Classification and evolutionary analysis of the basic helix-loop-helix gene family in the Japanese lamprey (*Lethenteron japonicum*)

Xiaoting Liu, Lei Liu *, Wei Liu, Juan Chen, Yanhua Sun,

Zi Li, Nannan Wang

School of Environment and Life Health, Anhui Vocational and Technical College,

Hefei 230011, Anhui, China

Abstract. Basic helix-loop-helix (bHLH) transcription factors are widely distributed in eukaryotic organisms ranging from yeast to mammals and are thought to be one of the largest families of regulatory proteins. They possess crucial functions in the control of a variety of developmental processes, such as cell proliferation and differentiation, cell lineage determination, myogenesis, neurogenesis, hematopoiesis, sex determination, gut development, as well as other essential processes. Lampreys are representatives of an ancient jawless vertebrates that diverged from our own ~500 million years ago and therefore are important for the study of vertebrate evolution. In this study, we conducted a genome-wide survey using the Japanese lamprey genomic database and identified 102 putative bHLH genes. Based on phylogenetic analysis, these Japanese lamprey genes were classified into 43 families, the identified LjbHLH genes were classified into 40 bHLH families with 42, 24, 18, 2, 13, and 0 member(s) in group A, B, C, D, E and F respectively, and 3 members categorized as “orphans”.

Keywords: Basic helix-loop-helix; vertebrate evolution; phylogenetic analysis; Japanese lamprey.

1. Introduction

The basic helix-loop-helix (bHLH) domain is a highly conserved peptide sequence that is an essential part of many transcription factors involved in a myriad of regulatory processes across Eukaryotic life, from neurogenesis in mammals[1-4], environmental response in plants [5] to metabolism in fungi [6]. This motif provides two of the crucial molecular roles for transcription factors, DNA binding and transcriptional regulation.

The tripartite bHLH domain is approximately 60 amino acids in length, with two highly conserved and functionally distinct segments, the basic region and the HLH region. The first 13 amino acid residues at the N-terminal of this motif comprise the basic domain, which generally contains five to six basic residues that facilitate DNA binding [4] . Many bHLH domains bind to the hexanucleotide sequence known as the E-box (CANNTG) or its degenerate forms in most bHLH proteins [4]. The HLH region at the C-terminal of bHLH motif facilitates dimerization and formation of homo- or heterodimeric complexes between different family members, consists of two amphipathic a-helices separated by a flexible loop structure [4].

Numerous bHLH proteins had been identified in animals, plants and fungi in succession since the first characterization of bHLH transcription factors was reported on the murine factors E12 and E47[4]. Phylogenetic analyses have classified the diversity of bHLH proteins into a number of distinct groups. Over 50 bHLH proteins are encoded in the genomes of most animals (metazoans) and are typically classified into six major groups (A–F), based on their ability to bind DNA and 45 bHLH families based on their different functions in the regulation of gene expression [3,7,8]. Group A consisted of 22 subfamilies that bind to CACCTG or CAGCTG core sequences of E boxes, and

their functions were the regulation of sex determination, trophoblast cell development and mesoderm formation. Group B contains 12 subfamilies and many ancient, highly conserved members such as Myc, Mad, Hairy, and Pho 4 that bind to CACGTG or CATGTTG core sequences of E boxes. They mainly controlled the expression of glucose-responsive genes and regulated the sterol metabolism. Plant and fungal bHLH proteins have been found to be most closely related to animal group B members, and are classified into 26–33 and 12 subgroups, respectively [6,9-14]. Group C had seven subfamilies, with one or two PAS domains following the bHLH motif, tend to bind the core sequence of ACGTG or GCGTG. They are responsible for the regulation of midline and tracheal development, circadian rhythms, and for the activation of gene transcription in response to environmental toxins. Group D had only one subfamily and act as an antagonist of group A bHLH proteins by forming inactive heterodimers which are incapable of binding target DNA. Group E contained two subfamilies and two characteristic domains in addition to the bHLH named “Orange” and “WRPW” peptide in the carboxyl terminus. They bind preferentially to sequences referred to as N boxes (CACGCG or CACGAG) and mainly regulate embryonic segmentation, somitogenesis and organogenesis. Group F consists of COE-bHLH proteins, which has an additional domain involved in both dimerization and DNA binding. The protein functions of this group were mainly regulation of head development and formation of olfactory sensory neurons [7,8,15].

Japanese lamprey (Lethenteron japonicum) are Northern hemisphere lampreys (subfamily Petromyzontidae) that diverged about 30–10 Mya [16]. Abstract Lampreys are eel-like jawless fishes evolutionarily positioned between invertebrates and vertebrates, and have been used as model organisms to explore vertebrate evolution.

2. Materials and Methods

2.1 Search of bHLH Sequences

For the purpose of obtaining candidate genomic sequences encoding bHLH motifs in the Japanese lamprey, Amino acids of the 45 representative bHLH motifs and the 114 mouse bHLH motifs obtained from the additional files of previous reports were used as queries to perform tBlastn searches against the integrated databases of the Japanese lamprey.

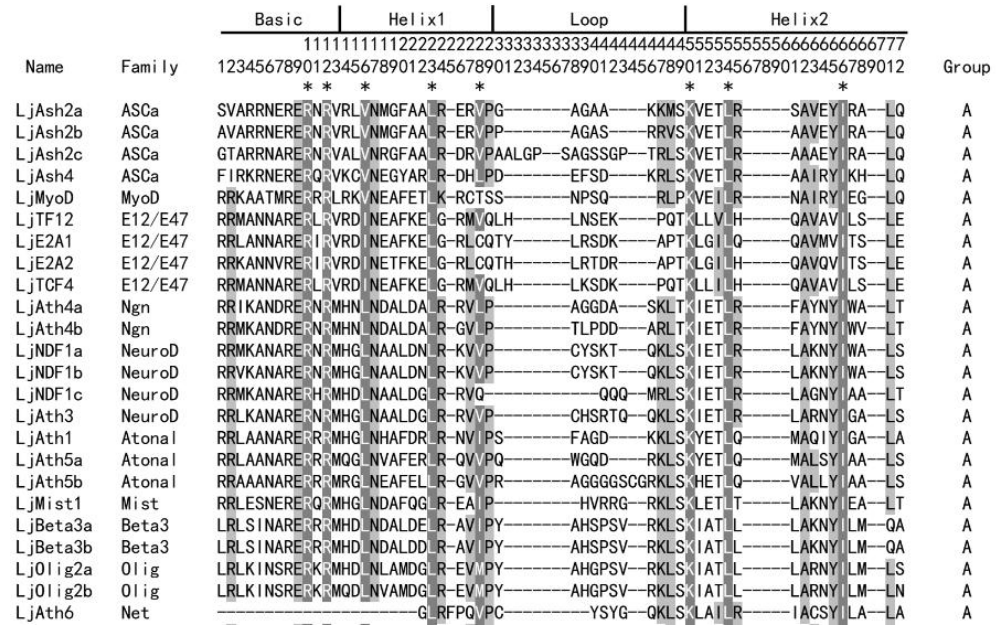
2.2 Phylogenetic analysis

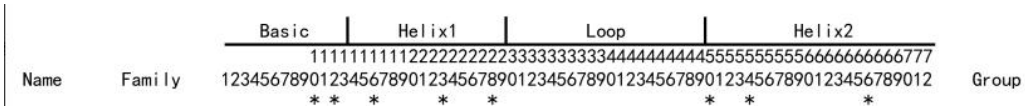
Evolutionary relationships among all identified bHLH motifs were examined using BioNJ, MP (maximum parsimony), and ML (maximum likelihood) methods. ML tree using PhyML program online with LG amino acid substitution model and other parameters optimized by ProtTest was constructed firstly with all the NvbHLHs and 59 DmbHLH motifs, so that we could know to which higher-order group a candidate bHLH sequence belonged. Then, in-group phylogenetic analysis was generated to identify homologs with ML method and two additional algorithms.

3. Result and discussion

3.1 Identifiation of bHLH proteins in the Japanese lamprey

Our search for bHLH-domain containing proteins identified 102 distinct bHLH transcription factors encoded in the Japanese lamprey is shown in Fig. 1 as described in Materials and methods. Most of the bHLH domains we obtained had more than 10 conserved amino acids among the 19 residues.





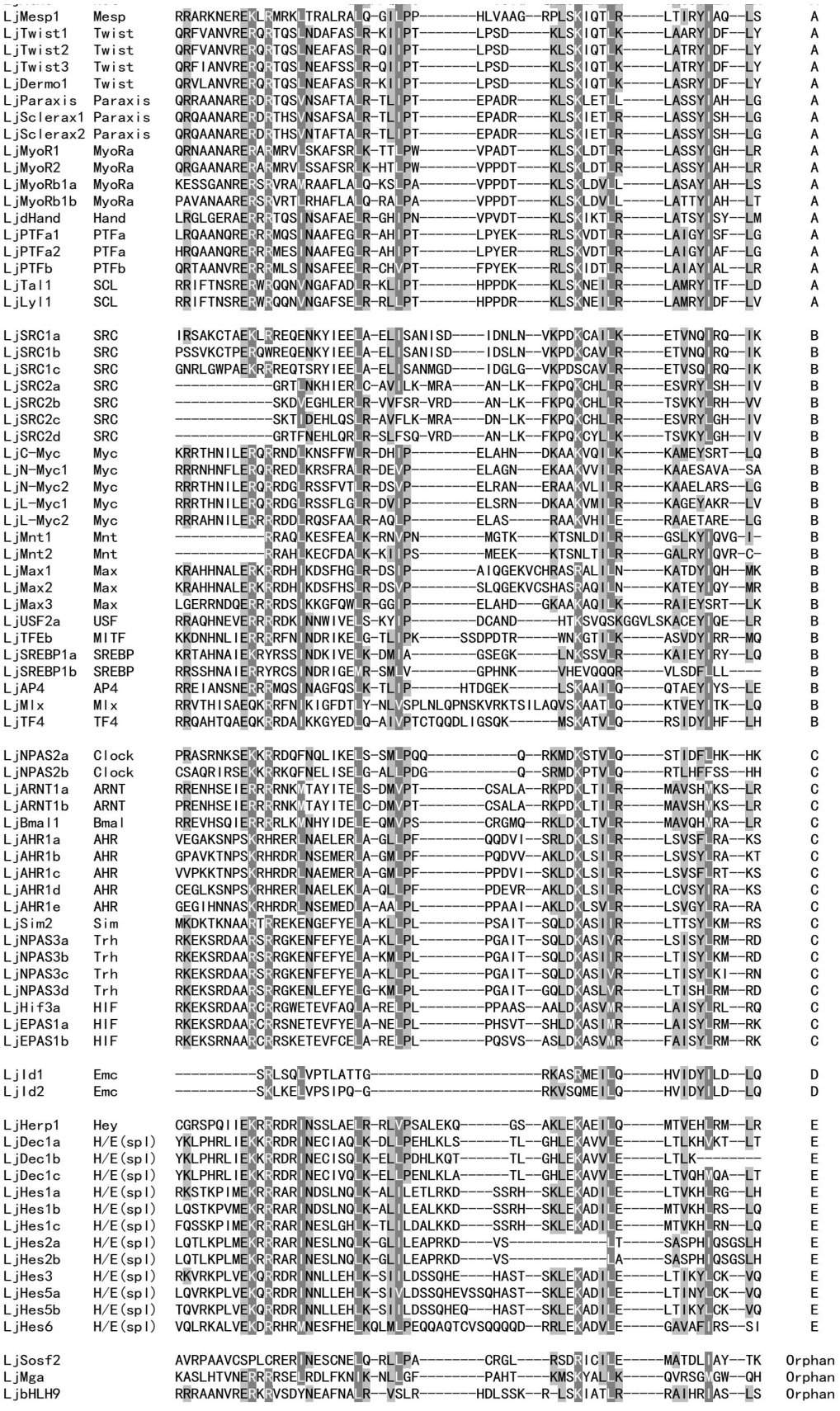


Fig. 1 Multiple alignment of the 102 Japanese lamprey basic helix-loop-helix motifs.

3.2 Orthologous Relationships with Mouse Proteins

Orthologous genes in two or more species are those that have evolved by vertical descent from a common ancestor [17]. Orthologue identification is conducive to further studies on structural and functional comparison with other organisms.

BLAST searching and in-group phylogenetic analysis revealed that 26, 22, 8, 1, 8,0, and 2 LjbHLH sequences could be assigned to their correspondent mouse bHLH homologs with sufficient bootstrap support (all NJ, MP, and ML bootstrap values ≥50 %) in groups A, B, C, D, E, F, and Orphan, respectively.

3.3 Identification of LjbHLH protein sequences

Protein sequences correspondent to 63 of the 102 identified LjbHLH motifs have been deposited in GenBank and those of the rest 40 are not available .

It has been reported that certain conserved domains or motifs are often present within related bHLH protein groups although amino acid sequences flanking the bHLH region are generally divergent, even in closely related proteins from the same species. In order to determine whether our classification to LjbHLH sequences is reliable, a separate phylogenetic tree was constructed based on an alignment of all LjbHLH motifs. Domains and motifs in AcbHLH protein sequences were then predicted using the online program SMART .

3.4 Intron/exon distribution within bHLH motif coding regions

Multiple introns interrupt the coding sequences in the great majority of genes in animals and plants, whereas intron densities in fungi and unicellular eukaryotes are highly variable. The coding regions, intron location and length of all 102 LjbHLH motifs are only 21 LjbHLH members with introns in their bHLH motifs.

4. Summary

The highly conserved bHLH proteins comprise a large superfamily of transcription factors. They are commonly distributed in large numbers within animal, plant, and fungal species. In this study, the Japanese lamprey (Lethenteron japonicum) genome was found to encode 102 bHLH genes. Phylogenetic analysis of 102 identified bHLH motifs permitted classification of these members into 43 families, with 42, 24, 18, 2, 13, and 0 member(s) in groups A, B, C, D, E, and F, respectively, with the remaining three members categorized as “orphans”.

Acknowledgment

This work was funded by the 2023 Key Quality Engineering Project of Anhui Vocational and Technical College (2023yjjyxm04), and the 2022 Anhui Provincial Quality Engineering Project (2022zygzsj035).

References

[1] Amoutzias GD, Robertson DL, Bornberg-Bauer E (2004) The evolution of protein interaction networks in regulatory proteins. Comp Funct Genomics 5: 79-84.

[2] Amoutzias GD, Robertson DL, Van de Peer Y, Oliver SG (2008) Choose your partners: dimerization in eukaryotic transcription factors. Trends Biochem Sci 33: 220-229.

[3] Jones S (2004) An overview of the basic helix-loop-helix proteins. Genome Biol 5: 226.

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

[5] Castillon A, Shen H, Huq E (2007) Phytochrome Interacting Factors: central players in phytochrome-mediated light signaling networks. Trends Plant Sci 12: 514-521.

[6] Sailsbery JK, Atchley WR, Dean RA (2012) Phylogenetic analysis and classification of the fungal bHLH domain. Mol Biol Evol 29: 1301-1318.

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

[8] Ledent V, Vervoort M (2001) The basic helix-loop-helix protein family: comparative genomics and phylogenetic analysis. Genome Res 11: 754-770.

[9] Atchley WR, Fernandes AD (2005) Sequence signatures and the probabilistic identification of proteins in the Myc-Max-Mad network. Proc Natl Acad Sci U S A 102: 6401-6406.

[10] Buck MJ, Atchley WR (2003) Phylogenetic analysis of plant basic helix-loop-helix proteins. J Mol Evol 56: 742-750.

[11] Carretero-Paulet L, Galstyan A, Roig-Villanova I, Martinez-Garcia JF, Bilbao-Castro JR, et al. (2010) Genome-wide classification and evolutionary analysis of the bHLH family of transcription factors in Arabidopsis, poplar, rice, moss, and algae. Plant Physiol 153: 1398-1412.

[12] Heim MA, Jakoby M, Werber M, Martin C, Weisshaar B, et al. (2003) The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. Mol Biol Evol 20: 735-747.

[13] Osborne TF, Espenshade PJ (2009) Evolutionary conservation and adaptation in the mechanism that regulates SREBP action: what a long, strange tRIP it's been. Genes Dev 23: 2578-2591.

[14] Pires N, Dolan L (2010) Origin and diversification of basic-helix-loop-helix proteins in plants. Mol Biol Evol 27: 862-874.

[15] Ledent V, Paquet O, Vervoort M (2002) Phylogenetic analysis of the human basic helix-loop-helix proteins. Genome Biol 3: RESEARCH0030.

[16] Kuraku S, Kuratani S (2006) Time scale for cyclostome evolution inferred with a phylogenetic diagnosis of hagfish and lamprey cDNA sequences. Zoolog Sci 23: 1053-1064.

[17] Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19: 99-113.