophobic core. Furthermore, in SREBP, the serine at site 51 makes a water-mediated hydrogen bond with a phosphate oxygen anchoring the dimer to DNA. Other bHLH proteins have histidine residues at site 51, and this side chain could make DNA contacts as well.

Our results suggest that two positions not originally considered contact sites by Ferre-D'Amare et al. (1993) may be such. Sites 17 and 24 have E values of 1.18 and 1.93, respectively (0.93 and 0.94 with residues classified into functional groups), which are well within the range of the E values for contact sites (table 1). Further examination of the bHLH structural data (A. R. Ferre D'Amare, personal communication) suggests that site 17 functions in a water-mediated DNA-protein contact. This interrelationship has apparently resulted in a high level of conservation of hydrophilic residues at this site (74% asparagine, 24% lysine or arginine). This interaction is clearly demonstrated in the sterol-regulatory-element-binding proteins (SREBPs) by Parraga et al. (1998).

Unfortunately, the situation with site 24 is more difficult to explain. Side chains in this position form part of an extension of the hydrophobic core of the bHLH dimer by burying under the flap formed by the loop component. The conservation of longer side-chain residues at site 24 may stem from their hydrophobic methylenes being partially buried and the hydrophilic tips being exposed (A. R. Ferre-D'Amare, personal communication). Additional high-resolution studies of the bHLH domain may provide a biological explanation for these quantitative observations on site 24.

Variability in Buried and Exposed Sites

The results reported here are in line with many previous observations that internal (buried) residues are less variable than external or exposed residues, and less variation means lower entropy values (e.g., Goldman, Thorne, and Jones 1998). Indeed, Atchley, Terhalle, and Dress (1999) previously tested this hypothesis and found that the entropy values of the buried sites are significantly smaller than those of the exposed sites.

Mutual Information

Tables 2 and 3 provide MI values describing the level of association among amino acid sites within the

Table 2
Mutual Information (MI) Values >1.0 Within and Between Components for Basic, Helix 1, and Helix 2 Sites

Basic region			Helix 1			Helix 2		
3	4	1.17	14	15	1.17	52	55	1.22
3	7	1.09	14	19	1.05	52	56	1.15
3	14	1.18	14	21	1.26	52	62	1.07
3	15	1.01	14	25	1.05	55	56	1.04
3	21	1.20	14	26	1.15	56	62	1.05
3	26	1.08	14	49	1.01			
3	52	1.08	14	51	1.01			
3	56	1.15	14	52	1.14			
3	62	1.13	14	55	1.11			
4	5	1.00	14	56	1.18			
4	7	1.11	14	62	1.16			
4	8	1.01	14	63	1.02			
4	14	1.08	15	21	1.19			
4	18	1.02	15	26	1.02			
4	21	1.10	15	52	1.03			
4	25	1.05	15	62	1.07			
4	52	1.10	18	21	1.05			
4	56	1.06	19	21	1.02			
4	62	1.07	21	25	1.05			
5	14	1.07	21	26	1.13			
5	52	1.01	21	52	1.20			
7	11	1.11	21	55	1.19			
7	14	1.09	21	56	1.19			
7	15	1.12	21	62	1.21			
7	56	1.04	25	56	1.05			
7	62	1.08	26	52	1.03			
8	14	1.09	26	56	1.02			
8	19	1.12	26	62	1.10			
8	21	1.00						
8	25	1.03						
8	52	1.05						
11	15	1.05						
11	21	1.10						
11	55	1.06						
11	62	1.08						

NOTE.—Only values >1.0, which reflect the top 5% of all MI values for the bHLH domain, are included. Basic region sites are 1–13, helix 1 sites are 14–28, and helix 2 sites are 50–64. Sites from the loop region are not included.

bHLH domain. Values >1.0 reported in table 2 constitute the top 5% of MI values for all bHLH domain sites as reported by Atchley, Terhalle, and Dress (1999). When the sites with MI > 1.0 are arranged in a network, specific patterns of association are made apparent (fig. 4). Within this network are subnetworks consisting of sets of sites for which each site has connections to all other sites in the subnetwork. These completely connected subnetworks correspond to the cliques previously defined in Atchley, Terhalle, and Dress (1999). A maximum clique corresponds to the largest maximally connected subnetwork. The two maximum cliques for the data from table 2 are presented in figure 4*A* and *B*.

Mutual Information Between the Basic Region and Helices 1 and 2

Table 2 describes the pattern in mutual information values between the basic DNA-binding region (sites 1–13) and helices 1 and 2. Within the basic component itself, MI values >1.0 are noted between sites (3, 4), (3, 7), (4, 5) (4, 7), (4, 8), and (7, 11).

Sites 3, 4, 7, and 8 in the basic region show high levels of association with specific sites in the two heli-

Table 3
Entropy (*E*) and Mutual Information (MI) Values for Sites in Helix 1 (H1) (Sites 14–28) and Helix 2 (H2) Sites (50–64)

		H2 SITES														
		50	51	52	***53***	54	55	56	***57***	58	59	***60***	61	62	63	64
H1	*E*	(*E* = 0.52)	(*E* = 2.46)	(*E* = 3.00)	(*E* = 1.24)	(*E* = 0.20)	(*E* = 2.84)	(*E* = 3.08)	(*E* = 1.23)	(*E* = 2.49)	(*E* = 3.16)	(*E* = 1.25)	(*E* = 1.46)	(*E* = 3.45)	(*E* = 3.36)	(*E* = 1.21)
14	3.31	0.17	**1.01**	**1.14**	0.29	0.03	**1.11**	**1.18**	0.47	0.95	0.83	0.55	0.47	**1.16**	**1.02**	0.34
15	3.08	0.20	0.88	**1.03**	0.42	0.03	0.95	0.93	0.29	0.92	0.69	0.42	0.38	**1.07**	0.95	0.40
16	2.03	0.11	0.45	0.73	0.10	0.06	0.67	0.60	0.21	0.43	0.44	0.22	0.26	0.58	0.46	0.19
17*	1.18	0.11	0.18	0.53	0.11	0.03	0.45	0.44	0.09	0.43	0.27	0.13	0.22	0.41	0.25	0.30
18	3.26	0.16	0.55	0.80	0.24	0.04	0.67	0.87	0.23	0.74	0.83	0.29	0.35	0.88	0.79	0.43
19	2.63	0.15	0.63	0.98	0.36	0.06	0.94	0.93	0.40	0.66	0.67	0.33	0.49	0.80	0.76	0.36
20	1.27	0.08	0.40	0.41	0.09	0.02	0.42	0.34	0.26	0.35	0.33	0.20	0.16	0.33	0.37	0.12
21	3.48	0.27	0.91	**1.19**	0.33	0.07	**1.19**	**1.19**	0.36	0.94	0.86	0.45	0.47	**1.21**	**1.11**	0.46
22	3.03	0.19	0.71	0.76	0.17	0.04	0.77	0.82	0.29	0.59	0.61	0.30	0.26	0.79	0.71	0.28
23	0.15	0.00	0.03	0.03	0.01	0.02	0.02	0.03	0.02	0.03	0.04	0.00	0.02	0.06	0.04	0.02
24*	1.93	0.11	0.60	0.77	0.23	0.05	0.64	0.62	0.23	0.65	0.37	0.41	0.25	0.60	0.45	0.21
25	2.90	0.16	0.78	0.98	0.37	0.04	0.81	**1.05**	0.37	0.70	0.73	0.37	0.43	0.94	0.83	0.34
26	3.09	0.14	0.83	**1.03**	0.25	0.03	0.94	**1.01**	0.28	0.80	0.86	0.50	0.38	**1.10**	0.86	0.36
27	2.27	0.07	0.50	0.60	0.14	0.03	0.56	0.77	0.16	0.52	0.45	0.39	0.21	0.71	0.49	0.20
28	1.98	0.05	0.47	0.65	0.17	0.01	0.63	0.78	0.13	0.41	0.54	0.43	0.22	0.61	0.54	0.17

NOTE.—Sites known to interact in the protein Max are underlined and in bold italic type. Two sites predicted from these analyses to interact are indicated with asterisks. MI values >0.56 reflect sites with high probabilities (*P* > 99%) of having associations due to structural and functional constraints. Values >1.0 represent the top 5% of all bHLH MI values and are given in bold type and underlined.

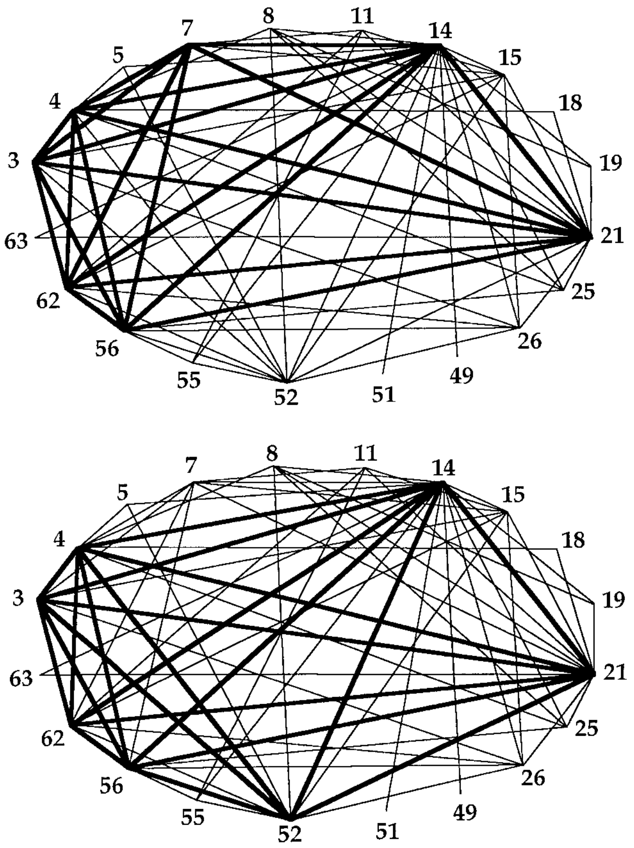


FIG. 4.—Network diagrams of the data from table 2 (pairs of sites with MI > 1.0). Maximum cliques are indicated by the heavy black lines. *A*, Maximum cliques containing site 7 from the basic region. *B*, Maximum clique containing site 52 from helix 2.

ces, primarily sites 14 and 21 in helix 1 and sites 52, 56, and 62 in helix 2. The helical wheel assumes a periodicity of 3.6 residues per helical turn, so site 14 is at the start of helix 1, and site 21 is two complete turns into the helix. In rank order, the MI values for the top 13 paired sites are as follows: 3 and 21 (1.20), 7 and 21 (1.19), 3 and 14 (1.18), 3 and 4 (1.17), 3 and 56 (1.15), 3 and 62 (1.13), 7 and 15 (1.12), 8 and 19 (1.12), 4 and 7 (1.11), 7 and 11 (1.10) 4 and 21 (1.10), 4 and 52 (1.10), and 11 and 21 (1.10).

Mutual Information Within and Between Helix 1 and Helix 2

Table 3 provides the *E* and MI values for interacting sites in helices 1 and 2. Packed or contact sites are underlined and in italics in table 3, and entropy values are provided for each site. These values provide a description of residue diversity at each site, ranging from $E = 0.15$ at site 23 (which is 98% leucine in this large database) to $E = 3.48$ at site 21. The maximum possible value for *E* is 4.32.

The remainder of the table provides pairwise MI values for these amino acid sites. Sites in this subset with MI values >1.0 are shown in bold type and underlined. As a point of reference, the largest observed MI value for the entire bHLH domain was 1.26 between sites 14 and 21.

From Table 3, it is clear that sites with low sequence diversity (small entropy values) also show little covariation with other amino acid sites. This is to be expected, since residues at paired sites do not covary if the individual sites themselves had little residue variation. Thus, the four primary sites in each helix previously shown to pack together (sites 16, 20, 23, and 27 in helix 1, and sites 50, 53, 57, and 60 in helix 2) exhibit very little residue variability (low entropy values), and there is very little covariability among the contact sites. The latter finding is reflected by the fact there are no MI values among these eight contact residues higher than 0.22 and, as will be seen below, none show significant covariation due to structural and functional constraints.

Several sites within the helices with higher residue diversity exhibit considerable mutual information with other variable sites. Thus, site 23 ($E = 3.48$) in helix 1 exhibits the highest observed MI values ($\mathrm{MI} = 1.19$) with sites 52, 55, and 56 in helix 2. In helix 2, site 52 likewise shows high MI values with sites 19, 21, 25, and 26 in helix 1.

Note, however, that while site 18 exhibits considerable residue diversity ($E = 3.26$), it does not necessarily exhibit high MI values with other sites. This demonstrates that high entropy does not necessarily produce high values of mutual information even in highly conserved protein domains.

Origins of Significant Mutual Information

There are several possible explanations for significant levels of association among many sites within the bHLH domain (other than simply chance associations). These include associations arising from evolutionary constraints, correlated mutations, functional associations, and structural constraints.

Simulation and the Partition of Observed Associations

A simulation was carried out using parametric bootstrap procedures to partition the observed covariation among amino acid sites into that due to evolutionary history (phylogeny) and stochastic events on the one hand, and that due to structural and functional constraints on the other. The distributions of MI values for the parametric bootstrap data and the empirical data were significantly different at $P < 0.001$. This suggests that there are significant associations among many amino acid sites in the bHLH domain over and above those due to stochastic and phylogenetic effects.

The distribution of MI values from 1,000 parametric bootstrap replicates, calculated using the neighbor-joining tree and the JTT substitution matrix, was compared with the distribution of MI values for 237 bHLH proteins (fig. 5). This comparison permits calculation of threshold values for distinguishing between structural/functional and phylogenetic/chance associations. In these analyses, sites having MI values above a given threshold value have a specific probability of covariation due to structural and functional constraints, rather than due to phylogenetic constraint or chance. The specific MI value used as this threshold has an associated prob-

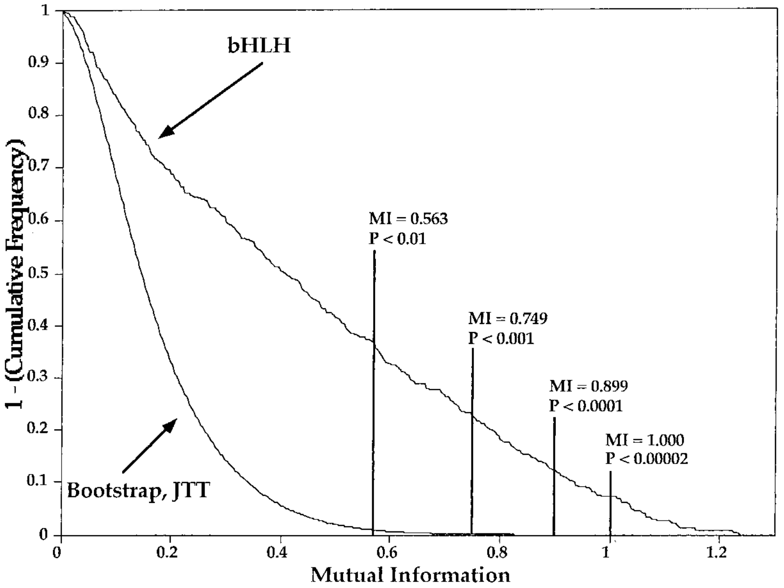


FIG. 5.—Inverse cumulative frequency distributions of *MI* values for the alignment of 237 bHLH protein sequences and 1,000 parametric bootstrap replicates using the JTT substitution matrix. Three probability thresholds ($P < 0.01$, $P < 0.001$, and $P < 0.0001$), their associated MI values, and the the probability associated with MI = 1.000 are indicated. These thresholds were calculated from the JTT bootstrap distribution. Because MI is a pairwise measure, each replicate consisted of $x(x - 1)/2$ values, where x is the number of sites in the alignment. For the alignment of 237 bHLH sequences, there were 32 sites without gaps, resulting in 496 MI values per replicate.

ability based on the number of values in the parametric bootstrap distribution that are greater than the threshold.

Based on the frequency distribution shown in figure 5, any pair of nongapped sites in the bHLH alignment that have MI values >0.563 will have a probability <0.01 that the covariation among sites is due to phylogeny or chance associations. With a threshold value of 0.749, the probability of this level of covariation being due to phylogeny or chance is reduced by an order of magnitude to $P < 0.001$. In table 2, sites are designated that have MI values >1.0 and it is noted that they reflect the top 5% of all MI values for the entire bHLH domain. These values >1.0 have a probability of $P <$ 0.00002 (from the parametric bootstrap simulations) that the associations are due to phylogeny or chance associations.

Among the contact sites, there is no significant covariation from structural and functional constraints. However, there is considerable significant covariation among noncontact sites that stems from structural and functional constraints (fig. 6). The latter is evident in those MI values larger than 0.56, the 1% critical value from the parametric bootstrap simulations.

Correlations in Hydropathy

From a structural and functional perspective, correlated amino acid replacements may have occurred among sites to maintain the same hydropathic relationships, similar electrostatic charges, size constraints, etc. It is difficult to examine statistical associations among sites with regard to electrostatic charge, since only five amino acids (K, R, H, D, and E) are charged and 15 residues have no charge. Since paired sites must show variability to exhibit correlation, attempts to compute correlation coefficients with invariant data (charge of zero in both variables) is undefined, as can be seen by equation (3). Consequently, we are unable to examine associations due to charge and have focused instead on the possibility of significant associations due to hydropathy and size constraints.

For 33 exemplar sequences, residues in helix 1 and helix 2 were coded as −1 if they were hydrophobic (A, C, G, I, L, M, F, P, and V), 1 if they were hydrophilic (R, N, D, E, Q, H, and K), and 0 if they were S, T, Y, or W. Then, pairwise product-moment correlation coefficients (r) were computed among all sites in the two helices to determine if significant associations existed among sites for hydropathy states. Several pairs of sites showed high correlations, including sites 20 and 23 ($r = 0.89$), 17 and 61 ($r = -0.81$), 50 and 61 ($r = -0.75$), and 27 and 61 ($r = 0.71$). (Any product-moment correlation >0.45 in this analysis is significant at $P < 0.01$) As can be seen in the data given in figure 7, the high positive correlations relate to strong association for hydrophobic residues at sites 20 and 23, as well as at sites 27 and 61 (fig. 8). The negative values refer to inverse relationships between hydrophobic and hydrophilic residues. In addition to these four pairs of sites, two pairs of sites (sites 23 and 27 and sites 25 and 50) had correlations >0.6.

Correlations in the Sizes of Amino Acid Residues

For these same 33 bHLH sequences, product-moment correlation coefficients were computed for the sizes (volume) of residues at pairs of sites for helices 1 and 2. Seven pairwise correlations occurred that were statistically significant. These included sites 23 and 50 ($r = 0.66$), 50 and 60 ($r = 0.64$), 53 and 54 ($r = -0.57$), 22 and 26 ($r = 0.57$), 23 and 60 ($r = 0.52$), 54 and 60 ($r = -0.51$), and 56 and 61 ($r = 0.50$). However, some

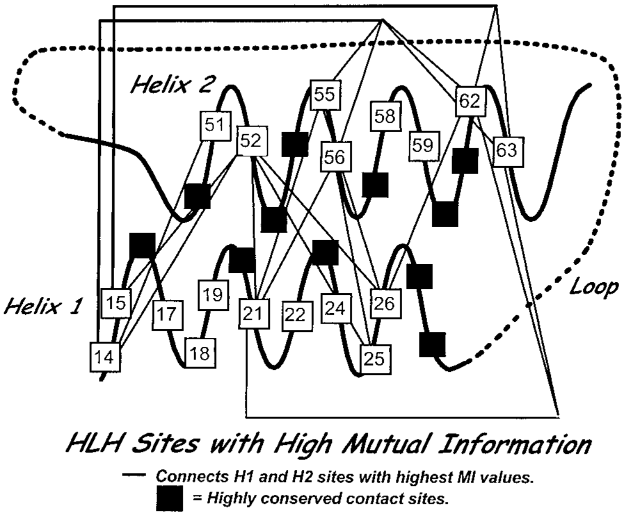

Helix 2
Helix 1
Loop
51
52
55
56
58
59
62
63
15
17
19
18
14
21
22
24
26
25
HLH Sites with High Mutual Information
— Connects H1 and H2 sites with highest MI values.
■ = Highly conserved contact sites.

FIG. 6.—Representation of high mutual information values among a subset of amino acid sites in bHLH proteins. The solid boxes are the conserved sites known to pack together (see fig. 1) from crystal structure studies. The lines between sites in helix 1 and helix 2 reflect the 12 sites with the highest mutual information (>0.9).

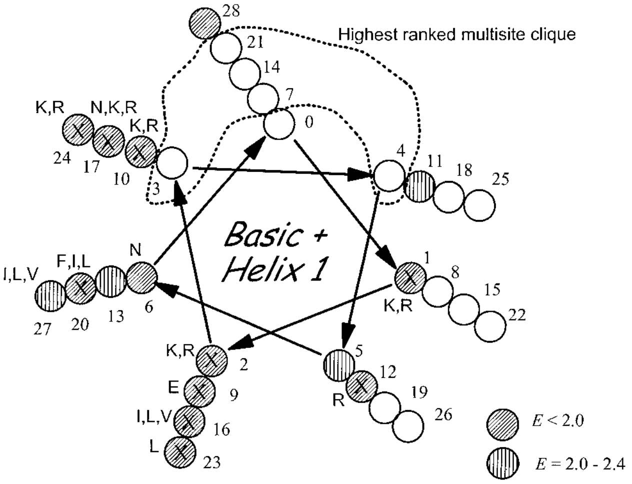


FIG. 7.—A helical wheel drawing of the basic DNA-binding region and helix 1. Highly conserved sites are shaded, and those sites involved in the predictive motif of Atchley, Terhalle, and Dress (1999) are denoted with an **X**. The five sites involved in the highest-ranked multisite clique are enclosed with a dotted line.

values must be viewed with caution because of computational difficulties associated with high levels of conservation (low entropy) and different residues with equivalent volumes.

In spite of these caveats, there are several interesting findings here involving associations based on size. The sites with the largest correlation coefficient (sites 20 and 53) are contact sites, and these observations indicate that there is a like association; i.e., large residues are paired with large residues. Thus, when F occurs at site 20, I, L, Y, or T occurs at site 53. Similarly, an L at site 20 pairs with an L, T, or V at site 53.

The largest negative value occurs between two adjacent sites (sites 53 and 54). It might be expected that significant inverse relationships would exist for the volume of adjacent residues. However, the difference between the entropy values for sites 53 (E = 1.24) and 54 (E = 0.20) stresses the need for caution in interpreting this correlation, since variable site 53 is paired with a largely unvaried site 54.

The correlation coefficient of 0.57 between sites 22 and 26 probably represents a meaningful structural/functional association, because there is considerable residue variability at both of these sites. Both sites occur away from the contact sites and, consequently, exhibit more variability.

Helical Wheels and Sequence Diversity

The helical wheel shown in figure 7 displays the helical distribution of residues including both the basic DNA-binding region and helix 1, while figure 8 provides a helical wheel for amino acid sites 50–64, which constitute helix 2. These figures summarize information for the 392 bHLH domain sequences in the database. The most prevalent residues at each site are shown in the figure. Amino acid cliques (*sensu* Atchley, Terhalle, and Dress 1999) shown in these figures are defined for each helix. Cliques are groups of amino acid positions all of which are more highly associated with each other than any are with a nonmember of the clique. Maximal cliques are those not contained in larger cliques. Finally, those sites involved in determining the predictive motif described in Atchley, Terhalle, and Dress (1999) are marked by an **X**. This predictive motif is a collection of 19 highly conserved sites whose amino acid compositions accurately discriminate bHLH-domain-containing proteins into groups A–D according to the evolutionary classification proposed by Atchley and Fitch (1997).

In Figure 7, sites are denoted that have entropy values <2.0, values between 2.0 and 2.4, and values >2.4. The most prevalent amino acid residues at each site are noted with the appropriate symbols where possible. Furthermore, five sites (sites 3, 4, 7, 14, and 21) that constitute the highest-ranked multisite clique in helix 1 are denoted.

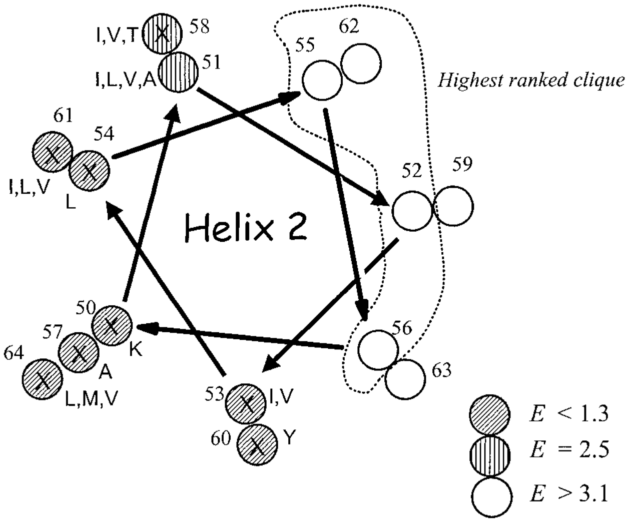
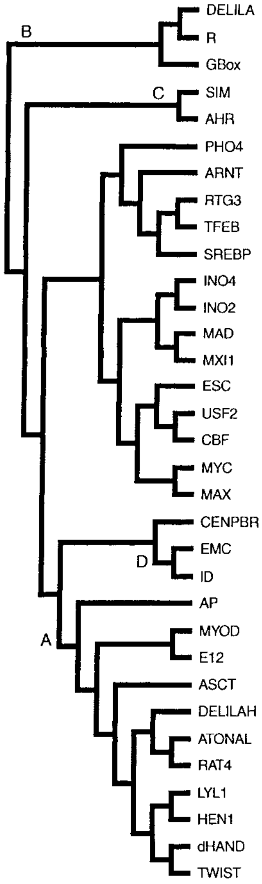


FIG. 8.—Helical wheel of helix 2 sites. Labeling conventions follow figure 7.



	HELIX 1															HELIX 2														
	14	21	28	15	22	16	23	17	24	18	25	19	26	20	27	50	57	64	51	58	52	59	53	60	54	61	55	62	56	63
DELILA	E	M	P	K	I	I	L	N	A	E	S	R	L	F	V	K	T	L	V	I	S	D	I	Y	L	L	D	R	H	G
R	E	L	P	K	V	L	L	N	K	E	S	M	L	F	L	K	T	L	A	I	S	A	I	Y	L	L	A	K	E	E
GBox	E	Y	P	K	A	L	L	N	R	Q	A	R	V	F	V	K	A	L	A	I	S	S	L	Y	L	I	G	N	D	E
SIM	E	Y	P	K	E	E	L	N	A	S	K	E	L	F	L	K	T	M	A	T	S	S	I	Y	I	L	R	K	L	-
AHR	D	D	P	R	R	L	L	N	A	T	S	E	L	L	L	K	S	K	L	V	S	S	V	Y	L	L	R	R	L	A
PHO4	N	H	P	R	E	L	L	A	A	V	S	P	L	L	I	K	A	L	A	C	T	R	T	Y	E	I	E	R	A	H
ARNT	N	T	P	K	E	M	L	T	S	A	D	Y	M	I	V	K	A	L	L	V	T	S	I	H	L	M	R	K	M	S
RTG3	E	K	P	L	E	I	L	K	G	Q	Q	K	L	I	V	K	T	L	G	V	I	E	I	Y	L	L	D	Q	R	Y
TFEB	F	K	P	N	E	I	L	N	G	D	M	R	L	I	I	K	S	M	G	V	T	D	I	Y	L	I	K	R	A	R
SREBP	S	I	V	S	E	I	L	N	K	D	D	K	L	I	V	K	A	L	S	I	A	D	V	Y	L	I	R	R	K	F
INO4	E	D	P	L	E	E	L	R	V	A	A	I	V	F	V	L	L	Y	I	S	I	Y	Y	L	L	S	K	W	S	L
INO2	I	E	R	N	R	T	L	K	I	E	K	A	S	F	V	K	V	I	H	M	I	N	L	D	L	I	T	K	C	S
MAD	A	E	P	H	K	L	L	R	K	L	G	C	L	L	V	T	A	L	L	K	S	L	L	H	L	I	T	K	K	K
MXI1	A	E	P	H	R	L	L	R	K	L	V	C	L	L	I	T	A	L	L	K	G	A	L	H	L	I	N	K	K	K
ESC	K	D	P	E	D	I	L	K	K	E	D	L	A	F	L	K	A	L	W	I	G	Q	L	Y	L	I	T	E	R	Q
USF2	D	V	P	K	Q	I	L	N	S	N	K	W	I	I	I	K	A	L	G	C	G	D	I	Y	L	I	S	R	K	E
CBF	E	N	P	N	V	I	L	N	S	T	D	A	L	I	L	K	A	L	A	A	A	E	I	Y	L	I	A	Q	R	K
MYC	N	F	P	E	A	L	L	K	R	R	D	S	Q	F	I	K	A	V	V	T	V	A	I	Y	L	I	K	L	K	S
MAX	D	H	P	H	S	I	L	K	R	D	D	S	S	F	V	R	A	M	A	T	Q	E	I	Y	L	I	D	Q	K	Y
CENPBR	P	I	P	D	A	L	R	R	R	K	F	G	N	E	I	K	E	S	R	R	A	K	I	Y	L	G	A	V	S	A
EMC	A	L	P	E	K	M	L	K	K	-	D	M	L	Y	V	K	V	L	L	I	E	D	I	Y	I	I	Q	C	H	D
ID	Y	S	P	D	R	M	L	N	K	G	E	C	L	Y	V	K	V	L	V	I	E	D	I	Y	L	I	Q	R	H	D
AP	Q	Q	P	S	S	I	L	N	K	A	T	G	L	F	I	K	T	L	A	A	A	E	I	Y	L	I	Q	F	Q	S
MYOD	S	E	S	K	T	V	L	N	K	E	R	A	C	F	T	K	A	L	V	I	E	R	I	Y	L	I	R	E	N	G
E12	R	K	Q	D	E	I	L	N	G	E	R	A	M	F	C	K	A	L	L	V	L	A	I	V	L	I	H	L	Q	S
ASCT	K	S	P	Q	Q	V	L	N	R	N	Q	G	H	F	I	K	A	L	V	V	S	E	T	Y	L	I	K	R	M	R
DELILAH	R	E	P	E	T	I	L	N	R	T	H	A	C	F	V	K	A	L	I	M	T	K	T	Y	L	I	R	T	L	M
ATONAL	Q	D	P	N	R	L	L	N	R	Q	Q	A	Y	F	L	K	A	L	H	Q	E	T	T	Y	L	I	Q	S	M	A
RAT4	H	D	P	D	N	L	L	N	R	A	K	A	V	L	V	K	A	L	I	K	E	N	T	Y	L	I	R	W	L	A
LYL1	Q	A	P	N	E	V	L	N	R	G	K	A	L	F	L	K	A	L	N	M	E	K	V	Y	L	I	R	G	L	F
HEN1	E	A	P	A	E	F	L	N	R	L	K	A	L	F	L	K	A	L	I	I	E	C	I	Y	L	I	R	S	L	Y
dHAND	Q	A	P	S	E	I	L	N	R	S	E	A	C	F	I	K	A	L	I	T	K	S	T	Y	L	I	R	A	L	Y
TWIST	Q	A	P	S	A	L	L	N	R	E	K	A	I	F	I	K	A	L	I	A	Q	R	T	Y	L	I	K	D	L	F

FIG. 9.—Neighbor-joining tree for 33 exemplar proteins and the distribution of residues for helix 1 and helix 2. Within a helix, residues are grouped by their positions in the helical wheel (figs. 6 and 7). In the tree, lineages representing Groups A–D of Atchley and Fitch (1997) are denoted.

In the DNA-binding region, there are five strongly conserved sites with $E < 1.7$ (sites 1, 2, 9, 10, and 12) (fig. 7), and four of these are highly basic in that they have K or R residues in great preponderance. The exception is site 9, which has a glutamic acid in 93% of all sequences. The glutamic acid at site 9 contacts the C in the E-box (CANNTG), and its presence indicates that DNA binding occurs. Those bHLH proteins lacking an E at site 9 do not bind DNA (groups C and D, *sensu* Atchley and Fitch 1997). The remaining highly variable sites in the basic region (5, 6, and 11) do not appear to show a systematic pattern of functional group amino acids.

The remainder of the sites shown in figure 7 (sites 14–28) constitute helix 1 in the bHLH domain. There are a number of highly conserved sites constituting one face on the helical wheel. These include sites 16, 17, 20, 23, 24, 27, and 28, and they comprise a conspicuous distribution. Sites 16, 20, 23, and 27 are hydrophobic sites with a high preponderance of I, L, V, and F amino acid residues.

According to Klingler and Brutlag (1994) and others, there is a hydrophobic periodicity that characterizes many α helices, i.e., the relative positions of amino acids in an amphipathic α helix may influence their interresidue correlation structure. Thus, an amino acid at position *i* may show a preference for similar types of amino acids at sites $i + 3$ and $i + 4$, which are on the same side of the helix. Analogously, an amino acid at position *i* may show a preference for dissimilar amino acid types at positions $i + 2$ and $i + 5$. More explicitly, if a hydrophobic residue occurs at site *i*, there is a greater expectation of seeing a hydrophobic residue at sites $i + 3$ and $i + 4$. Thus, the more coincident two residues are on one side of an α helix, the more likely they are to be of the same hydropathy. Conversely, the closer to a 180° separation two residues are, the more likely they are to be of opposite hydropathies (Klingler and Brutlag 1994).

Sites 16, 20, 23, and 27 are three to four sites apart, in accordance with the $i + 3$ and $i + 4$ pattern of hydrophobic periodicity described by Klingler and Brutlag (1994). Thus, this set of four sites provides a highly hydrophobic face for the helix. Sites 17 and 24, on the other hand, are conserved sites with hydrophilic residues

(K, R, N) which provide the hydrophilic face indicative of amphipathic helices.

Finally, site 28 has a high frequency of P residues (63%), indicative of the last site of an α helix. At site 32, 35% of the sequences have P residues. Some proteins, like MyoD, continue helix 1 another turn, and the protein is turned out of the helix into the loop one turn later than in other bHLH proteins.

Figure 8 shows the helical wheel for helix 2. The helix starts with site 50, which is a highly conserved K residue (93% of all sequences). Hydrophobic periodicity is easily seen relative to sites 57 and 64, which are predominantly hydrophobic residues, as are the $i + 3$ and $i + 4$ sites. However, the $i + 2$ and $i + 5$ sites to sites 57 and 60 are not necessarily hydrophobic and are rather diverse in their amino acid compositions. This is also the case for sites 54 and 61. On average, helix 2 appears to be more hydrophobic than helix 1. The predictive motif proposed by Atchley, Terhalle, and Dress (1999) to discriminate bHLH proteins employs sites that fall on these highly conserved faces of the two helices.

Phylogenetic Aspects of the Helical Configuration

To explore the phylogenetic aspects of an α helix configuration, we combined a neighbor-joining tree of 33 bHLH domains with the distribution of their amino acids along a helical wheel (fig. 9). The proteins chosen for these analyses were simply some typical representatives of the various clades and evolutionary lineages as reported by Atchley and Fitch (1997). This tree delimits those proteins that belong to groups A–D. Group A and B bHLH domain proteins are most prevalent in the literature and databases, followed by those of group C. Group A proteins bind to a CAGCTG E-box configuration, while group B proteins bind to the CACGTG E-box configuration. Group C has a more complex DNA-binding behavior, and group D does not bind DNA.

Atchley, Terhalle, and Dress (1999) derived a 19-element predictive motif that accurately identifies bHLH-domain-containing proteins. Groups A and B show the smallest deviation from this motif, while groups C and D deviate considerably, mostly in the basic DNA-binding region, as would be expected. Proteins that fit this predictive motif most closely include MyoD and Achaete-Scute (no mismatches) and LYL1, d-HAND, and SREBP (one mismatch). The most divergent proteins are INO4 and CENP-B, with 10 mismatches, while INO2 has 8 mismatches. (EMC also has 10 mismatches, but this is to be expected since it is a group D protein which does not have a typical basic DNA-binding component.)

Let us consider the structures of those proteins that deviate most from the predictive motif. CENP-B was originally described as a bHLH protein (Sullivan and Glass 1991; Sugimoto, Muro, and Himeno 1992). However, Atchley and Fitch (1997) suggested that it deviated considerably in its primary sequence from more typical bHLH proteins. In addition to considerable deviation from the predictive motif for bHLH domains (Atchley, Terhalle, and Dress 1999), the first residue in helix 1 is a proline. Proline is an amino acid generally associated with breaking of α helices, and, indeed, the last residue in helix 1 and the first one in the loop regions are prolines. Analysis of the secondary structure of the CENP-B bHLH domain by the PSA algorithm described in *Materials and Methods* indicates that the stretch of residues considered to be the helix 1 region have low probabilities of being α helices. The probability values range from 0.34 for the first residue (proline) to 0.56 (an isoleucine midway in helix 1). The means and standard deviations for the probabilities of being α helices for the residues in the basic, H1, L, and H2 regions are 0.67 (±0.16), 0.37 (±0.15), 0.40 (±0.15), and 0.55 (±0.15), respectively. An analysis of variance of the basic and H1 regions to test the null hypothesis that the basic and helix 1 regions have an equal probability of being α helices is rejected at $P < 0.001$.

Recently, it was suggested that CENP-B is not a helix-loop-helix protein but, rather, should be classified as helix-turn-helix (Iwahara et al. 1998). Our analyses here suggest that CENP-B probably should not be classified as an HLH protein.

Another protein that deviates considerably from the predictive model is INO4, which is believed to positively regulate the coordinate expression of phospholipid structural genes in yeast (Nikoloff and Henry 1994). The fit of INO4 to the predictive motif in helix 2 is particularly bad (five mismatches). However, in spite of the sequence differences, the PSA analyses indicate that INO4 fits the α helix model rather well. Our analyses suggest that a more detailed analysis of the structure of INO4 might be fruitful.

The extent of conservation at particular sites and pairs of sites can now be more clearly seen, together with deviations from these patterns of conservation. For example, all of the proteins except four have proline residues at site 28. The exceptions include the proteins ADD1, INO2, MyoD, and E12. Furthermore, only a single sequence (PHO4) has a residue other than L or I at site 54. In PHO4, the residue at this site is an E.

Discussion

A clear understanding of interactions among amino acid sites is fundamental to producing a comprehensive model for the structure, function, and evolution of proteins. Because of the limited number of secondary and tertiary structures that proteins can assume (Sternberg 1996), it can be argued that evolution at the amino acid sequence level is constrained by higher-order structure. This ‘‘phylogenetic inertia’’ (*sensu* Felsenstein 1985), which determines the rate and direction of evolutionary change at the sequence level, is reflected in the variability and covariability among primary sequence elements. Covariance structure and phylogenetic inertia are important concepts in Darwinian evolution that have not been adequately explored at the molecular level. The analyses reported here deal with the magnitude and origins of covariability among amino acid sites and their relevance to protein structure and evolution.

In the *Introduction,* we listed several important topics or questions that we wished to provide information about in these analyses. We discuss these topics below.

Origins of Associations among Amino Acid Sites

We have shown elsewhere that considerable amounts of covariation occur among amino acid sites in the bHLH domain (Atchley, Terhalle, and Dress 1999). One impetus for the present study was to explore whether such covariation could be partitioned into those effects due to phylogenetic, structural/functional, and stochastic causes. Such partitions are critical to understanding the origin of sequence and structural variability and the evolution of protein structure.

To resolve this question about partitioning covariation, we simulated sequence data using a parametric bootstrap procedure. Generation of the simulated data sets requires two underlying models: (1) a phylogenetic model that dictates the amounts of change (branch lengths) and the grouping of changes (tree topology), and (2) an evolutionary model describing the probability of change from one residue to another. Results derived from these simulated data are dependent on the characteristics of these underlying models. Because the JTT matrix is based on a generalized model of protein evolution that accounts for phylogeny, it is an appropriate model for the calculation of residue changes.

The approach employed here permits residue changes to be generated in a controlled manner. The pattern of clustering and the number of residue changes between nodes (calculated from the branch lengths) constrain the magnitude of correlations between sites. Long branches leading to clades with many taxa produce high MI values. Conversely, a pattern of short internal branches coupled with long terminal branches leads to small MI values. Therefore, as one would expect, the distribution of MI values reflecting only stochastic and phylogenetic constraint will be quite dependent on the characteristics of the tree used in the parametric bootstrap procedure.

The analyses described here for the highly conserved bHLH domain clearly demonstrate that the observed covariation among amino acid sites can be partitioned into those associations due to common evolutionary history versus those due to structural/functional constraints. We showed in tables 1 and 2 that there are significant associations arising from phylogeny, structure, and function origins in all the components of the bHLH domain. With regard to structural and functional constraints, there are significant associations among amino sites within the DNA-binding region, between the binding and dimerization regions, and between the dimerization regions. Some of these significant associations can be attributed to particular structural and functional attributes of the protein, including significant associations among amino acid sites due to hydropathy relationships and the sizes of residues. However, the basis for other significant associations await further clarification from critical experimental analyses. One of the purposes of large quantitative studies like this one is to provide hypotheses about structural and functional relationships which can be explored by subsequent experimental studies.

Entropy and Site-Directed Mutagenesis Studies

Site-directed mutagenesis is a powerful tool for elucidating protein structure. With this approach, particular amino acids are perturbed in specific ways in order to assess the impact of sequence changes on protein structure. Unfortunately, the number of possible sites to perturb is quite high, the relationships among sites is not well understood, and consequently there is always a quandary about which sites to experimentally alter.

Quantitative data from entropy and MI calculations may provide valuable insight into this problem. For example, perturbing amino acid sites exhibiting low entropy values or sites sharing significant amounts of mutual information with other sites may generate quite different results from those obtained by perturbing sites with high diversity or low mutual information. Obviously, protein stability, folding, and functionality are dynamic multidimensional and integrated phenomena. It follows, then, that information about variability and covariability among sites would provide valuable input for mutagenesis experiments and for the subsequent development of robust models for protein structure and function.

Entropy and Classes of Amino Acid Sites

Analyses of the basic DNA-binding region and the two α-helical regions of the bHLH domain suggest that there are three classes of sites. The first class includes amino acid sites with low entropy and low mutual information with other sites. Epitomizing this class are the contact sites between the two α helices that comprise the hydrophobic core of these domains. These contact sites had entropy values varying between 0.2 and 2.3, a range of values that differed significantly from (and did not overlap with) the entropy values for the noncontact sites. If the amino acid residues at each site are transformed into functional groups of amino acids as described in Atchley, Terhalle, and Dress (1999), the entropy relationships are even more pronounced. These contact sites exhibited very low levels of correlation in residue composition with other sites in the bHLH domain. Such low levels of mutual information are to be expected, because two variables must exhibit variation before they can exhibit shared or common variation as reflected by the MI values (which can be demonstrated algebraically, as, e.g., in Atchley, Terhalle, and Dress 1999).

The second class of amino acid sites involves those with higher levels of sequence diversity (entropy) and high levels of mutual information. We described a number of sites with higher entropy values where residue composition was highly correlated with that at other sites. Many of these sites are involved in important structural and/or functional attributes in these proteins.

The third class of amino acid sites includes those with high levels of entropy but low levels of mutual information. Thus, variability at these sites is apparently unrelated to variability at other amino acid sites. This independence could simply stem from stochastic variation at these sites unrelated to any functional or structural considerations in the protein or with regard to other sites. Alternatively, the variability at these individual sites could be of functional or structural significance, but these sites function in a manner orthogonal to other sites.

Entropy, Conserved Sites, and Protein Structure

A basic tenet of protein structural analyses is that the information contained in the primary sequence is sufficient to dictate the three-dimensional structure (Strait and Dewey 1996). Consequently, another impetus for these analyses was to integrate information theoretic analyses about sequence diversity with attributes of the proteins that had been elucidated by experimental studies. If quantitative approaches using techniques like entropy measures and mutual information are to be successful, we must be able to relate sequence characteristics to structural and functional attributes over large numbers of proteins.

The relationship between sequence covariability, packing, and protein structure is an essential part of understanding protein evolution. Native proteins assume a particular packing density. If the maximum possible packing density is assumed, then sequence evolution could be very difficult, because each mutation of an interior residue would require one or more simultaneous and compensating mutations to maintain the dynamics of such a high density (Richards 1992). In this case, amino acid substitutions would involve paired or higher-order changes to maintain packing relationships. Packing restrictions on the surface residues would be weaker than those in the interior.

There is considerable heterogeneity in the amount of residue diversity at various sites. However, this residue diversity seems to correlate well with surface accessibility of the positions, with the interior positions being much more conserved. Indeed, our results show systematic relationships between covariability, packing, and structure. Amino acid sites from the α helices known to pack together in the interior of the protein are highly conserved and, as a consequence, exhibit low variability. The entropy values for these contact sites are low, indicating low diversity at the contact sites. Those sites within the α helix that are buried and constitute the hydrophobic core are significantly less variable than exposed and hydrophilic sites. In addition, they show very little covariability in residue composition between sites. In contrast, sites away from this hydrophobic core show significantly more sequence diversity, as reflected by their entropy values.

Furthermore, there are highly conserved sites among diverse bHLH proteins which are therefore highly predictive for bHLH proteins. As a consequence, these residues discriminate the bHLH domain with highly accuracy (Atchley, Terhalle, and Dress 1999). These highly invariant sites show very little intercorrelation with other sites with regard to their constitutive residues. This lack of correlation stems largely from the fact that invariant sites cannot exhibit covariation with other sites. To be most effective, predictive motifs need to exhibit high stability among individual elements and a lack of intercorrelation among the elements, i.e., independence among the component elements. Such low variability and covariability is the case for the elements of the bHLH predictive motif.

The observed relationships between entropy measures and protein interactions described for the bHLH domain are very intriguing. They suggest that certain structural and functional attributes of proteins can be predicted from quantitative measures of sequence diversity and association. However, analyses such as these need to be carried out on other groups of proteins before generalities can be made. One conclusion from these theoretical and experimental findings is that efforts to model protein structure must take into consideration the simultaneous covariation among amino acid sites. Data on covariation among sites is necessary to understand the multidimensional structural, functional, and evolutionary dynamics in proteins.

Acknowledgments

We are particularly indebted to Adrian Ferre-D'Amare for his helpful discussions and interpretations with regard to his bHLH crystal structure studies and for providing us with unpublished observations. Jeff Thorne, James Rosinski, and Bill Bruno provided helpful comments on the manuscript. The research was generously supported by the National Institutes 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).

LITERATURE CITED

APPLEBAUM, D. 1996. Probability and information. Cambridge University Press, Cambridge, England.

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

———. 1997. A natural classification of the basic helix-loop-helix class of transcription factors. Proc. Natl. Acad. Sci. USA **94**:5172–5176.

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

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

BROWNLIE, P., T. CESKA, M. LAMERS, C. ROMIER, G. STIER, H. TEO, and D. SUCK. 1997. The crystal structure of an intact human Max-DNA complex: new insights into mechanisms of transcriptional control. Structure **5**:509–520.

BRUNO, W. 1996. Modeling residue usage in aligned protein sequences via maximum likelihood. Mol. Biol. Evol. **13**: 1368–1374.

CLARKE, N. D. 1995. Covariation of residues in the homeodomain sequence family. Protein Sci. **4**:2269–2278.

EFRON, B., and R. J. TIBSHIRANI. 1993. An introduction to the bootstrap. Chapman and Hall, New York.

Ellenberger, T., D. Fass, M. Arnaud, and S. C. Harrison. 1994. Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. Genes Dev. **15**:970–980.

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. **125**:1–15.

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

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

Goldman, N. 1993. Simple diagnostic statistical tests of models for DNA substitution. J. Mol. Evol. **37**:650–661.

Goldman N, J. L. Thorne, and D. T. Jones. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics **149**:445–458.

Herzel, H., and I. Gross. 1995. Measuring correlations in symbol sequences. Physica A **216**:518–530.

Huelsenbeck, J. P., D. M. Hillis, and R. Jones. 1996. Parametric bootstrapping in molecular phylogenies: applications and performance. Pp. 19–45 *in* J. D. Ferraris and S. R. Palumbi, eds. Molecular zoology: advances, strategies, and protocols. Wiley-Liss, New York.

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

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS **8**:275–282.

Klingler, T. M., and D. L. Brutlag. 1994. Discovering structural correlations in alpha-helices. Protein Sci. **3**:1847–1857.

Korber B. T., R. M. Farber, D. H. Wolpert, and A. S. Lapedes. 1993. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc. Natl. Acad. Sci. USA **90**:7176–7180.

Koshi, J. M., and R. A. Goldstein. 1997. Mutation matrices and physical-chemical properties: correlations and implications. Proteins **27**:336–344.

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

Morgenstern, B., and W. R. Atchley. 1999. Modular evolution of the bHLH family of transcription factors. Mol. Biol. Evol. **16**:1654–1663.

Murre, C., G. Bain, M. A. van Dijk, I. Engel, B. A. Furnari, M. E. Massari, J. R. Matthews, M. W. Quong, R. R. Rivera, and M. H. Stuiver. 1994. Structure and function of helix-loop-helix proteins. Biochim. Biophys. Acta **1218**:129–135.

Nikoloff, D. M., and S. A. Henry. 1994. Functional characterization of the INO2 gene of Saccharomyces cerevisiae. J. Biol. Chem. **269**:7402–7411.

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

Richards, F. M. 1992. Folded and unfolded proteins: an introduction. Pp. 1–58 *in* T. E. Creighton, ed. Protein folding. W. H. Freeman and Co., New York.

Roman-Roldan, R., P. Bernaola-Gavan, and J. L. Oliver. 1996. Application of information theory to DNA sequence analysis: a review. Patt. Recog. **29**:1187–1194.

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

Schneider, T. D. 1996. Reading of DNA sequence logos: prediction of major groove binding by information theory. Methods Enzymol. **s**:445–455.

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

Sokal, R. R., and F. J. Rohlf. 1995. Biometry. Freeman and Sons, New York.

Sternberg, M. J. E. 1996. Protein structure prediction. IRL Press, Oxford, England.

Strait, B. J., and T. G. Dewey. 1996. The Shannon information entropy of protein sequences. Biophys. J. **71**:148–155.

Sugimoto, K., Y. Muro, and M. Himeno. 1992. Anti-helix-loop-helix domain antibodies: discovery of antibodies that inhibit DNA binding activity of human centromere protein B (CENP-B). J. Biochem. **111**:478–483.

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

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. Hillis, C. Moritz, and B. K. Mable, eds. Molecular systematics. 2nd edition. Sinauer, Sunderland, Mass.

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

White, J. V., C. M. Stultz, and T. F. Smith. 1994. Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. Math. Biosci. **119**:35–75.

Wollenberg, K., and W. R. Atchley. 2000. Separation of phylogenetic and functional associations in biological sequences using the parametric bootstrap. Proc. Natl. Acad. Sci. USA (in press).

Yang, Z. 1997. Phylogenetic analysis by maximum likelihood (PAML). Version 1.3. Department of Integrative Biology, University of California at Berkeley.

Charles Aquadro, reviewing editor

Accepted October 11, 1999