# Molecular Architecture of the DNA-Binding Region and Its Relationship to Classification of Basic Helix–Loop–Helix Proteins

*William R. Atchley and Jieping Zhao*

Department of Genetics and Center for Computational Biology, North Carolina State University

Multivariate statistical analyses are used to explore the molecular architecture of the DNA-binding and dimerization regions of basic helix–loop–helix (bHLH) proteins. Alphabetic amino acid data are transformed to biologically meaningful quantitative values using a set of 5 multivariate "indices." These multivariate indices summarize variation in a large suite of amino acid physiochemical attributes and reflect variability in polarity–accessibility–hydrophobicity, propensity for secondary structure, molecular size, codon composition, and electrostatic charge. Using these index score data, discriminant analyses describe the multidimensional aspects of physiochemical variation and clarify the structural basis of the prevailing evolutionary classification of bHLH proteins. A small number of amino acids from both the binding dimerization domains, when considered simultaneously, accurately distinguish the 5 known DNA-binding groups. The relevant sites often have well-documented structural and functional characteristics.

## Introduction

Eukaryotic genes contain specific short DNA sequences that regulate gene expression by binding to a special class of proteins called transcription factors. Variation in transcriptional machinery underlies the myriad of gene expression patterns responsible, in large measure, for the observed patterns of evolutionary diversity. Understanding the evolution of gene expression patterns requires knowledge of the "molecular architecture" of DNA–transcription factor interactions, that is, the relative composition of amino acids, their physiochemical properties, the structure of the DNA–protein interaction, binding affinities among protein variants, and a myriad of other aspects. Information on architecture is needed to resolve questions about how proteins recognize specific DNA regions, how the necessary binding specificity is achieved to regulate the myriad biological processes that take place within the cell, and what structural and functional aspects of each amino acid facilitate specificity of expression.

Variation in transcription factors stems from the nonrandom distribution of the component amino acids that, in turn, arises from various underlying phylogenetic, structural, functional, and stochastic causes (Atchley et al. 2000; Wollenberg and Atchley 2000). Consequently, to explore the architecture of complex molecular phenomena, the relevant system components must be dissected out in order to ascertain the underlying causes of their variability, portray the relative biological impact of their variability, and describe the evolution of the various components and their interactions. Computational biology provides the requisite quantitative methodology for such tasks.

Statistical analyses can partition sequence variation into major biologically meaningful underlying causal components. One model is as follows:

$$\begin{aligned}\text{sequence variation} &= \text{evolutionary constraints} + \text{structural constraints} \\ &\quad + \text{functional constraints} + \text{interactions} \\ &\quad + \text{stochastic effects}\end{aligned} \tag{1}$$

(Atchley et al. 2000, 2005; Wollenberg and Atchley 2000; Buck and Atchley 2005). Achieving such partitions was hampered previously because sequence elements are represented as alphabetic letters with no natural underlying metric. Atchley et al. (2005) circumvented this "sequence metric problem" through multivariate statistical analyses of the latent covariance structure of almost 500 physiochemical attributes of amino acids. The resultant information was summarized in 5 multidimensional and biologically interpretable "indices" or "factor scores." These scores numerically position each of the 20 coding amino acids in 5 highly interpretable multivariate patterns of physiochemical variability. The factor scores are statistically independent variables reflecting highly interpretable patterns of covariation for 1) polarity, accessibility, and hydrophobicity (***pah***); 2) propensity for secondary structure (***pss***); 3) molecular size or volume (***ms***); 4) codon composition (***cc***); and 5) electrostatic charge (***ec***). The proportions of the common physiochemical variation explained by factors 1–5 are 42.3%, 25.7%, 17.2%, 10.2%, and 4.6%, respectively.

Observed variation in the *i*th amino acid ($V_i$) in a large collection of aligned protein sequences can be partitioned and its variability modeled as

$$\begin{aligned}V_i = (b_1)\mathbf{pah} + (b_2)\mathbf{pss} + (b_3)\mathbf{ms} \\ + (b_4)\mathbf{cc} + (b_5)\mathbf{ec} + \varepsilon,\end{aligned} \tag{2}$$

where $b_j$ is the relative contribution of the *j*th physiochemical factor and $\boldsymbol{\varepsilon}$ is the residual or unique variation not explained by the model. The magnitude and sign of the various *b* values determine their relative contribution to the model.

Factor scores have well-understood statistical properties and, consequently, can be used as new synthetic variables in subsequent analyses. A multiple sequence alignment can be transformed into a numerical database where each amino acid letter code becomes an array of 5 numbers corresponding to the 5 physiochemical factor scores. These new numerical data can be used in subsequent statistical analyses to explore important questions about

Key words: computational biology, bHLH proteins, molecular architecture, multivariate statistics, DNA-binding specificity, molecular evolution.

E-mail: bill@atchleylab.org.

*Mol. Biol. Evol.* 24(1):192–202. 2007
doi:10.1093/molbev/msl143
Advance Access publication October 13, 2006

biosequence variation and its relationship to protein structure, function, and evolution. Further, patterns of covarying sites within each of these 5 multidimensional physiochemical attributes can be explored in detail. For example, there are a number of amino acids sites within the basic helix–loop–helix (bHLH) domain that cluster into separate groups based on ***pah*** scores and provide important information about the evolution of structure and function.

Herein, a multivariate statistical approach is used to explore the molecular architecture of DNA binding in the bHLH family of transcriptional regulators. The bHLH proteins regulate a broad array of developmental processes in eukaryotic organisms including cell proliferation, differentiation, neurogenesis, myogenesis, hematopoiesis, sex determination, environmental sensing, and many other important processes (Murre et al. 1994; Atchley and Fitch 1997; Massari and Murre 2000; Ledent et al. 2002; Kewley et al. 2004).

The evolutionarily highly conserved bHLH domain is tripartite and consists of a basic DNA-binding region and 2 α-helices separated by a variable length loop (Ferre-D'Amare et al. 1993). Dimerization produces a left-handed, 4-helix bundle with a hydrophobic core that is stabilized by van der Waals and other interactions in the HLH region of each monomer (Beltran et al. 2005). These interactions facilitate contact between the basic DNA-binding region and the major groove of the DNA (Ferre-D'Amare et al. 1993). DNA–protein interactions occur at a highly conserved hexanucleotide CANNTG sequence "E-box," where the center base pair defines phylogenetically related bHLH sequences. Bases outside the E-box may also be involved in determining the specificity of DNA binding.

Crystal structures are available for 6 bHLH proteins, including Max (Ferre-D'Amare et al. 1993; Brownlie et al. 1997), E47 (Ellenberger et al. 1994), USF (Ferre-D'Amare et al. 1994), MyoD (Ma et al. 1994), PHO4 (Shimizu et al. 1997), and SREBP (Parraga et al. 1998). Superimposition of these 6 structures gives a close fit for the α-carbons over much of the domain, suggesting that we can generalize the structural attributes of the various proteins (Shimizu et al. 1997). It is assumed here that major structural and functional features of the bHLH domain, as described in these 6 structural studies, are common to all 288 proteins examined herein.

Certain amino acids have been identified as either buried in the hydrophobic core or exposed on the surface. Also noted are amino acids that contact DNA bases or the phosphate backbone, together with those that pack against other amino acids or that exhibit various other interactions, such as salt bridges and hydrogen bonds (table 1).

DNA-binding activity in bHLH proteins involves a transition from a largely unfolded to primarily α-helical configuration of the bHLH region (Turner et al. 2004). All bHLH proteins undergo a basic region random coil to α-helix folding transition on specific DNA binding (Parraga et al. 1998). There are numerous base-specific and phosphate backbone contacts with amino acids in the basic region as well as in the dimerization region. The precise characteristics of helical configuration are apparently defined by the physiochemical attributes of individual amino acids.

**Table 1**
**Structural Attributes of bHLH Domain Elements**

| Site | Domain | Structural Attributes | *E* |
|---|---|---|---|
| 1 | B1 | Contacts DNA (A7′) or phosphate backbone; beginning of basic region | 0.521 |
| 2 | B2 | Contacts phosphate backbone in PHO4, E47 | 0.443 |
| 3 | B3 | | 0.862 |
| 4 | B4 | | 0.808 |
| 5 | B5 | Contacts DNA (G3′) | 0.647 |
| 6 | B6 | Contacts phosphate backbone | 0.588 |
| 7 | B7 | | 0.806 |
| 8 | B8 | Contacts phosphate backbone—MyoD, E47; contacts DNA in E47 | 0.680 |
| 9 | B9 | Contact with DNA (C3 and A2) | **0.303** |
| 10 | B10 | Contacts phosphate backbone | **0.396** |
| 11 | B11 | Contacts phosphate backbone—SREBP | 0.757 |
| 12 | B12 | Contacts phosphate backbone | **0.249** |
| 13 | B13 | Contacts DNA (central base) and phosphate backbone (contacts central base in E-box) | 0.526 |
| 14 | H1 | | 0.834 |
| 15 | H2 | Contacts phosphate backbone in PHO4 | 0.756 |
| 16 | H3 | Buried site; side chain packs against 20 | 0.544 |
| 17 | H4 | Contacts phosphate backbone—SREBP | **0.293** |
| 18 | H5 | | 0.796 |
| 19 | H6 | | 0.661 |
| 20 | H7 | van der Waals contacts with H2 side chains (especially sites 50, 53, 54′); buried site | 0.428 |
| 21 | H8 | | 0.806 |
| 22 | H9 | | 0.716 |
| 23 | H10 | Packs against 53, 57; buried site | **0.027** |
| 24 | H11 | | 0.466 |
| 25 | H12 | | 0.751 |
| 26 | H13 | | 0.678 |
| 27 | H14 | Packs against 60, 61 (in Max) | 0.505 |
| 28 | H15 | End of helix; *P* turns strand; packs against 60; buried site | **0.372** |
| 29 | L1 | Beginning of loop region | 0.841 |
| 30 | L2 | | 0.849 |
| 31 | L3 | | 0.659 |
| 32 | L4 | | 0.810 |
| 46 | L5 | | 0.722 |
| 47 | L6 | Stabilizes loop path, salt bridge, contacts phosphate backbone | 0.430 |
| 48 | L7 | | 0.539 |
| 49 | L8 | Contacts phosphate backbone in PHO4 | 0.557 |
| 50 | H21 | Begin H2, packs against 63 and 64′, contacts phosphate backbone, DNA C3; anchors H2 to DNA; stabilizes H1 with H2 by packing with 20 | **0.132** |
| 51 | H22 | | 0.635 |
| 52 | H23 | | 0.665 |
| 53 | H24 | Packs against 50 and 54′ (Max) | **0.334** |
| 54 | H25 | Packs against 50 and 53 (Max) | **0.107** |
| 55 | H26 | | 0.602 |
| 56 | H27 | | 0.745 |
| 57 | H28 | | **0.349** |
| 58 | H29 | | 0.629 |
| 59 | H210 | | 0.740 |
| 60 | H211 | | **0.341** |
| 61 | H212 | Packs against symmetry mate 61 and 60 (Max) | **0.319** |
| 62 | H213 | | 0.796 |
| 63 | H214 | | 0.861 |
| 64 | H215 | Interacts with symmetry mate (Max) | **0.354** |

NOTE.—Structural explanations are based on Ferre-D'Amare et al. (1993) for Max. Site numbering scheme is based on the human sequences of Max. The last column is the normalized site entropy for the complete 288 sequences in this database. Values closer to zero are more highly conserved.

**Table 2**
**Conventional Classification of bHLH DNA-Binding Groups Based on the Amino Acid Composition of the Basic Region**

| | | Amino Acid Composition by Site | | | | |
|---|---|---|---|---|---|---|
| Binding Group | E-Box | 5 | 6 | 8 | 9 | 13 |
| **A** (MyoD, Ac-S) | CA***GC***TG | A,N,T | N,T | ***R(K)*** | ***E*** | M,L,V |
| **B** (Myc, SREBP | CA***CG***TG | ***R,K*** | N,V,S,K | I,V,L | ***E*** | ***R*** |
| **C** (Ahr) | None | R,S,K,Q,T | N,D,S | P,A | K,S,A | ***R*** |
| **D** (ID, Emc) | None | D,V,S,T | E,A,T,M | D,E,M | P,V,T,S | L,N,Q |
| **E** (Hairy) | CACG***CG***,CACG***AG*** | ***K*** | ***P*** | L,M,V | ***E*** | ***R*** |

NOTE.—DNA-binding groups **A**–**E** are provided together with the E-box composition and the most frequent amino acid composition at amino acid sites 5, 6, 8, 9, and 13. Diagnostic amino acids at these sites are shown in bold and italic font and underlined.

For purposes of the present work, modulation of DNA-binding specificity in the cell is hypothesized to be hierarchical. That is, there is a coarse gradient of bHLH specificity at the entry level (i.e., DNA-binding groups) and a finer gradient of much more specific affinities recognized at lower levels (enhanced binding in individual clades). Enhanced specificity may involve simultaneous variation of multiple amino acids as well as interactions with other proteins. This latter hypothesis is evaluated in more detail in a subsequent paper.

Amino acid sequence attributes of the basic region have been used to classify bHLH proteins into 5 natural (phylogenetic) groups that reflect general DNA-binding behavior (Atchley and Fitch 1997; Ledent et al. 2002) (table 2). Three groups of proteins directly bind DNA (groups **A**, **B**, and **E**). The fourth group (**C**) is the so-called bHLH-PAS proteins that do not directly bind DNA but rather heterodimerize to various group **B** proteins that, in turn, bind to the DNA. The fifth group (**D**) has no basic DNA-binding region and does not bind DNA. Rather, these proteins act as dominant negative regulators by heterodimerizing with bHLH proteins and preventing them from binding DNA.

Herein, multivariate statistical analyses on 288 bHLH domain–containing proteins are used to explore a series of questions about the molecular architecture of DNA binding. 1) Which amino acids in the bHLH domain best discriminate the 5 DNA-binding groups? 2) Are the best discriminating amino acids always found in the DNA-binding region? 3) What structural aspects of amino acids are reflected by these best diagnostic amino acids? 4) What relationships exist between these discriminating amino acids and DNA–protein interactions? Further, this report provides a paradigm for how multivariate statistical analyses like discriminant analysis can be applied to amino acid sequence data to explore multidimensional questions in protein evolution and structure.

Hereafter, to conserve space, sequence element (amino acid) *n* is described as site *n*. DNA-binding groups are denoted in boldface font (e.g., group **A**), and residue abbreviations are in italics (e.g., residue *A*; amino acids are coded by conventional single letter symbols).

## Materials and Methods

### Data Structure

The totality of variability in multivariate sequence data can be statistically partitioned as

$$\text{total variability} = \text{among-groups variation} + \text{within-groups variation}.$$

The "groups" category could include clades within a phylogenetic tree, DNA-binding groups, promoters, domains, or other biologically realistic groupings. In the present analyses, observed variability in a diverse array of bHLH sequences is partitioned into that variability "among" well-defined proteins exhibiting specific DNA-binding patterns relative to variability "within" each of the binding groups. We will briefly summarize the multivariate covariation patterns in the 288 sequences without any group designation as reported by Atchley and Buck (Atchley WR, Buck MJ, submitted paper). Then we will classify these sequences into the 5 experimentally defined DNA-binding groups and explore the patterns of sequence covariation that distinguish among them.

### The bHLH Data

A total of 288 sequences containing the bHLH domain were aligned using both global (ClustalX) and local (Dialign) alignment algorithms and any differences resolved by eye. Variability in each sequence element is described by a normalized Boltzmann–Shannon entropy value (Atchley et al. 1999, 2000; Wang and Atchley 2006) where values closest to zero are highly conserved and those closest to unity are highly variable. The dynamics for entropy values for the bHLH region for these data are given in figure 1.

Our analyses depend heavily on the use of equivalent traits (homologous amino acids), so some amino acids in the loop region that are difficult to homologize are excluded. Only amino acids 1–31 and 45–64 in the numbering scheme of Atchley and Fitch (1997) were analyzed.

### Overall Sequence Covariation

Atchley and Buck (Atchley WR, Buck MJ. The Multidimensional Nature of Amino Acid Correlation in Basic Helix-Loop-Helix (bHLH) Proteins. Submitted) used pairwise mutual information values (Atchley et al. 1999, 2000) to describe the extent of pairwise correlation among amino acid sites in 288 bHLH sequences. These correlations were computed without regard for any group structure among sequences. Factor analysis was then used to describe the latent patterns of amino acid covariability inherent to the bHLH domain, and these patterns related to structural

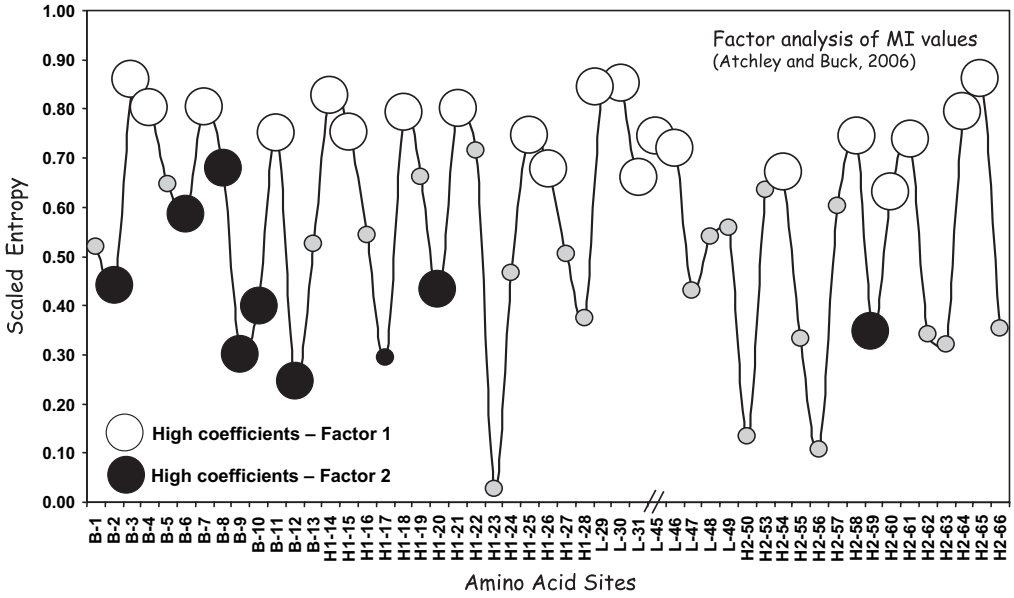


FIG. 1.—Dynamics of amino acid variability by site in the bHLH domain. The Boltzmann–Shannon entropy was computed on amino acid frequency at each amino acid site and the result plotted. Higher values for $E$ imply more variability or uncertainty. The numerical values of the factor coefficients from Atchley WR and Buck MJ (unpublished data) are included to show the orthogonality of the patterns of amino acid variability. See the text for more details.

and functional entities. The first pattern, for example, delimited the amino acids on the hydrophilic surface of the domain, whereas the second pattern focused on the DNA-binding region and related sites. Subsequent patterns related to important structural and functional components of the domains.

## Sequence Classification

The conventional evolutionary classification of the 5 DNA-binding groups using the 13 amino acids in the basic region is shown in table 2 (Atchley and Fitch 1997; Ledent et al. 2002). Herein, this classification is evaluated using decision tree analysis on the 13 amino acids in the DNA-binding region (Breiman 1998). Decision trees are widely used in data mining to graphically describe alternative steps in a decision scheme. They can be used to represent all possible outcomes of a decision process and the paths by which decisions are made in classification tasks. The top layer consists of input nodes, and the decision nodes determine the order of progression. Leaves on the decision tree are all possible outcomes or classifications, whereas the root is the final outcome.

## Data Transformation

An important prerequisite to most multivariate statistical analysis of sequence variability is to transform the alphabetic amino acid codes to quantitative variables that accurately summarize the physiochemical attributes of the amino acids. This transformation is necessary because alphabetic amino acid codes have no realistic or meaningful underlying sequence metric.

Atchley et al. (2005) carried out a multivariate statistical analysis of variation in 495 amino acid physiochemical properties. They used common factor analysis (not principal components) to transform the multidimensional and highly intercorrelated variability in $a$ physiochemical attributes into $k$ orthogonal physiochemical patterns. The reduction in dimensionality in this instance was from $a = 495$ physiochemical attributes to $k = 5$ independent and biologically highly interpretable physiochemical indices. This transformation reconciles the high level of redundancy in the original physiochemical attributes and produces much smaller, statistically independent, and well-conditioned variables for subsequent statistical analyses.

Factor analysis partitions the overall amino acid physiochemical variability into common and unique components (Johnson and Wichern 2002; Atchley et al. 2005). The former is that common variability shared among variables ($r^2$), whereas the latter includes information that is unique to each amino acid and random noise ($1 - r^2$). The resultant common factors are "latent variables" that describe the major patterns of the underlying covariability in the original physiochemical variables. In a factor analysis sense, the unique variance is that variance not explained by the latent structure model.

A set of new synthetic variables called factor scores are generated by this procedure that position each amino acid in the various axes of latent variability. These factor scores or "physiochemical indices" are normally distributed with mean zero and variance (and standard deviation) of unity.

Next, each alphabetic amino acid code in the aligned sequence database was transformed to a $1 \times 5$ vector of factor scores. Thus, each amino acid is described by a set of 5 numerical values reflecting their relative position in these multidimensional physiochemical factors. This transformation permits data to be analyzed one physiochemical

factor at a time or with all factors considered simultaneously. The set of factor scores are multivariate normal with a well-conditioned covariance matrix.

### Discriminant Analysis

Discriminant analysis is a statistical approach for analyzing the multivariate patterns of covariation among a priori defined groups overall and above the variability within groups (Johnson and Wichern 2002). This approach defines the latent structure of among-groups covariation and determines the subset of attributes that best separate a set of a priori defined groups. It provides multivariate information about how groups have diverged, the extent of overlap in the classification system, the sets of variables that best define the groups, and related questions. This powerful method is widely used in many disciplines but seldom has been used in biosequence analyses, primarily because of the alphabetic nature of sequence data.

Herein, discriminant analysis is used in 2 ways to study the DNA-binding group classification. First, we use stepwise discriminant analysis (SWDA) to order amino acid sites in terms of their ability to discriminate the DNA-binding groups. Second, we use multiple group discriminant analysis (=canonical variate analysis [CVA]) to ascertain a set of linear combinations of the quantitative variables "simultaneously" that best reveal the differences among a priori defined classes.

SWDA ranks individual amino acids in the bHLH domain by their ability to discriminate the 5 DNA-binding groups. A step-up amino acid site selection procedure begins with no sites in the model. Then, at each step, an amino acid site is added from the aligned sequences that contributes most to discriminating power of the model, as measured by Wilks' lambda likelihood ratio criterion (Johnson and Wichern 2002). Then, the next best discriminating amino acid is added. At each step, covariances among the amino acids are readjusted. Typically, this procedure continues until all the amino acid sites have been added (complete discriminant analyses) or until a defined level of explained among-groups variance has been achieved.

Statistics provided for each step include the following. 1) $F$ statistic for entering the variable from analysis of covariance. 2) Partial $r^2$ for predicting the variable in question while controlling for the effects of variables already selected for the model. The higher the $r^2$, the more complete the discrimination among groups. 3) Average squared canonical correlation (ASCC) describing the relative distinctiveness of the groups at that step in the model. An ASCC value of 1.0 implies complete discrimination. Here, variables were added until they accounted for a squared partial correlation ($R^2$) value of 90% for attribute factor 1 or 75% for attribute factors 2–5. 4) Pairwise product–moment correlation coefficients among the best discriminating sites are provided for each factor-transformed data set (but not shown because of space limitations). These correlations provide information about the interrelationships in polarity–accessibility–hydrophobicity, for example, among the discriminating sites prior to the discriminant analysis. Negative coefficients indicate inverse relationships between amino acid sites and are useful for elucidating instances of compensatory change (Atchley et al. 2000). Because of multiple tests, a very conservation critical value of $r = 0.4$ (df = 286) was chosen. Correlations among discriminating variables are partitioned out at the relevant steps by the discriminant analysis.

Next, we assessed the discrimination of these groups using all amino acids or factors simultaneously. This type of discriminant analysis is usually called CVA (=canonical variate analysis). For all practical purposes, the 5 factor scores are independent of each other (orthogonal) (Atchley et al. 2005) and therefore can be considered independent measures of amino acid physiochemical variation. That is, each amino acid site now had 5 numerical descriptors, giving a resultant database with a total of 255 variables (5 factors × 51 amino acid sites).

Finally, a CVA was carried out on each of the 5 factor-transformed variables. Thus, for the polarity–accessibility–hydrophobicity factor–transformed data, a CVA was carried out on all 51 bHLH amino acid sites. Each of these analyses produced 4 eigenvectors that described the simultaneous subset of sites that best discriminated the 5 groups. Plots are provided for several of these vectors, which show the extent of discrimination in each dimension. Also, the square root of the Mahalanobis pairwise distance (Johnson and Wichern 2002) provides the overall divergence for each physiochemical attribute dimension in multivariate standard deviation units.

The statistical procedures employed here are standard methodology typically found in most multivariate statistical software packages. The Statistical Analysis System (SAS) software package was used for these analyses, and the SAS documentation provides extensive discussion of the various procedures.

## Results

### Overall Sequence Variability

Figure 1 shows the dynamic pattern exhibited by the entropy values for the bHLH domain for ungrouped data (Wang and Atchley 2006). The most variable sites are those on the hydrophilic surface, whereas the least variable are usually the packed sites in the highly conserved hydrophobic core (Atchley et al. 2000). This pattern of alternating variable and conserved sites is to be expected in an amphipathic α-helix configuration. The ramifications of this pattern are discussed below.

### Decision Tree

A bifurcating decision tree (Breiman 1998) shows the efficacy of specific sites to produce the classification shown in table 2 (fig. 2). The entry point provides the number of sequences in the a priori defined binding groups. The step 1 node is site 8 where >98% of the sequences in group **A** have a basic residue (*R* or *K*). Crystal structure studies on group **A** proteins (MyoD and E47) indicate that site 8 contacts the phosphate backbone in both proteins and a DNA base in E47. Such contacts apparently do no occur in group **B** bHLH proteins for which crystal structures are available.

In step 2, site 9 is 100% glutamic acid (*E*) in groups **A**, **B**, and **E**. This amino acid contacts the CA base component

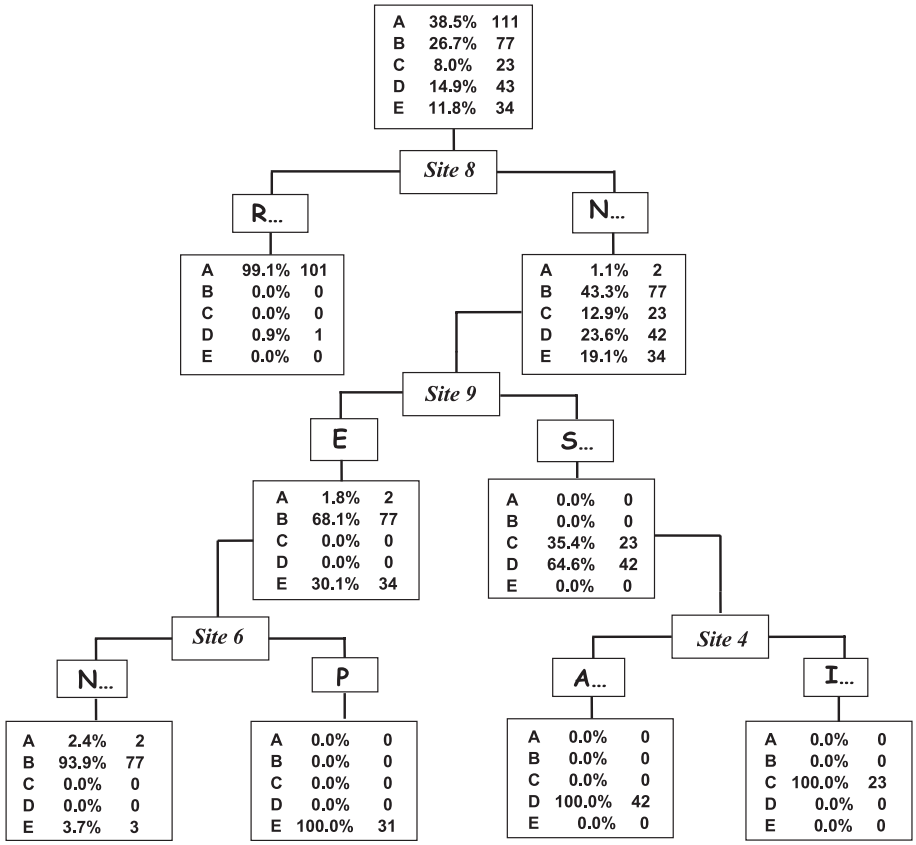


FIG. 2.—Decision tree describing the separation of the 5 DNA-binding groups as described in table 1. Within each box, the sample size and proportion is provided for each of the 5 groups. This result can be analogized to a dichotomous taxonomic key found in the plant or insect taxonomy.

of the hexanucleotide E-box (CANNTG). All bHLH proteins that bind DNA have an *E* at site 9. Site 6 has >91% proline (*P*) residues in group **E**. Site 6 contacts the phosphate backbone (table 2) in those group **A** and **B** proteins for which crystal data are available. Group **E** proteins were apparently derived from group **B** (Atchley and Fitch 1997). Thus, the helix-deforming effects of proline may have been important in the evolutionary divergence of group **E** proteins.

Site 4 that separates groups **C** and **D** has no highly conserved amino acid. Group **C** proteins are the so-called bHLH-PAS proteins like the dioxin receptor Ahr that bind DNA by heterodimerizing with a group **B** protein like Arnt. Group **D** proteins, like ID, do not bind DNA but rather act as negative regulators by heterodimerizing with other bHLH and preventing DNA binding.

Other decision trees can be generated that reflect the discreteness of these DNA-binding groups and substantiate the conclusion that these 5 groups are natural evolutionary and not an artifact of a simplistic classification system.

## Discriminating Sites

The decision tree analysis confirms that a small number of amino acids can accurately define the conventional phylogenetic classification. However, the decision tree analysis provides little information about the molecular architecture underlying classificatory decisions or about interactions among various amino acids. However, discriminant analysis can provide highly quantitative information about the molecular details of this classification.

First, a SWDA is used to determine those sites most diagnostic of the 5 groups for each multidimensional physiochemical factor. SWDA ranks individual sites according to their overall ability to simultaneously separate all 5 binding groups. The biological function of the diagnostic sites is interpreted from previously reported crystal and magnetic resonance imaging analyses (table 2).

Second, multiple group discriminant analysis, commonly called CVA, is carried out. CVA considers all sites simultaneously when discriminating among a priori defined binding groups (Johnson and Wichern 2002). CVA computes the eigenvectors of the among-groups covariance matrix and assesses the relative discriminatory ability of variables along statistically independent multivariate patterns of variation. The magnitude and sign of the canonical variate (CV) coefficients describe the simultaneous combinations of most important sites. Because CVA considers all variables simultaneously, compensatory (inverse) relationships can be depicted. Results for the SWDA and CVA are given next for each of the factor score variables.

### *Polarity, Accessibility, Hydrophobicity (**pah**) Index*

SWDA suggests that 11 sites account for 90% of the ***pah*** variability (table 3). Eight of these sites (2, 3, 5, 8, 9,

**Table 3**
**Separate Stepwise Discriminant Function Analyses for Each Set of Factor Scores to Distinguish between 5 DNA-Binding Groups in bHLH Proteins**

| | Factor 1—***pah*** | | | Factor 2—***pps*** | | | Factor 3—***ms*** | | | Factor 4—***cc*** | | | Factor 5—***ec*** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Step | Enter | $r^2$ | ASCC | Enter | $r^2$ | ASCC | Enter | $r^2$ | ASCC | Enter | $r^2$ | ASCC | Enter | $r^2$ | ASCC |
| 1 | Site 9 | 0.97 | 0.24 | Site 20 | 0.77 | 0.19 | Site 61 | 0.63 | 0.16 | Site 19 | 0.68 | 0.17 | Site 12 | 0.87 | 0.22 |
| 2 | Site 8 | 0.94 | 0.47 | Site 49 | 0.53 | 0.32 | Site 12 | 0.56 | 0.30 | Site 5 | 0.59 | 0.31 | Site 8 | 0.76 | 0.39 |
| 3 | Site 10 | 0.81 | 0.64 | Site 16 | 0.51 | 0.40 | Site 54 | 0.58 | 0.41 | Site 54 | 0.52 | 0.44 | Site 54 | 0.61 | 0.52 |
| 4 | Site 49 | 0.52 | 0.75 | Site 17 | 0.44 | 0.51 | Site 9 | 0.43 | 0.45 | Site 20 | 0.50 | 0.54 | Site 13 | 0.56 | 0.57 |
| 5 | Site 12 | 0.48 | 0.78 | Site 12 | 0.43 | 0.55 | Site 20 | 0.34 | 0.51 | Site 16 | 0.32 | 0.58 | Site 9 | 0.43 | 0.59 |
| 6 | Site 64 | 0.35 | 0.80 | Site 52 | 0.39 | 0.61 | Site 16 | 0.38 | 0.56 | Site 24 | 0.38 | 0.63 | Site 64 | 0.40 | 0.63 |
| 7 | Site 24 | 0.32 | 0.85 | Site 55 | 0.32 | 0.65 | Site 8 | 0.34 | 0.61 | Site 49 | 0.33 | 0.66 | Site 5 | 0.33 | 0.67 |
| 8 | Site 3 | 0.28 | 0.87 | Site 51 | 0.30 | 0.68 | Site 21 | 0.27 | 0.65 | Site 14 | 0.24 | 0.68 | Site 52 | 0.30 | 0.73 |
| 9 | Site 13 | 0.28 | 0.88 | Site 53 | 0.27 | 0.71 | Site 13 | 0.24 | 0.68 | Site 48 | 0.20 | 0.69 | Site 56 | 0.27 | 0.77 |
| 10 | Site 2 | 0.23 | 0.89 | Site 57 | 0.22 | 0.72 | Site 10 | 0.18 | 0.69 | Site 12 | 0.18 | 0.71 | | | |
| 11 | Site 5 | 0.21 | 0.90 | Site 64 | 0.22 | 0.74 | Site 15 | 0.17 | 0.70 | Site 50 | 0.19 | 0.73 | | | |
| 12 | | | | Site 6 | 0.19 | 0.76 | Site 52 | 0.22 | 0.71 | Site 56 | 0.16 | 0.74 | | | |
| 13 | | | | | | | Site 51 | 0.19 | 0.74 | Site 57 | 0.15 | 0.75 | | | |
| 14 | | | | | | | Site 55 | 0.16 | 0.75 | | | | | | |

NOTE.—The data reflect alphabetic sequence data transformed to numerical values representing 5 multivariate physiochemical attribute scales. The stepping order describes the relative discriminatory power of each amino acid site for each physiochemical attribute factor. Each $F$ value is significant at $P < 0.0001$. Sites 1–13 are the DNA-binding region, 14–28 are helix 1, 29–50 are the variable length loop region, and 51–64 are helix 2.

10, 12, and 13) are from the DNA-binding region, and 3 (24, 49, and 64) are from the α-helical dimerization regions. This result clearly demonstrates importance of the aspects of polarity, hydrophobicity, and accessibility in those amino acids that interface with the DNA. However, at least 3 amino acids from outside the DNA-binding region are also important to this classification.

Three contiguous sites from the DNA-binding region, that is, 8, 9, and 10, account for 64% of the variability in ***pah*** (table 3). All 3 sites contact the DNA bases or phosphate backbone (table 2). Site 9 is always glutamic acid (*E*) in bHLH proteins known to bind DNA (groups **A**, **B**, and **E**). Site 10 is predominantly acidic and hydrophilic with amino acids *K* and *R* in **A**, **B**, **C**, and **E**, whereas hydrophobic residues are often found for **D**. Amino acid composition at site 8 distinguishes group **A** (residue *R* or *K*) from groups **B** and **E** (not *R* or *K*) as seen in the decision tree results.

The remaining highly diagnostic sites (in order of their discriminatory power) include 49, 12, 64, 24, 3, 13, 2, and 5 (table 3). Of these, sites 2, 5, 12, and 13 contact bases and the phosphate backbone. The contrast of basic and hydrophilic amino acids versus other amino acid attributes predominates group classification decisions at many of these sites. Groups **A**, **B**, **C**, and **E** have a high preponderance of basic hydrophilic residues (*K*, *R*) in sites 2 and 12, whereas group **D** residues are highly hydrophobic (*L*, *F*). Site 64 is the last amino acid in the bHLH domain, and groups **A**, **B**, **D**, and **E** have a preponderance of hydrophobic residues (*L*, *M*), whereas group **C** has basic hydrophilic residues (*K*, *R*). At site 13, groups **B**, **C**, and **E** have a high frequency of *R* residues not found in groups **A** and **D**. Site 24 lies within the first α-helix and is both acidic and hydrophilic (**A**, **B**, **D**, and **E**) or hydrophobic in group **C**. Site 49 is the site immediately before the start of the second helix. Groups **C** and **E** have only acidic and hydrophilic residues (*D*, *E*), whereas residues *S* and *T* predominate in groups **A** and **D**.

Figure 3 shows a summary of the SWDAs where the most important discriminating amino acid sites are projected onto a helical wheel model of the bHLH domain. The amphipathic nature of the domain can be seen by the localization to one face of the wheel of the packed sites. These packed sites are indicative of the hydrophobic core. Figure 3 clearly shows that the best discriminating sites for these multivariate patterns of physiochemical are located on the hydrophobic face in both helix 1 and helix 2 and in the part of the loop immediately adjacent to helix 2.

CVA on ***pah*** produces these same 11 best discriminating sites from SWDA multidimensional nature of variation in ***pah***. All 4 eigenvectors for ***pah*** are important in differentiation of the 5 binding groups. The first CV for ***pah*** accounts for 62% of the total variation and separates the binding groups into 2 sets, that is, groups **A**, **B**, and **E** and groups **C** and **D**. The rank order of the best discriminating sites is 9, 10, 8, 12, 6, and 50. All 6 sites contact bases or the phosphate backbone, 5 are from the DNA-binding region and the sixth is the first element in the second α-helix. Site 8 has a negative canonical coefficient, indicating that it varies inversely to 6, 9, 10, 12, and 50 with regard to ***pah*** attributes. Site 8 appears to contact bases and the phosphate backbone only in MyoD and E47 (group **A**).

The second CV for ***pah*** explains 25% of the variance and separates out group **A**. Sites 8, 13, 6, 12, 48, 2, and 19 have the highest ranked coefficients. Only sites 19 and 48 do not contact bases or phosphate backbone. There is an inverse relationship between sites 13 and 48 versus sites 8, 6, 2, and 19. The third CV accounts for about 9% of the variability and separates out group **C**. There are 7 sites with coefficients of approximately equal size. There is an inverse relationship with negative values for sites 9, 24, and 52 and positive values for 8, 27, 57, and 64.

Figure 4*A* provides a plot of the projection of the sequence data onto CVs 1 and 2 together with a matrix of the square root of the Mahalanobis pairwise distance. There is a high degree of separation in the 5 groups, and the pairwise distances shows that the centroids are all statistically significant at $P < 0.0001$. CV 1 distinguishes groups **C** and **D** from **A**, **B,** and **E**. CV 2 discriminates group **A** from the others.

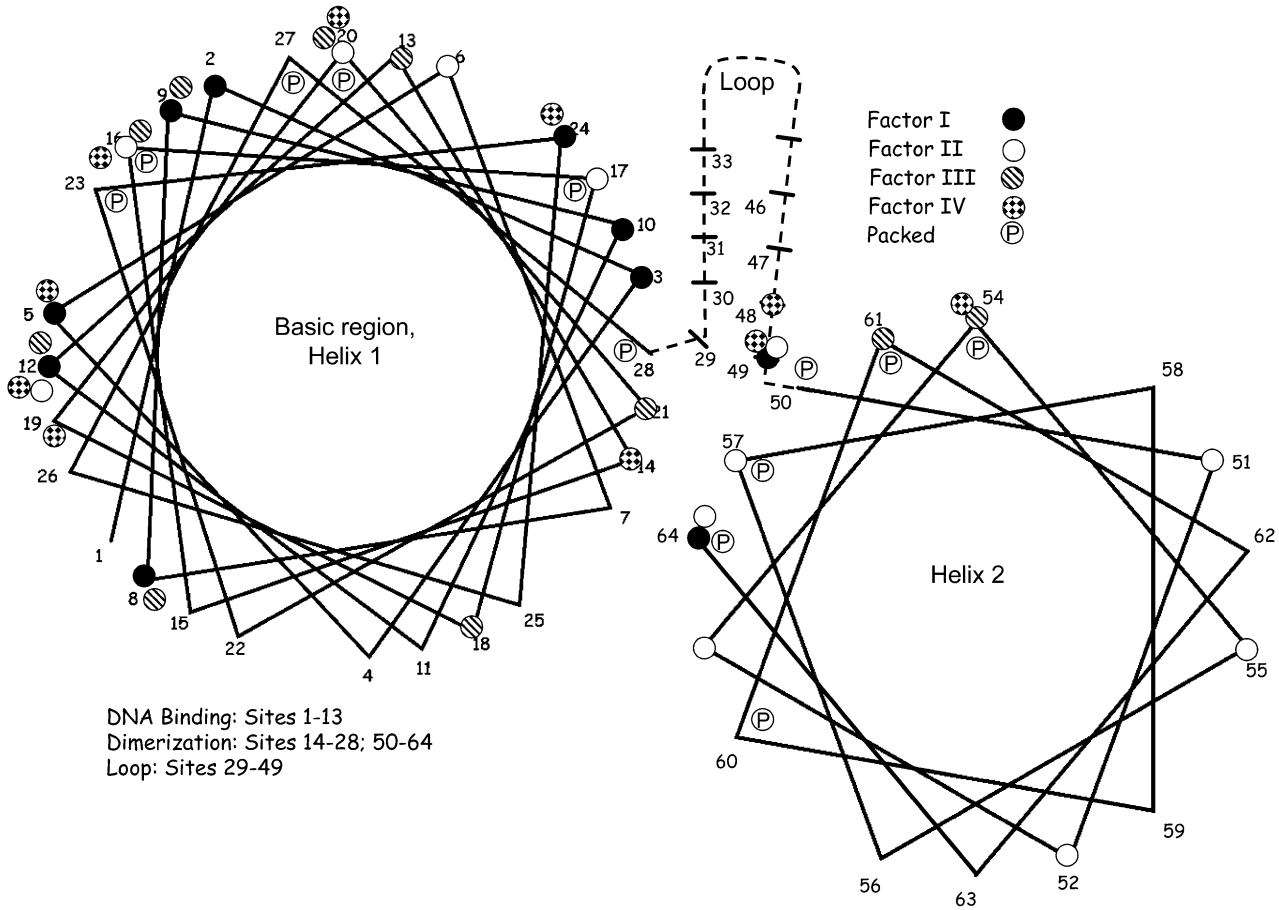


FIG. 3.—Projection of the stepwise discriminant function results onto helical wheels to show the relative placement of amino acid sites with large discriminant coefficients.

*Propensity for Secondary Structure (**pps**)*

SWDA shows that 12 sites account for 76% of the variability in ***pps*** among the 5 DNA-binding groups. Site 20 from within the first α-helix is the best discriminator followed by sites 49, 16, 17, and 12. Unlike ***pah***, only 2 of the 12 best discriminating sites for ***pps*** lie within the DNA-binding region. Significant pairwise correlations between discriminating include 12 and 16 (0.51), 12 and 20 (−0.66), and 16 and 20 (−0.45).

All these 12 sites are important structurally. Site 20 is a buried site that makes numerous van der Waals contacts with residues in the second helix. As expected, the residue at site 20 is highly hydrophobic (*F*, *L*, or *I*) in all 5 groups. In group **D**, the residue is tyrosine.

Site 49 occurs immediately before initiation of the second α-helix. It contains predominantly *S*, *T*, and *P* residues in groups **A**, **B**, and **D** and *D* and *E* in groups **C** and **E**. Residues *R*, *S*, and *T* have low values for the helix–coil equilibrium constant, relative frequency in an α-helix, and high values for the frequency of a turn and the Chou–Fasman parameter for coil conformation. Interestingly, site 49 is predominantly *D* (group **C**) and *E* (group **E**), which are acidic residues but with divergent values on the secondary structure scales given for the other 3 groups. Site 16 is a buried amino acid where the amino acid side chain packs against a site 20 in Max. There is a hydrophobic residue in groups **A**, **B**, **D**, and **E**. (*I*, *L*, *M*, *V*), whereas the residue in group **C** is predominantly glutamic acid. Site 12 in the DNA-binding region contacts the phosphate backbone. It has high frequencies of *R* except for group **D,** which has *L* and *Q*.

Five significant discriminatory sites in the second α-helix are 51, 52, 53, 55, and 57. Sites 53 and 54 pack against sites 50, 53, and 54. Site 51 has residues *V*, *I*, *L*, and *M* predominating in groups **A**, **B**, and **D** and residue *A* in groups **C** and **E**. All these amino acids have low values on the Chou–Fasman parameter for coil conformation. Site 52 has a high frequency of *E* residues in groups **A** and **D**, *Q* in group **E**, and *S* in group **C**. Glutamic acid (*E*) exhibits high average relative probabilities of a helix, low normalized frequencies for a turn, and low values for free energy in an α-helix region. Serine (*S*) has low relative probabilities of a helix and high Chou–Fasman coil configuration values. Site 55 has high frequency of basic amino acids (*K*, *R*) in groups **A**, **B**, and **C**. However, group **D** is all *Q*, whereas group **E** is all *E*.

CVA shows, as before, that the best discriminating variables are distributed among all 4 CVs. The first CV separates out group **D** from the others, particularly from group **E**. The sites with the largest CV coefficients are 16 and 20

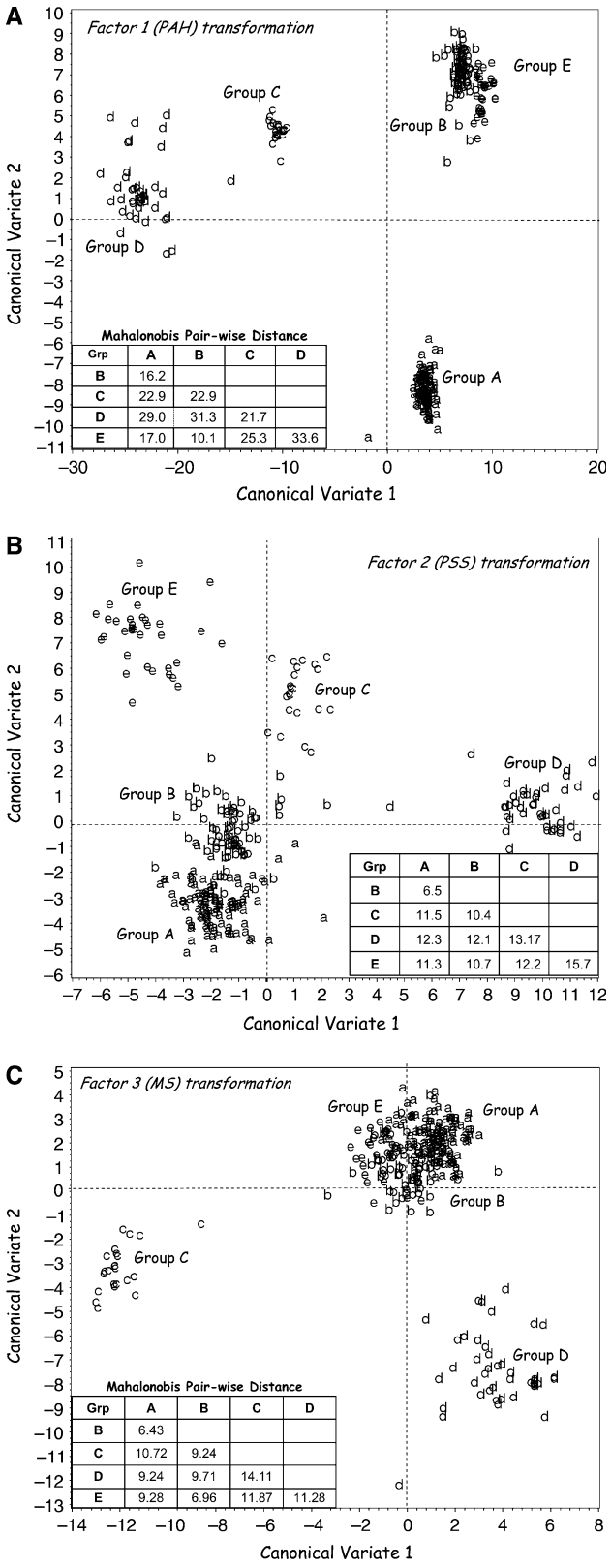


FIG. 4.—Projection of individual sequences onto canonical vectors for physiochemical attribute factors 1–3. The plots contain the first 2 canonical vectors for each factor together with the square root of the Mahalanobis pairwise distance computed for all 4 canonical vectors; (*A*) reflects the (***pah***) physiochemical attribute, (*B*) involves the ***pps*** factor, and (*C*) shows the effects of molecular size (***ms***).

with 16 being negative, indicating an inverse relationship between the 2 sites. Both 16 and 20 are buried sites in the first helix and interact with amino acids in the second helix. The second CV separates groups **C** and **E** from the others. There are bipolar coefficients with negative values for sites 24, 27, 28, and 53 but positive values for sites 13, 46, and 57. Sites 27 and 28 are buried sites that pack against amino acids in the second helix. Site 28 is typically a *P*, which signals the end of the first helix and the beginning of the loop. Site 53 is in the second helix and packs against sites 60 and 64.

Figure 4*B* documents the patterns and the extent of discrimination among the binding groups for propensity for secondary structure. All centroids are significantly different at $P < 0.001$, although there is less overall divergence among the groups compared with ***pah***, particularly between groups **A** and **B**.

### *Molecular Size (**ms**)*

Five sites account for 50% of the among-groups variability in molecular size (table 3). These 5 sites and their ASCC values include 61 (16%), 12 (14%), 54 (11%), 9 (4%), and 20 (6%). Within the top 10 sites (accounting for almost 70% of the variability), 8 are functionally or structurally defined sites. Sites 16, 20, 54, and 61 are buried and pack against sites in the other monomer or are involved with various interactions. Atchley et al. (2000) showed that within these interactive sites, compensatory changes involving molecular size were very important to maintain the structural geometry of the molecule. Sites 9 and 12 contact the DNA base or phosphate backbone, where again size relationships are expected to be important. Only 54 and 61 show a significant correlation (−0.42) among the first 10 discriminating sites.

The first CV strongly separates group **C** from the others and has high coefficients on sites 9, 52, 54, 57, and 61. Three of these sites (9, 54, and 61) are buried sites that either contact the DNA (9) or pack against other amino acids. Site 61 has a negative coefficient, whereas the others are positive. The second CV has highest coefficients on sites 5, 6, 12, 13, 54, and 55. Sites 5, 6, 12, and 13 all contact either the DNA base or the phosphate backbone. Site 54 is a buried site that packs against sites 60 and 63.

Figure 4*C* provides a depiction of the extent of divergence among the 5 groups. Clearly, the molecular size physiochemical factor is most involved with separation of groups **C** and **D**. All pairwise distances are statistically significant at $P < 0.001$.

### *Codon Composition (**cc**)*

Four sites explain over 50% of the among-groups variation in codon composition, and 13 sites explain 75%. The top 4 sites and their ASCC values are 19 (17%), 5 (14%), 54 (13%), and 20 (10%). Significant correlations occur for only 2 pairs of sites, that is, 16 and 19 (0.64) and 24 and 54 (−0.41).

### *Electrostatic Charge (**ec**)*

Nine sites explain 77% of the variability in electrostatic charge, of which 3 sites (12, 8, and 54) account

for over 50%. Significant correlations exist between 8 and 12 (0.47), 90 and 12 (−0.43), and 5 and 54 (−0.46).

### All Factors Considered Simultaneously

When all the factors and sites are considered simultaneously, a concise multidimensional picture is provided that clarifies the relative importance of all the amino acids and their physiochemical attributes. This analysis clarifies several points: 1) various physiochemical aspects of only 5 sites account for 86% of the among-groups variation. These sites are 8, 9, 10, and 12 from the DNA-binding region and site 49 from the loop immediately before the second α-helix. 2) Of this 86%, variability in factor 1 (polarity, accessibility, and hydrophobicity) is responsible for all but 8%. Most of this 8% arises from factor 2 (propensity for secondary structure).

## Discussion

Multivariate statistical analyses are a powerful tool for integrating sequence, structural, and functional information. Herein, we have provided a paradigm for how multivariate statistical, structural, and sequence analyses can be integrated for understanding the molecular architecture and geometry of DNA binding. The results employ a large battery of physiochemical attributes to give a highly multidimensional quantitative description of protein structure, function, and evolution. This approach permits rigorous statistical inference on biologically important questions using complex sequence data.

Atchley and Buck (submitted) carried out a multivariate factor analysis on pairwise correlations (computed as mutual information values) among the amino acids in these 288 bHLH sequences. Their analyses focused on the covariance structure without regard to hierarchical structure within the data. They dealt with questions about dimensionality of the overall covariance structure (how many independent patterns of covariation occur in the bHLH domain) and whether the resultant patterns of amino acid variability could be interpreted in a meaningful biological manner.

Atchley and Buck (submitted) found 7 major patterns of multivariate covariation that accounted for much of the common variation among amino acids in the bHLH domain. The patterns themselves, which summarize many millions of years of evolutionary change over the entire transcription factor family, retained considerably biological interpretability. For example, the first and largest covariance pattern included those more highly variable amino acids found on the hydrophilic surface of the proteins. This pattern exhibited a high degree of concordance with binding group membership, clade membership, and loop length. This finding suggested that these amino acids exhibited a strong phylogenetic signal and was highly meaningful in understanding evolutionary variability in this protein family. The second covariance pattern related to the structural and functional aspects of DNA binding.

In the present paper, sequences were grouped into 5 experimentally defined DNA-binding groups. Thus, we can ask about the underlying latent structure of among-groups variability as it relates to understanding the mechanisms of evolutionary diversification among groups that differ in their DNA binding. The alphabetic amino acid codes were transformed into an array of 5 numerical values representing their physiochemical attributes and the resultant data subjected to discriminant analysis.

In the Introduction, we raised 4 fundamental questions. First, which amino acids in the bHLH domain best discriminate these 5 a priori defined DNA-binding groups? All analyses point to approximately 6 sites (8, 9, 10, 12, 24, and 49) as giving >90% discrimination when considered simultaneously. Transforming the alphabetic amino acid codes to the numerical factor score variables of Atchley et al. (2005) elucidates the underlying physiochemical attributes that explain these evolutionary patterns.

Second, are the best discriminating amino acids always found in the basic DNA-binding region? The answer is no, there are amino acids outside the binding region with significant discriminatory power, particularly sites 20, 24, and 49. Site 20 is a packed site with many van der Waal contacts with helix 2 sites, and site 49 occurs just before the start of helix 2. Clearly, variation of the amino acid composition of both sites has considerable potential structural impact.

Third, which physiochemical attributes of amino acids are reflected by these best discriminators? These analyses show that there is significant discriminating power in all 5 multivariate physiochemical attributes; however, the most important is clearly the polarity–accessibility–hydrophobicity (***pah***) component followed by propensity for secondary structure (***pss***).

Fourth, what are the relationships between these discriminating amino acids and DNA–protein interactions? The results clearly show that many of the best discriminating sites are those constituent amino acids of known structural and functional roles.

These results provide detailed structural and physiochemical information to clarify the general nature of DNA binding in bHLH proteins. They suggest the relative importance of the various amino acids inside the basic DNA-binding region as well as the α-helical dimerization region.

The best binding group discrimination is found in variables relating to polarity, accessibility, and hydrophobicity attributes (***pah***) in those amino acid sites that contact the DNA, that is, 9, 8, 10, 12, and 49. These 5 sites account for 78% of the among-groups variation for the ***pah*** pattern. Out of the 11 sites that are most important in discriminating among the groups for the ***pah*** data, only amino acid 24 has not been shown to contact the DNA. Hence, we know not only which amino acids best distinguish these binding groups but also what physiochemical attributes are most important. With regard to other physiochemical attributes, amino acid sites for which the propensity for secondary structure pattern has the most discriminatory power are found in the 2 α-helices rather than the DNA-binding region.

Examining the spatial distribution of important discriminating sites provides additional interesting information. Atchley and Buck (submitted) found that the hydrophilic face on these proteins was most variable. However, for data structured by DNA-binding groups, figure 3 shows that the best discriminating amino acids predominate on the hydrophobic face. The latter coincides with factor 2 in the study by Atchley and Buck (submitted).

These results provide new information about molecular architecture that can be integrated into new structural analyses on bHLH proteins, subsequent molecular analyses of the 5 major DNA-binding groups including molecular dynamics simulations, and hypothesis testing about the molecular basis of DNA-binding specificity. The use of the factor score transformations has greatly facilitated the multivariate statistical analyses described in this work. The use of these transformations provide for more rigorous statistical analyses of the underlying causes of amino acid variation. These transformations were used in analyses of the patterns and causes of periodicity of amino acid variation in bHLH proteins (Wang and Atchley 2006).

Understanding the combinatorial nature of variation in DNA-binding specificity provides evolutionary insight into the functional divergence in the groups of transcriptional regulators. The information provided by these analyses provides insight into the evolutionary divergence in structure and function that has potentiality at important nodes in the evolutionary history of the bHLH proteins as seen in phylogenetic trees estimated by Atchley and Fitch (1997), Ledent et al. (2002), and others. One of the problems with estimating phylogenetic divergence with trees is the difficulty is perceiving multivariate change occurring at important nodes in the trees. The results provided here shown that multivariate statistical analyses can provide important new information about molecular architecture as well as the components of evolutionary divergence in proteins.

These results provide a useful paradigm for statistical analyses of structural, functional, and evolutionary aspects of protein variation. One can now meaningfully partition amino acid sequence variability into a small set of underlying physiochemical components that should greatly facilitate many types of protein structural analyses and provide a much deeper understanding of the underlying biological meaning for observed amino acid variability.

## Acknowledgments

The authors are indebted to Steve Spiker for his critical comments on the manuscript. This work was supported by National Institutes of Health Grant GM45344 and by funds from North Carolina State University.

## Literature Cited

Atchley WR, Fitch WM. 1997. A natural classification of the basic helix-loop-helix class of transcription factors. Proc Natl Acad Sci USA. 94(10):5172–5176.

Atchley WR, Terhalle W, Dress A. 1999. Positional dependence, cliques, and predictive motifs in the bHLH protein domain. J Mol Evol. 48(5):501–516.

Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. 2000. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol. 17(1):164–178.

Atchley WR, Zhao JP, Fernandes A, Duke T. 2005. Solving the protein sequence metric problem. Proc Natl Acad Sci USA. 102(18):6395–6400.

Beltran AC, Dawson PE, Gottesfeld JM. 2005. Role of DNA sequence in the binding specificity of synthetic basic-helix-loop-helix domains. Chembiochem. 6(1):104–113.

Breiman L. 1998. Classification and regression trees. Boca Raton (FL): Chapman & Hall/CRC.

Brownlie P, Ceska TA, Lamers M, Romier C, Stier G, Teo H, Suck D. 1997. The crystal structure of an intact human Max-DNA complex: new insights into mechanisms of transcriptional control. Structure. 5(4):509–520.

Buck MJ, Atchley WR. 2005. Networks of coevolving sites in structural and functional domains of serpin proteins. Mol Biol Evol. 22(7):1627–1634.

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(8): 970–980.

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(1):180–189.

Ferre-D'Amare AR, Prendergast GC, Ziff EB, Burley SK. 1993. Recognition by Max of its cognate DNA through a dimeric B/Hlh/Z domain. Nature. 363(6424):38–45.

Johnson RA, Wichern DW. 2002. Applied multivariate statistical analysis. Upper Saddle River (NJ): Prentice Hall.

Kewley RJ, Whitelaw ML, Chapman-Smith A. 2004. The mammalian basic helix-loop-helix/PAS family of transcriptional regulators. Int J Biochem Cell Biol. 36(2):189–204.

Ledent V, Paquet O, Vervoort M. 2002. Phylogenetic analysis of the human basic helix-loop-helix proteins. Genome Biol. 3(6):1–18.

Ma PCM, Rould MA, Weintraub H, Pabo CO. 1994. Crystal-structure of Myod Bhlh domain-DNA complex—perspectives on DNA recognition and implications for transcriptional activation. Cell. 77(3):451–459.

Massari ME, Murre C. 2000. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. Mol Cell Biol. 20(2):429–440.

Murre C, Bain G, van Dijk MA, Engel I, Furnari BA, Massari ME, Matthews JR, Quong MW, Rivera RR, Stuiver MH. 1994. Structure and function of helix-loop-helix proteins. Biochim Biophys Acta. 1218(2):129–135.

Parraga A, Bellsolell L, Ferre-D'Amare AR, Burley SK. 1998. Co-crystal structure of sterol regulatory element binding protein 1a at 2.3 angstrom resolution. Structure. 6(5):661–672.

Shimizu T, Toumoto A, Ihara K, Shimizu M, Kyogoku Y, Ogawa N, Oshima Y, Hakoshima T. 1997. Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. EMBO J. 16(15):4689–4697.

Turner EC, Cureton CH, Weston CJ, Smart OS, Allemann RK. 2004. Controlling the DNA binding specificity of bHLH proteins through intramolecular interactions. Chem Biol. 11(1):69–77.

Wang Z, Atchley WR. 2006. Spectral analysis of sequence variability in basic helix-loop-helix (bHLH) protein domains. Evol Bioinformatics (Online) 2:201–210.

Wollenberg KR, Atchley WR. 2000. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc Natl Acad Sci USA. 97(7):3288–3291.

William Martin, Associate Editor

Accepted October 5, 2006