Molecular evolution of the MyoD family of transcription factors

WILLIAM R. ATCHLEY*, WALTER M. FITCH†, AND MARIANNE BRONNER-FRASER‡

*Center for Quantitative Genetics, Department of Genetics, North Carolina State University, Raleigh, NC 27695-7614; and Departments of †Ecology and Evolutionary Biology and ‡Cell and Developmental Biology, University of California, Irvine, CA 92717

Contributed by Walter M. Fitch, August 17, 1994

ABSTRACT—Myogenesis in skeletal muscle is a cascade of developmental events whose initiation involves the MyoD family of transcription factors. Evolutionary analyses of amino acid sequences of this family of transcriptional activators suggest that the vertebrate genes MyoD, myf-5, Myog (myogenin), and myf-6 were derived by gene duplications from a single ancestral gene. A common genetic origin predicts some functional redundancy between MyoD and myf-5 and between MyoD and myf-6. Experimental studies have suggested that these pairs of genes can substitute for each other during myogenesis. Separate analyses of the conserved basic helix-loop-helix and nonconserved flanking elements yield similar branching sequences but show evolutionary change in the basic helix-loop-helix region has occurred at a much slower rate.

Evolutionary changes in development and morphology occur, in part, as a result of heritable changes in patterns of gene expression. Because these changes are under transcriptional control, the evolution of transcriptional activators may be an important component in evolutionary change. We have examined the molecular evolution of the MyoD gene complex, a small family of transcription factors involved in myogenesis. Myogenesis is a developmental cascade whose initiation involves determination of multipotential mesodermal stem cells to a myogenic lineage and their subsequent differentiation into functional myocytes. This regulatory gene family includes MyoD (1, 2), Myog (or myogenin and myf-4) (3-6), myf-5 (7), and myf-6 (8-10).

The MyoD complex is part of the basic helix-loop-helix (bHLH) family of transcriptional regulators that control cell type-specific transcription, proliferation, and transformation (6). The bHLH family exhibits a highly conserved motif of ~60 amino acids concerned with protein dimerization and DNA binding. The basic region constitutes the DNA binding motif and the contiguous helix-loop-helix region is a dimerization motif permitting multimerization with other bHLH proteins. bHLH proteins have several common characteristics. They form heterodimers with ubiquitously distributed E proteins (which also belong to the bHLH family) and they recognize a specific consensus sequence in DNA known as the E box. In the myogenic lineage, the E box is present in most skeletal-muscle-specific genes and in the other bHLH transcription factors. In addition to the MyoD genes, the bHLH family contains several gene complexes involved in cell determination and differentiation. These include achaete-scute (neurogenesis), twist (mesoderm formation), daughterless (sex determination and peripheral nervous system formation), and myc (protooncogenes).

Reports on Caenorhabditis elegans (11), sea urchin (12), and Drosophila (13) indicate that only a single member of the MyoD gene family occurs in invertebrates. However, there are four separate skeletal muscle regulatory genes with overlapping functions in vertebrates. When transfected into fibroblasts, each of the four transcription factors in the vertebrate MyoD family can activate the complete myogenic program of gene expression, converting cultured fibroblasts to skeletal myocytes (6, 14, 15). Further, the genes autoregulate and cross-activate each other in vitro (6, 14-18).

In contrast, some long-term muscle cell lines only express subsets of the MyoD family (19) and do not cross-activate. During embryogenesis, each gene is apparently activated at a slightly different time (17). In primary muscle cultures, members of the MyoD family activate the complete myogenic program in a defined sequence producing multinucleate myotubes with fully assembled contractile apparatus in a manner indistinguishable from primary myocytes (20). In some cell lines, on the other hand, only subsets of muscle-specific genes may be transiently expressed (6).

METHODS AND MATERIALS

To gain a better understanding of the developmental roles and consequences of these genes in myogenesis, we examine the evolutionary relationships among 29 amino acid sequences of the MyoD gene family from 12 species. The sequences were aligned using the CLUSTAL multiple alignment computer program (21) followed by manual correction by eye. Genetic divergence among aligned sequences is expressed as pairwise genetic distances defined as the proportion of amino acids by which any two sequences differ (gapped positions excluded). A neighbor joining tree (22) was computed to summarize the evolutionary relationships among sequences and the results were analyzed using the bootstrap method to provide confidence levels for the tree topology (23). This was done for the complete sequence, the bHLH sequence alone, and the complete sequence minus the bHLH region.

RESULTS

The neighbor joining tree for the MyoD gene family is given in Fig. 1. Table 1 provides some representative genetic distances between selected sequences. There is little similarity between invertebrate and vertebrate species and it was not possible to align them except in the region of the bHLH motif. Over the entire sequence, the average proportional differences between invertebrate and vertebrate amino acid sequences among all genes (26 comparisons) are 56% for sea urchin, 63% for Drosophila, and 70% for C. elegans.

Fig. 1 appears statistically reliable because only 6 of the 26 cladest have bootstrap values less than 95%. A low bootstrap value may occur where the tree appears incorrect. For example, Myf5 in Xenopus diverges after, rather than before, the bird lineage, and this arrangement has a bootstrap value of only 83%, suggesting it is not statistically well supported. An important exception occurs in the MyoD1 lineage where mammals diverge first rather than last, yet the bootstrap value is high (>95%). This implies that another gene duplication has occurred; however, supporting evidence in the

Abbreviation: bHLH, basic helix-loop-helix.
form of the other duplicated gene of the pair has not yet been found.

Fig. 1 suggests that a single ancestral myogenesis transcription factor initially split into two lineages early in the evolution of vertebrates. MyoD1 and myf-5 evolved from one of these vertebrate lineages while the second lineage produced Myog and myf-6. Thus, MyoD1 and myf-5 arose from a common gene as did Myog and myf-6. One possible outcome of such a pattern of evolution by gene duplication in the MyoD family is that MyoD1, myf-5, Myog, and myf-6 may have preserved some redundancy in function. Instances of proven and inferred paralogy (homology by gene duplication) are indicated in Fig. 1. Paralogy is proven each time two genes from the same taxon trace back to a different node; it is inferred whenever a statistically well-supported clade (a monophyletic subtree) implies a relationship grossly at odds with well-substantiated biological opinion.

In humans, MyoD1 and Myog are located on human chromosomes 11 and 1, respectively, while myf-5 and myf-6 are located on chromosome 12 with their translational start codons only 8.5 kb apart (4, 5, 18). In spite of their close physical proximity, Fig. 1 indicates that myf-5 and myf-6 are not each other’s closest relatives and, thus, probably did not arise simply by tandem duplication. If they did arise by tandem duplication, then at least two of the genes have migrated away from the tandem duplication site to other chromosomes and the flanking sequences have accumulated sufficient substitutions to destroy any evidence of such duplications. Thus, three duplication events probably occurred that successively increased the number of paralogous genes from one to two to three to four MyoD family genes. However, there is little evidence favoring a scenario of two rather than three tandem duplications with gene numbers going from one to two to four MyoD family genes. (i) Fig. 1 shows that first split (duplication) divides the four lineages two and two not one and three. (ii) The distances from the node showing the first duplication (having an 87 to its right in Fig. 1) to the other two duplications are nearly equal (18- and 20-amino acid replacements), indicating that the duplications could have occurred simultaneously, producing four genes from two. The numbers 18 and 20 come from the multiplication of 45 and 52 replacements per 1000 sites on Fig. 1 by
Table 1. Some representative distances between myogenesis transcription factors

<table>
<thead>
<tr>
<th></th>
<th>XenoMyoD1</th>
<th>ChickMyoD</th>
<th>MouseMyoD</th>
<th>MouseMyf5</th>
<th>ChickMyf5</th>
<th>MouseMyf5</th>
<th>ChickMyf6</th>
<th>MouseMyf6</th>
<th>ChickMyog</th>
<th>MouseMyog</th>
<th>SeaUMyoD</th>
<th>FlyMyoD</th>
<th>CelMyoD</th>
</tr>
</thead>
<tbody>
<tr>
<td>XenoMyoD1</td>
<td>0.48</td>
<td>0.48</td>
<td>0.55</td>
<td>0.58</td>
<td>0.60</td>
<td>0.60</td>
<td>0.64</td>
<td>0.69</td>
<td>0.30</td>
<td>0.25</td>
<td>0.16</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>ChickMyoD</td>
<td>0.21</td>
<td>0.47</td>
<td>0.55</td>
<td>0.58</td>
<td>0.60</td>
<td>0.55</td>
<td>0.58</td>
<td>0.60</td>
<td>0.56</td>
<td>0.44</td>
<td>0.24</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>MouseMyoD</td>
<td>0.30</td>
<td>0.47</td>
<td>0.55</td>
<td>0.58</td>
<td>0.60</td>
<td>0.55</td>
<td>0.55</td>
<td>0.58</td>
<td>0.56</td>
<td>0.44</td>
<td>0.24</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>MouseMyf5</td>
<td>0.48</td>
<td>0.47</td>
<td>0.55</td>
<td>0.58</td>
<td>0.60</td>
<td>0.55</td>
<td>0.55</td>
<td>0.58</td>
<td>0.56</td>
<td>0.44</td>
<td>0.24</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>ChickMyf5</td>
<td>0.48</td>
<td>0.47</td>
<td>0.55</td>
<td>0.58</td>
<td>0.60</td>
<td>0.55</td>
<td>0.55</td>
<td>0.58</td>
<td>0.56</td>
<td>0.44</td>
<td>0.24</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>MouseMyf5</td>
<td>0.45</td>
<td>0.44</td>
<td>0.55</td>
<td>0.58</td>
<td>0.60</td>
<td>0.55</td>
<td>0.55</td>
<td>0.58</td>
<td>0.56</td>
<td>0.44</td>
<td>0.24</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>ChickMyf6</td>
<td>0.55</td>
<td>0.56</td>
<td>0.58</td>
<td>0.60</td>
<td>0.55</td>
<td>0.55</td>
<td>0.55</td>
<td>0.58</td>
<td>0.56</td>
<td>0.44</td>
<td>0.24</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>MouseMyf6</td>
<td>0.58</td>
<td>0.59</td>
<td>0.62</td>
<td>0.57</td>
<td>0.55</td>
<td>0.53</td>
<td>0.57</td>
<td>0.60</td>
<td>0.56</td>
<td>0.44</td>
<td>0.24</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>ChickMyog</td>
<td>0.60</td>
<td>0.61</td>
<td>0.61</td>
<td>0.58</td>
<td>0.58</td>
<td>0.50</td>
<td>0.54</td>
<td>0.57</td>
<td>0.56</td>
<td>0.44</td>
<td>0.24</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>MouseMyog</td>
<td>0.60</td>
<td>0.60</td>
<td>0.59</td>
<td>0.55</td>
<td>0.56</td>
<td>0.50</td>
<td>0.54</td>
<td>0.57</td>
<td>0.56</td>
<td>0.44</td>
<td>0.24</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>SeaUMyoD</td>
<td>0.55</td>
<td>0.55</td>
<td>0.56</td>
<td>0.53</td>
<td>0.54</td>
<td>0.53</td>
<td>0.56</td>
<td>0.57</td>
<td>0.55</td>
<td>0.44</td>
<td>0.24</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>FlyMyoD</td>
<td>0.64</td>
<td>0.64</td>
<td>0.67</td>
<td>0.60</td>
<td>0.57</td>
<td>0.63</td>
<td>0.62</td>
<td>0.66</td>
<td>0.65</td>
<td>0.44</td>
<td>0.24</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>CelMyoD</td>
<td>0.69</td>
<td>0.69</td>
<td>0.71</td>
<td>0.72</td>
<td>0.68</td>
<td>0.71</td>
<td>0.70</td>
<td>0.68</td>
<td>0.70</td>
<td>0.66</td>
<td>0.71</td>
<td>—</td>
<td></td>
</tr>
</tbody>
</table>

Table elements are the proportion of amino acids that differ between pairs of sequences. The elements below the diagonal are the differences between the full sequences (392 amino acids). Standard errors for the full sequence distances are ~0.03 for all pairwise elements. Values above the diagonal are the proportional differences between amino acid sequences for the bHLH motif only (59 amino acids). Standard errors for these latter values are ~0.06. SeaU, sea urchin; fly, Drosophila melanogaster; Cel, Caenorhabditis elegans; Xeno, Xenopus laevis.

DISCUSSION

In vertebrates, skeletal muscle cells arise from mesodermal structure called somites, which are balls of epithelial cells with the potential to form dermis, cartilage, and muscle. The portion of the somite closest to the epidermis differentiates into the dermomyotome, which subsequently divides into the dermatoine (presumptive dermis) and myotome (presumptive muscle). Cells of the medial dermomyotome form axial and back muscles. Cells of the lateral dermomyotome form the intercostal, ventral abdominal, and limb musculature.

The myogenic genes have differential patterns of expression in the developing somites. In mouse, myf-5 mRNA is expressed at the earliest time in the epithelial somite, followed by Myog, myf-6, and later by MyoD RNAs, which are expressed consecutively in the myotome (24-27). Surprisingly, the pattern of expression of orthologous genes in avian embryos is somewhat different from the mouse. For example, MyoD1 (qmf1) is the first myogenic gene to be expressed and is present in the epithelial somite. This is followed by Myog (qmf2) and myf-5 (qmf3) expression, which are first observed in the dorsomedial portion of the somite (28) and later throughout the myotome. However, these differences in expression pattern between mouse and bird are less significant if one considers MyoD1 and myf-5 as a functionally equivalent pair of genes with a common evolutionary origin.

The trees provide an explanation for phenotypes observed after functional inactivation of myogenic genes. For example, MyoD1 and myf-5 arose from a more recent common gene, which suggests that they are more closely related and, therefore, might more readily substitute for one another in muscle development. Experimental support for the idea that MyoD1 and myf-5 may be functionally similar with respect to muscle formation comes from mice carrying null mutations in either gene. These mice have apparently normal skeletal muscle, suggesting that either MyoD1 or myf-5 alone is sufficient to produce muscle differentiation (29, 30). This is consistent with the possibility that the two genes have overlapping functions in myogenesis and may substitute for each other. Alternatively, they both could be required for differentiation of different subsets of precursors, but the loss of one may be compensated by proliferation of the other. When MyoD1 is absent, there is a 3.5-fold increase in the amount of myf-5 mRNA (31), suggesting some up-regulation of myf-5 in the absence of MyoD1. However, the myf-5 knock-out mice die of severe rib deficiencies and malformations, whereas the MyoD1 knock-out mice are viable. Thus, the two genes cannot be completely functionally interchangeable. One would predict that loss-of-function of both MyoD1 and myf-5 would produce a severe phenotypic effect. Indeed, mice carrying null mutations at both MyoD1 and myf-5 are completely devoid of any skeletal muscle (31).

The evolutionary analyses also suggest overlapping functions for Myog and myf-5. Mice lacking Myog have many muscle precursors that are blocked from fusing and forming muscle fibers (32, 33); however, when transfected with high levels of MyoD1, fibroblast cells from myogenic mutant mice can differentiate into normal muscle, suggesting a transcription factor other than myogenin can induce muscle differentiation (33). Thus, MyoD1/myf-5 may be expressed earlier in the myogenesis cascade than Myog/myf-6. The temporally
distinct expression patterns of myogenin and MyoD in muscle precursor cells may partially explain the more severe phenotypic effects of knocking-out Myog alone compared with MyoD1 alone. During normal development, MyoD1 may be turned off before Myog/myf-6 is required. Thus, in the Myog knock-out mice, there may be insufficient spatiotemporal overlap for the endogenous levels of MyoD1 to rescue muscle differentiation when later expressed genes like Myog are functionally perturbed. An important test of the redundancy of Myog and myf-6 would be to transfect the mutant cells with myf-6 to see whether this can rescue the myogenic differentiation program.

bHLH proteins may be important in determining cell fate and/or differentiation in a number of developmental systems; for example, a number of bHLH DNA binding proteins, including achaete, scute, asense, and lethal of scute, are involved in differentiation of the neuronal lineage in Drosophila. There are some interesting similarities between bHLH genes involved in vertebrate myogenesis and Drosophila neurogenesis (34). Ectopic expression of any of the four genes of the achaete–scute complex results in ectopic neurogenesis. Furthermore, these genes exhibit distinct but overlapping patterns with other related genes. In addition, members of the achaete–scute complex utilize ubiquitous cofactors as both positive and negative regulators, as do the myogenic bHLH genes. Furthermore, there appears to be cross-regulation between different bHLH genes in both fly neurogenesis and vertebrate myogenesis. Thus, the bHLH motif may be a reiterated theme in development, perhaps arising from a common ancestral gene component.

A model of vertebrate myogenesis proposed by Jan and Jan (34) suggests that the cascade of regulation by bHLH proteins in myogenesis may be quite similar to that occurring in Drosophila neurogenesis. In the case of neurogenesis, there is a hierarchy of bHLH function such that achaete and scute, which are very similar in sequence and function, regulate another homologous pair, asense and deadpan. For myogenesis, both MyoD1 and myf-5 show overlapping functions by inducing muscle precursor cells (undifferentiated mesodermal cells) to undergo differentiation. Their target genes, Myog and myf-6, are expressed later in myogenesis and may function to maintain this differentiated state. This model is consistent with our phylogenetic results which shows that MyoD1, myf-5, Myog, and myf-6 evolved from common gene lineages. Thus, no single member of the bHLH family functions as a master regulator in either vertebrate myogenesis or Drosophila neurogenesis. Rather, a group of proteins have evolved that may work in concert to provide genetic control of a developmental cascade.

We thank Drs. Charles Ordahl, Brian Williams, Susan Bryant, Eric Olson, and Michael Miyamoto for helpful comments on the manuscript and Helene Van for technical assistance. One of us (W.R.A.) is supported by the Alfred P. Sloan Foundation and by grants from...
the National Institutes of Health (GM-45344) and the National Science Foundation (BSR-910718). W.M.F. is supported by a grant from the National Science Foundation (DEB9096152). M.B.-F. was supported by Muscular Dystrophy Association and by grants from the National Institutes of Health (HD-15527, HD-25138, and DE-10066).