Review

An overview of the basic helix-loop-helix proteins
Susan Jones

Address: Department of Biochemistry, School of Life Sciences, University of Sussex, Falmer, Brighton BN1 9QG, UK.
E-mail: S.Jones@sussex.ac.uk

Published: 28 May 2004

Genome Biology 2004, 5:226

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/6/226
© 2004 BioMed Central Ltd

Abstract

The basic helix-loop-helix proteins are dimeric transcription factors that are found in almost all eukaryotes. In animals, they are important regulators of embryonic development, particularly in neurogenesis, myogenesis, heart development and hematopoiesis.

The basic helix-loop-helix (bHLH) proteins form a large superfamily of transcriptional regulators that are found in organisms from yeast to humans and function in critical developmental processes, including sex determination and the development of the nervous system and muscles. Because of their functional diversity and importance, this superfamily has been the subject of a number of recent reviews covering many species [1,2], and also a number of reviews specific to individual species, including Saccharomyces cerevisiae [3], Drosophila [4,5], human [6] and Arabidopsis [7-9]. The main emphasis in the recent literature has been on phylogenetic sequence analysis of bHLH families. This article gives an overview of how bHLH proteins are classified by sequence and summarizes their structures and functions.

Classifications of bHLH proteins by sequence

Members of the bHLH superfamily have two highly conserved and functionally distinct domains, which together make up a region of approximately 60 amino-acid residues. At the amino-terminal end of this region is the basic domain, which binds the transcription factor to DNA at a consensus hexanucleotide sequence known as the E box. Different families of bHLH proteins recognize different E-box consensus sequences. At the carboxy-terminal end of the region is the HLH domain, which facilitates interactions with other protein subunits to form homo- and hetero-dimeric complexes. Many different combinations of dimeric structures are possible, each with different binding affinities between monomers. The heterogeneity in the E-box sequence that is recognized and the dimers formed by different bHLH proteins determines how they control diverse developmental functions through transcriptional regulation [10].

The bHLH motif was first observed by Murre and colleagues [11] in two murine transcription factors known as E12 and E47. With the subsequent identification of many other bHLH proteins, a classification was formulated on the basis of their tissue distributions, DNA-binding specificities and dimerization potential [12]. This classification, which divides the superfamily into six classes, was initially based on a small number of HLH proteins but has since been applied to larger sets of eukaryotic proteins [1]. More recently, an approach using evolutionary relationships was used to classify bHLH proteins into four major groups (A-D) [13], taking into account E-box binding, conservation of residues in the other parts of the motif, and the presence or absence of additional domains. The sequencing of new genomes has led to the identification of additional bHLH families, and this evolutionary classification has now been extended to include two additional groups (E and F; Table 1) [6]. Parsimony analysis by Atchley and Fitch [13] of a phylogenetic tree derived from 122 sequences suggested that an ancestral HLH sequence most probably came from group B, and group B proteins are indeed the most prevalent type of bHLH proteins in animals. The situation is similar in the Arabidopsis genome, in which the G-box-binding bHLH proteins (part of group B) are the most abundant group [7].
One basis for the evolutionary classification shown in Table 1 is the presence or absence of additional domains, of which the most common are the PAS, orange and leucine-zipper domains. PAS domains, located carboxy-terminal to the bHLH region, are 260-310 residues long and function as dimerization motifs [14]. They allow binding with other PAS proteins, non-PAS proteins, and small molecules such as dioxin. The PAS domain is named after three proteins containing it: *Drosophila* Period (Per), the human aryl hydrocarbon receptor nuclear translocator (Arnt) and *Drosophila* Single-minded (Sim) [15]. The domain is itself made up of two repeats of approximately 50 amino-acid residues (known as PAS A and PAS B) separated by about 150 residues that are poorly conserved [16]. PAS-domain-containing bHLH proteins (bHLH-PAS proteins) form phylogenetic group C. A distinct additional domain, the orange domain, is a 30-residue sequence that is also located carboxy-terminal to the PAS region, from which it is separated by a short, variable length of sequence. Transcription factors with this additional domain, designated bHLH-O and forming part of phylogenetic group E, include the hairy-related proteins, called HEY1, HEY2 and HEYL in mouse and humans [17]. The molecular function of the orange domain is still unclear; it has been proposed that it mediates specificity and transcriptional repression [18], but there is also evidence that it can play a role in dimerization [17].

A number of bHLH protein families, mostly in phylogenetic group B, have a leucine-zipper domain contiguous with the second helix of the HLH domain; like the HLH domain, this mediates dimerization. Proteins that have only a leucine-zipper domain coupled with a basic domain (denoted bZIP) and no HLH domain are a separate family of DNA-binding proteins in their own right (reviewed in [19]). The sequence of the zipper consists of a repeating heptad, with hydrophobic and apolar residues occurring at the first and fourth positions and polar and charged residues at the remaining positions. Leucine is the residue that predominates at position 4; it thus lends its name to the zipper motif. One bHLH protein that has a leucine-zipper domain (and that is therefore denoted a bHLHZ protein) is Max, which forms the hub of a network of bHLH transcription factors. Max is known to form homodimers and heterodimers with the group B proteins Myc, Mad, Mnt and Mga, and these complexes each have sequence-specific DNA-binding and transcriptional functions [20].

The additional domains in bHLH proteins, such as the leucine zipper, are always carboxy-terminal to the bHLH region. The position of the bHLH and additional domains within the complete sequence of the protein varies widely between different families, however. This variable pattern of domain positioning has led to the proposal that bHLH proteins have undergone modular evolution by domain shuffling, a process that involves domain insertion and rearrangement [21].

### Structures of bHLH proteins
In comparison with the volume of sequence data, structural data for the bHLH superfamily of transcriptional regulators are still relatively sparse. Just nine bHLH protein structures
have been deposited to date in the Protein Data Bank (PDB; see Table 2) [22]. The CATH [23] and SCOP [24] protein-structure classifications classify eight of these structures into one superfamily (Table 2; SREBP-2 has not been classified). A number of the structures (PDB codes 1an2, 1hlo, 1nlw, 1nkp, and 1am9) include an additional zipper domain that is carboxy-terminal to the HLH region. Two of the structures solved are heterodimers: a Max-Myc complex (PDB code 1nkp) and a Max-Mad complex (PDB code 1nlw). The remaining complexes are homodimeric, and all but one include the structure of the bound DNA double helix, giving insights into the binding specificity at the E box. Representatives of these bHLH structures are shown in Figure 1.

The structure of MyoD (Figure 1a) is typical of many bHLH proteins, comprising two long α helices connected by a short loop, which in the case of MyoD is 8 residues in length. The first helix (H1) includes the basic domain, which makes contact with the major groove of the DNA. MyoD is a homodimer in which the two monomers make identical contacts with the DNA. Comparisons of this structure with that of Max (which includes an additional leucine zipper domain; Figure 1d-f) reveal that the presence or absence of this domain does not significantly affect the structure of the bHLH segment [25].

Two interesting features revealed by the three-dimensional structure of the Pho4 bHLH domain (Figure 1b) are the existence of a short stretch of α-helix in the loop region that links helix H1 to helix H2 and the recognition of DNA bases outside the E-box sequence [26]. The Pho4 protein binds DNA as homodimer, and its two subunits form a parallel four-helix bundle (Figure 1b). The short α-helix region in the loop lacks the stabilizing hydrogen-bonding network observed in other bHLH proteins. In the Pho4 structure, each half-site of the symmetrical E box is recognized by a triad of residues, but bases beyond the E box, including a GG sequence at the 3′ end, are also recognized [26]. Base recognition outside the E box is also observed for MyoD, but in this structure it occurs at the 5′ end of the E box [25].

Sterol regulatory element binding protein 1a (SREBP-1a; Figure 1c) is an example of a bHLH structure that includes one of the additional domains, the leucine zipper. SREBPs are bHLHZ transcription activators that bind to a DNA target site as a homodimer and are essential for cholesterol metabolism.

### Table 2

| PDB code | Protein Chains | Protein name | Species | Group | SCOP superfamily | CATH homologous superfamily | CATH sequence family |
|----------|----------------|--------------|---------|-------|------------------|----------------------------|---------------------|
| 1mdy     | ABCD           | MyoD bHLH    | Mouse   | A     | Helix-loop-helix DNA-binding domain | MyoD basic-helix-loop-helix domain, subunit B | 4.10.280.10.1 |
| 1an4     | AB             | USF bHLH     | Human   | B     | Helix-loop-helix DNA-binding domain | MyoD basic-helix-loop-helix domain, subunit B | 4.10.280.10.2 |
| 1an2     | AC             | Max bHLHZ    | Mouse   | B     | Helix-loop-helix DNA-binding domain | MyoD basic-helix-loop-helix domain, subunit B | 4.10.280.10.2 |
| 1hlo     | AB             | Max bHLHZ    | Human   | B     | Helix-loop-helix DNA-binding domain | MyoD basic-helix-loop-helix domain, subunit B | 4.10.280.10.2 |
| 1nlw     | BE             | Max bHLHZ    | Human   | B     | Helix-loop-helix DNA-binding domain | MyoD basic-helix-loop-helix domain, subunit B | 4.10.280.10.2 |
| 1nkp     | BE             | Max bHLHZ    | Human   | B     | Helix-loop-helix DNA-binding domain | MyoD basic-helix-loop-helix domain, subunit B | 4.10.280.10.2 |
| 1am9     | AD             | Muc prot-oncogene bHLHZ | Human | B | Helix-loop-helix DNA-binding domain | MyoD basic-helix-loop-helix domain, subunit B | 4.10.280.10.2 |
| 1am9     | ABCD           | SREBP-1a bHLHZ | Human | B | Helix-loop-helix DNA-binding domain | MyoD basic-helix-loop-helix domain, subunit B | 4.10.280.10.3 |
| 1ukd     | CDEF           | SREBP-2 HLHZ | Human   | B     | Helix-loop-helix DNA-binding domain | NC | NC |
| 1a0a     | AB             | Pho4 bHLH    | S. cerevisiae | B | Helix-loop-helix DNA-binding domain | MyoD basic-helix-loop-helix domain, subunit B | 4.10.280.10.4 |
| 1nlw     | AD             | Mad bHLHZ    | Human   | B     | Helix-loop-helix DNA-binding domain | NC | NC |

The PDB codes and protein names for nine bHLH proteins deposited in the PDB are shown with their superfamily names from the CATH [23] and SCOP [24] classifications and their sequence family numbers from CATH. Max has more than one structure solved, including two complexes (1nkp, Max-Myc and 1nlw, Max-Mad). Structures shown in Figure 1; NC, the protein is not yet included in the CATH or SCOP protein structure classifications. SREBP, steroid response element binding protein. Protein chains indicates the chain identification letter assigned to individual subunits in the PDB files.
metabolism [27]. Unlike other bHLH proteins that recognize a symmetrical E box, SREBP-1a recognizes an asymmetrical sterol regulatory element. This asymmetric recognition is possible because of the presence of a tyrosine residue in the basic domain. The tyrosine replaces the arginine observed in other bHLH proteins such as Max, and this change results in the loss of polar interactions with the DNA [27]. Recently, a crystal structure of another SREBP, SREBP-2, has been solved [28], in which SREBP-2 is bound in a complex with importin-β, a molecule that mediates the transport of molecules into and out of the nucleus; the structure reveals that SREBP-2 is imported into the nucleus as a homodimer.

Two of the most interesting structures to be solved to date are those of the Max-Mad (Figure 1d) and Max-Myc (Figure 1e) heterodimer complexes bound to double-stranded DNA [29]. In each monomer, the amino-terminal α helix is a continuous secondary-structural element that includes the basic region and the α helix H1, and the carboxy-terminal α helix is made up of two continuous α-helical segments, helix H2 and the leucine-zipper region. The Myc-Max and Mad-Max complexes are quasi-symmetric heterodimers that have interfaces made up of hydrophobic and polar interactions involving residues in helices H1 and H2 and the leucine zipper. Mutation studies suggest that dimer specificity is controlled by the amino acids Gln91 and Asn92 (in the Max numbering) in the Myc-Max dimer. The studies also show that Glu125 controls Mad-Max heterodimer formation [29]. One interesting feature of the Myc-Max crystal structure (Figure 1f) is the presence of two heterodimers in the asymmetric unit of the crystals. The two structures form a heterotetramer in which the head-to-tail assembly of leucine zippers from different heterodimers results in the formation of an antiparallel four-helix bundle (Figure 1f). It has been shown previously that Myc-Max heterodimers can form higher multimeric structures [30], and there is evidence to suggest that the tetramer observed in the crystal also exists under physiological conditions [29].

**Functions of bHLH proteins**

The heterogeneity of DNA sequences recognized and dimers formed by the bHLH proteins enable them to function as a diverse set of regulatory factors. The bHLH proteins can be divided into those that are cell specific and those that are widely expressed. The cell-type-specific members of the superfamily are involved in cell-fate determination in many different cell lineages and form an integral part of many
processes, including neurogenesis, cardiogenesis, myogenesis, and hematopoiesis (Table 3). The bHLH proteins involved in neurogenesis include \textit{Drosophila Atonal} and other ‘proneural’ proteins [31]. In vertebrates, Mash-1, Math-1 and the neurogenins are important in the initial determination of neurons, whereas Nero-D, NeuroD2, MATH-2 and others are differentiation factors [32]. The bHLH transcription factors dHAND and eHAND are important in cardiac development in vertebrates [33]. The myogenic regulatory factors, including MyoD, MRF-4, Myf-5 and myogenin, together regulate both the establishment and differentiation of the myogenic lineage [34]. The stem cell leukemia (SCL) protein is a bHLH transcription factor that is essential for hematopoiesis and is associated with acute T-cell leukemia [35].

One family of bHLH proteins that is widely expressed in many different cell types is the Myc family. The \textit{Myc} genes are among the most frequently affected genes in human tumors [36]. Myc proteins are known to regulate translation initiation [37] and they also function as transcriptional activators when they form heterodimers with Max proteins (also members of group B) [38]. There is some evidence, however, that these dimers may also operate as negative regulators of transcription (reviewed in [39]). Max is also known to form homodimers and heterodimerize with other bHLH proteins including Mad [38]. This dimerization network of Myc/Max/Mad transcription proteins has a large number of target genes involved in the cell cycle, and the network has been considered to function as a transcription module [20].

In summary, the bHLH superfamily constitutes a large and diverse class of proteins, with over 125 different proteins identified in humans and 145 in \textit{Arabidopsis}. The discovery of their diverse functions in the cell cycle, cell-lineage development and tumorigenesis has elevated the interest in them in the 15 years since they were first identified by Murre and co-workers [11]. So what do the coming years hold in store for this superfamily? With the sequencing of more genomes, it is expected that further superfamily members and new sequence families will be identified. With an increasing number of proteins targeted and solved by structural-genomics consortia, the structural data available for this superfamily will also grow. The knowledge gained from new sequences and novel high-resolution structures will offer further insights into the mechanisms by which they control such diverse processes. This increasing knowledge base may make them good targets for new drug therapies for conditions including heart disease and cancer.

\textbf{Acknowledgements} 
I would like to thank Mario Garcia, Hugh P. Shanahan and Janet M. Thornton (European Bioinformatics Institute, UK) for their help in extracting and analyzing the structural data on the bHLH proteins.
226.6 Genome Biology 2004, Volume 5, Issue 6, Article 226 Jones

4. Moore AW, Barbel S, Jan LY, Jan YN: A genomewide survey of basic helix-loop-helix factors in Drosophila. Proc Natl Acad Sci USA 2000, 97:10346-10341.

5. Peyretire S, Kahn D, Haenlin M: New members of the Drosophila Myc transcription factor subfamily revealed by a genomewide examination for basic helix-loop-helix genes. Mech Dev 2001, 104:99-104.

6. Ledent V, Paquet O, Vervoort M: Phylogenetic analysis of the human basic helix-loop-helix proteins. Genome Biol 2002, 3:research0030.1-0030.18.

7. Toledo-Ortiz G, Hug E, Quail PH: The Arabidopsis basic/helix-loop-helix transcription factor family. Plant Cell 2003, 15:1749-1770.

8. Heim MA, Jakoby M, Werber M, Martin C, Weisshaar B, Bailey PC: The basic helix-loop-helix transcription factor family in plants: a genomewide study of protein structure and functional diversity. Mol Biol Evol 2003, 20:735-747.

9. Morganstern B, Archelley WR: Phylogenetic analysis of plant basic helix-loop-helix proteins. J Mol Evol 2003, 56:742-750.

10. Fairman R, Beran-Steed RK, Anthony-Cahill SJ, Lear JD, Stafford WF, Degrado VF, Benfai PA, Brenner SL: Multiple oligomeric states regulate the DNA-binding of helix-loop-helix peptides. Proc Natl Acad Sci USA 1993, 90:10429-10433.

11. Murre C, Mc Caw PS, Baltimore D: A new DNA binding and dimerizing motif in immunoglobulin enhancer binding, Daughterless, MyoD and Myc proteins. Cell 1989, 77:577-783.

12. Murre C, Bain G, Vandijk MA, Engel I, Furnari BA, Massari ME, Matthews JR, Quong MV, Rivera RB, Snuier PH: Structure and function of helix-loop-helix proteins. Biochim Biophys Acta 1994, 1218:129-135.

13. Archelley WR, Fitch WM: A natural classification of the basic helix-loop-helix class of transcription factors. Proc Natl Acad Sci USA 1997, 94:1725-1727.

14. Kewley RJ, Whitelaw ML, Chapman-Smith A: The PAS domain confers target specificity of Drosophila bHLH/PAS proteins. Genes Dev 1997, 11:2079-2089.

15. Parraga A, Bellsolell L, Ferré-D’Amare AR, Burley SK: The basic-region/leucine zipper family of DNA binding proteins. Nucleic Acids Mol Biol 1997, 22:629-651.

16. Pfirter S, Archelley WR: Evolution of bHLH transcription factors: modular evolution by domain shuffling? Mol Biol Evol 1999, 16:1654-1663.

17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. Nucleic Acids Res 2000, 28:235-242.

18. Oshimi Y, Shimizu M, Kyogoku Y, Ogasawara N, Oshima Y, Hakoshima T: Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. EMBO J 1997, 16:4689-4697.

19. Perricaudet M, Archelley WR: The basic region/helix-loop-helix-leucine zipper domain of Myc proto-oncoproteins: function and regulation. Oncogene 1999, 18:2955-2966.

20. Schoffit EV: The role of c-myc in regulation of translation initiation. Oncogene 2004, 23:3217-3221.

21. Nair SK, Burley SK: X-ray structures of Myc-Max and Mad-Max recognizing DNA: molecular bases of regulation by proto-oncogenic transcription factors. Cell 2003, 112:193-205.

22. Davis RL, Turner DL: Vertebrate hairly and Enhancer of split related proteins: transcriptional repressors regulating cellular differentiation and embryonic patterning. Oncogene 2001, 20:8342-8357.

23. Steidl C, Leimeister C, Klamt B, Maier M, Nanda I, Dixon M, Clarke R, Schmid M, Gessler M: Characterization of the human and mouse HEY1, HEY2, and HEY3 genes: cloning, mapping, and mutation screening of a new bHLH gene family. Genomics 2000, 66:195-203.

24. Hu JC, Sauer RT: The basic-region leucine-zipper family of DNA binding proteins. Proc Natl Acad Sci USA 1997, 94:11266-11271.