*Proc. Natl. Acad. Sci. USA*
Vol. 90, pp. 659–663, January 1993
Evolution

# Molecular evolution of the myosin family: Relationships derived from comparisons of amino acid sequences

HOLLY V. GOODSON AND JAMES A. SPUDICH

Departments of Biochemistry and Developmental Biology, Beckman Center for Molecular and Genetic Medicine, Stanford University Medical School, Stanford, CA 94305

*Contributed by James A. Spudich, October 2, 1992*

ABSTRACT **To examine the evolutionary relationships between members of the myosin family, we have used two different phylogenetic methods, distance matrix and maximum parsimony, to analyze all available myosin head sequences. We find that there are at least three equally divergent classes of myosin, demonstrating that the current classification of myosin into only two classes needs to be reexamined. In the myosin II class, smooth muscle myosin is more closely related to nonmuscle myosin than to striated muscle myosin, implying that smooth muscle and skeletal muscle myosins were independently derived from nonmuscle myosin and suggesting that similarities between these types of muscle are the result of convergent evolution. The grouping of head sequences produced by phylogenetic analysis is consistent with classifications based on enzymology and structural localization and is generally consistent with grouping based on common tail structure elements. This result demonstrates that specific head sequences are tightly coupled to specific tail sequences throughout evolution and challenges the idea that myosin heads are freely interchangeable units whose unique function is determined only by the tail structure to which it is attached.**

The myosins are a family of mechanochemical proteins whose members participate in activities as diverse as cytokinesis, muscle contraction, and organelle motility. For many years, the term "myosin" referred to the two-headed, filament-forming protein with a coiled-coil tail, now known as myosin II (for review see refs. 1 and 2). More recently, a number of proteins were identified which contain a region homologous to the myosin II head but lack the coiled-coil tail. These proteins are collectively called myosin I, or unconventional myosin (for review see refs. 3–5). Study of myosins I has generated a great deal of interest because it is thought that these proteins play important roles in both cell motility and organelle transport. Initially it seemed possible that myosins I were evolutionary novelties, specific to the organism in which they were found. However, members of at least two myosin I subclasses have been found in organisms as divergent as yeast and mammalian cells (reviewed in refs. 4 and 5), strongly implying that these types of myosin I, like myosin II, are universally present in eukaryotic cells.

As additional myosin I proteins have been identified, it has become clear, both through biochemistry and sequence analysis, that the subgroups of myosin I are actually much more divergent than the subgroups of myosin II. Are myosins I more related to each other than to myosin II, or are there several completely independent types of myosin? Do subgroupings derived from evolutionary analysis agree with those derived from biochemical characterization? Do relationships derived from phylogenetic study of myosin II head domains agree with those previously done on tail sequences? Can evolutionary information derived from sequence data allow prediction of biochemical characteristics of otherwise uncharacterized proteins? With these questions in mind we undertook the following phylogenetic analysis of myosin head sequences.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

## MATERIALS AND METHODS

We utilized two phylogenetic methods in the analysis presented here: distance matrix, as implemented by the ensemble of programs for progressive alignment and phylogenetic tree construction by Feng and Doolittle (6), and maximum parsimony, as performed by the PROTPARS program of the PHYLIP package of Felsenstein (7). Protein sequences were used in order to avoid artifacts caused by codon bias and to include data derived from protein sequencing. All sequences were truncated to begin and end at conserved amino acid positions to avoid biasing scores by the presence of unrelated sequences; these amino acids correspond to positions 15 and 812 in the alignment given by Pollard *et al.* (4).

The Feng and Doolittle programs first align all sequences progressively and then calculate a distance score for each pair of aligned sequences, thus creating a distance matrix. Trees are constructed by connecting the most related sequences stepwise [by the method of Fitch and Margoliash (8)] and adjusting the branch lengths so that they are as consistent as possible with the distances in the matrix [by the least-squares approach of Klotz and Blanken (9)]. Distance matrix methods can artifactually shorten longer branches when no correction is made for multiple substitutions (10). However, as we have attempted to measure only topology and not time of divergence, this is not expected to affect our results. Sequences were entered in alphabetical order by organism; altering input order did not change branching order or significantly change branch length (data not shown).

PROTPARS, like other maximum parsimony methods, assumes that the simplest path of evolution is the one followed and thus attempts to find the branching topology requiring the fewest possible mutations to get from a single unknown ancestral sequence to the present array. The multiple alignments needed as input into PROTPARS were generated by the CLUSTAL V package (11). Confidence intervals were estimated by the technique of bootstrap resampling (12): the SEQBOOT program of the PHYLIP package was used to create multiple ($n$ = 50) randomly sampled data sets from the original alignment. PROTPARS was then run on these data sets, and the resulting phylogenetic trees were then compared by the CONSENSE program of the PHYLIP package to give a measure of the robustness of the data producing the various nodes. Sequences were entered into the alignment programs in alphabetical order by organism, but PROTPARS trials were conducted with randomized order of sequence entry. Bootstrapping trials were also conducted with randomized order of entry, thus testing for both entry-order artifacts and robustness of the data.

System constraints limited PROTPARS analysis to 30 sequences at a time. The tree shown in Fig. 2 is actually a

composite of two trees, one which included all sequences except the striated muscle myosins RrEFSk, CkEFSk, RrC$\alpha$, RrC$\beta$, HuC$\beta$, CeHCB, CeHCC, CeHCD, and OvBW (see legend to Fig. 1 for abbreviations), and another which included all myosins II and used ScMY4 as an outgroup. The relationship between striated and other myosins was identical in both trees. The final tree was constructed by replacing the incomplete striated muscle myosin branch from the first tree with the complete striated muscle myosin branch from the second tree at the node marked "R" in Fig. 2.

## RESULTS AND DISCUSSION

The evolutionary relationships between amino acid sequences of head domains of known myosin genes were analyzed by two phylogenetic methods, distance matrix and maximum parsimony, which are based on different principles and operate under somewhat different assumptions (10). Figs. 1 and 2 show the trees of the myosin family derived from these analyses. Nodes (branch points) found on these trees may represent either divergence of proteins (gene duplication events) or divergence of species. Interpretations based on these data (or any other phylogenetic tree) must keep this ambiguity in mind. The labels U, A, B, and C mark analogous nodes on the two trees, as discussed below. A first conclusion to be drawn from these trees is that the myosin family can be separated into at least three equally unrelated classes: the myosins II, the *dilute* class of myosin I (composed of mouse dilute, chicken brain p190, and yeast MYO2 and MYO4), and the classic myosins I (comprised of the intestinal brush-border myosins, brain brush-border-like myosin, the amoeboid myosins I, and *S. cerevisiae* MYO3). This result is strongly supported by both methods and contrasts sharply with the traditional view of the myosin family, in which the greatest division is between myosin I and myosin II. Neither method can place *Drosophila* ninaC, *Drosophila* 95F unconventional myosin, or *Acanthamoeba* high molecular weight myosin I into the framework of the three main classes (see below), implying that there could be as many as six independent classes of myosin in the sequences represented here.

The grouping produced by the analysis of head-domain sequence agrees with that expected from similarities in tail sequence (compare Figs. 1 and 2 with Fig. 3). Both methods clearly group yeast MYO3 with *Acanthamoeba* IB and *Dic-*

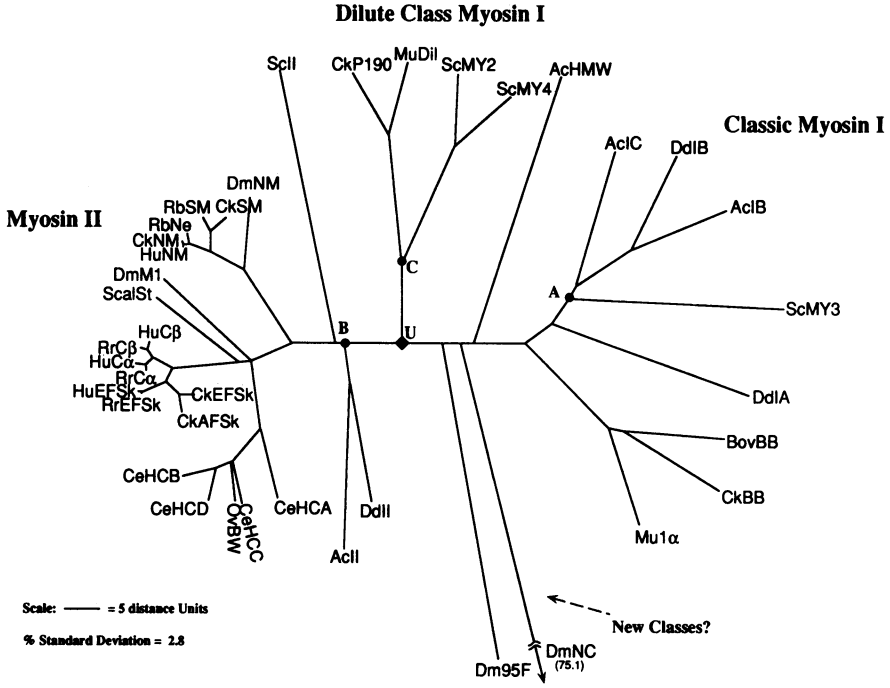


FIG. 1. Phylogenetic trees obtained from distance matrix analysis of myosin head protein sequences. Branch lengths are drawn to scale indicated in units of distance as calculated in the distance matrix. The percent standard deviation (lower left corner) gives an estimate of the error in branch length and position. This tree is drawn unrooted (without definition of the position of the "trunk"), as is proper since one cannot know which sequence is closer to the ancestral myosin gene. The significance of nodes U, A, B, and C is explained in the text. Published protein sequences were obtained from the Protein Identification Resource data bank (PIR), version 32, or the University of Geneva protein sequence data bank (Swiss-Prot), version 22, unless otherwise noted. Accession numbers and abbreviations are as follows: *Acanthamoeba* myosin II (AcII), Swiss-Prot P05659; *Acanthamoeba* high molecular weight myosin I (AcHMW), PIR A23622; *Acanthamoeba* myosin IB (AcIB), Swiss-Prot P19706; *Acanthamoeba* myosin IC (AcIC), Swiss-Prot P10569; bovine brush-border myosin I (BovBB), Swiss-Prot P10568; *Caenorhabditis elegans* myosin heavy chain A (CeHCA), Swiss-Prot P12844; *C. elegans* myosin heavy chain B (CeHCB), Swiss-Prot P02566; *C. elegans* myosin heavy chain C (CeHCC), Swiss-Prot P12845; *C. elegans* myosin heavy chain D (CeHCD), Swiss-Prot P02567; chicken brain P190 (CkP190), PIR S19188; chicken embryonic fast skeletal muscle myosin (CkEFSk), Swiss-Prot P02565; chicken adult skeletal muscle myosin (CkAFSk), Swiss-Prot P13538; chicken gizzard smooth muscle myosin (CkSm), Swiss-Prot P10587; chicken nonmuscle myosin (CkNM), Swiss-Prot P14105; chicken brush-border myosin I (CkBB), PIR A33620; *Dictyostelium* myosin II (DdII), PIR A26655; *Dictyostelium* myosin IB (DdIB), PIR A33284; *Dictyostelium* myosin IA (DdIA), Swiss-Prot P22467; *Drosophila* ninaC (DmNC), PIR A29813; *Drosophila* muscle myosin (DmM1) was spliced by hand from sequence PIR A32491 and splice junction information for cDNA cD301 (13); *Drosophila* nonmuscle (DmNM), PIR A36014; *Drosophila* 95F unconventional myosin (Dm95F), EMBL X67077; human cardiac $\alpha$ isoform (HuC$\alpha$) was entered from ref. 14; human cardiac $\beta$ isoform (HuC$\beta$), Swiss-Prot P12883; human nonmuscle myosin (HuNM), PIR M81105; human embryonic fast skeletal (HuEFSk), Swiss-Prot P11055; mouse brain brush-border-like myosin I (Mu1$\alpha$) sequence was kindly provided by Elliott Sherr; mouse dilute locus (MuDil), PIR S13652; *Onchocerca volvulus* body wall myosin (OvBW), PIR M74066; rabbit smooth muscle (RbSM), PIR M77812; rabbit neuronal myosin (RbNe), EMBL X62659; rat cardiac $\alpha$ isoform (RrC$\alpha$), PIR S06005; rat cardiac $\beta$ isoform (RrC$\beta$), Swiss-Prot, P02564; rat embryonic fast skeletal (RrEFSk), Swiss-Prot P12847; scallop striated (ScalSt), PIR S13557; *Saccharomyces cerevisiae* MYO1 (ScII), PIR S12323; *S. cerevisiae* MYO2 (ScMY2), PIR A38454; *S. cerevisiae* MYO3 (ScMY3), unpublished results; *S. cerevisiae* MYO4 (ScMY4), GenBank M90057.

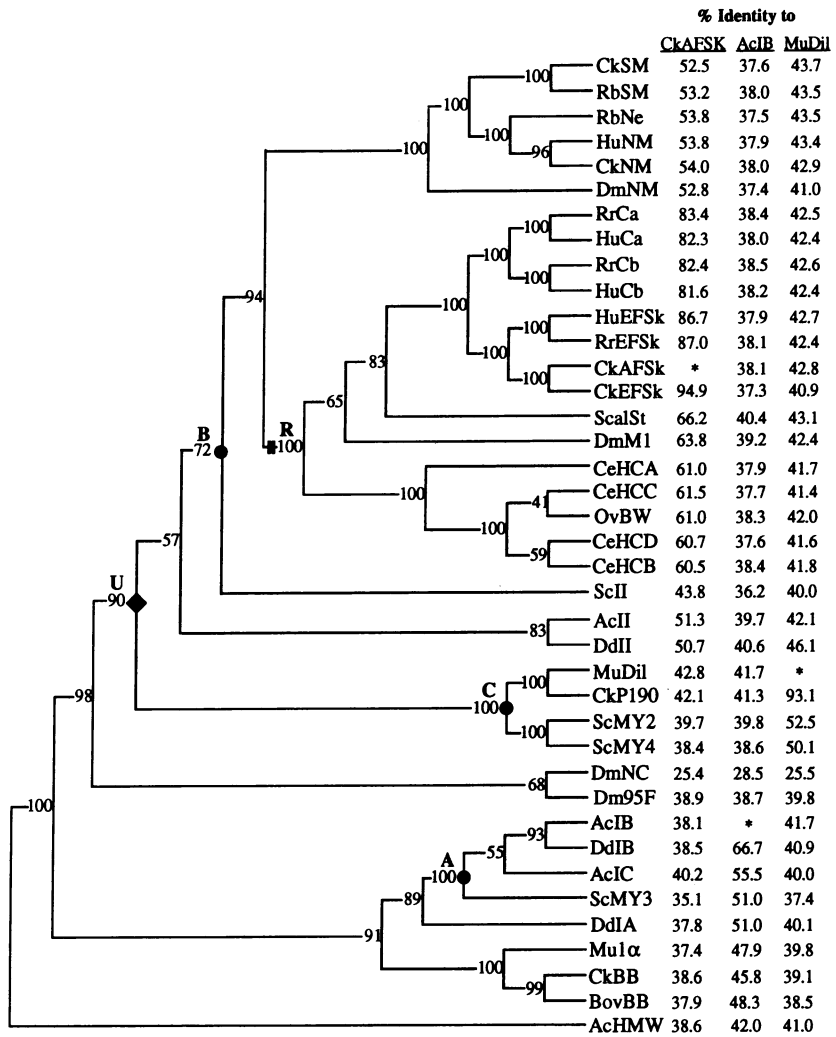


FIG. 2. Phylogenetic tree obtained from maximum parsimony analysis of myosin head protein sequences. This tree has been arbitrarily rooted at AcHMW for the purposes of more easily viewing topology. Branch lengths have no meaning. Numbers beside nodes are the percentage of bootstrapping trials in which an identical node was produced (see *Materials and Methods*) and thus are a measure of the robustness of the data generating that particular node. The significance of nodes U, A, B, C, and R is explained in the text. As an additional, albeit less sensitive, means of comparison, numbers to the right of the sequence names give the percent identity of each sequence to representative members of the three major classes. References and abbreviations are the same as in Fig. 1.

*tyostelium* IB, which have previously been subclassed as amoeboid myosins. These proteins have been postulated to play a role in cell motility (15). While the discovery of an amoeboid myosin in yeast does not rule out this possibility, the degree of conservation between these proteins suggests that they operate in similar environments in these different organisms and challenges us to describe this environment. *Acanthamoeba* IC, which has a slightly different tail structure (see Fig. 3), is also grouped with these proteins. The low bootstrap percentage in this part of the maximum parsimony tree results from the inability of the method to determine whether the gene duplication leading to *Acanthamoeba* IC occurred before or after the divergence of lines leading to yeast and *Acanthamoeba*. Distance matrix analysis does not clarify the situation: the long branch lengths between these sequences mean that the relative position of this branch point is poorly determined on this tree as well. It is interesting that the brush-border-type myosins diverge from the classic myosin I branch long before the separation of the yeast and amoeboid protozoa, implying that this subclass of classic myosin I was present in primordial eukaryotes and may well be ubiquitous. *Dictyostelium* myosin IA diverges nearby. This proximity, coupled with similarity in tail structure between these proteins and the errors inherent in estimating branch points between long branches, suggests that these proteins may have a common ancestor.

Figs. 1 and 2 show that smooth muscle myosins are more closely related to nonmuscle myosin than to striated muscle myosin, in agreement with previous analysis (22). This relationship implies that striated muscle is the more ancient form of muscle. In fact, *Drosophila* nonmuscle myosin diverges before the separation of vertebrate smooth and nonmuscle myosins, suggesting that smooth muscle postdates the divergence of fly and vertebrate lines. In contrast, the worm, fly, and scallop striated proteins are clearly derived from the same ancestors as the vertebrate striated myosins, which indicates that striated muscle predates the divergence of these organisms. The close relationship between smooth and nonmuscle myosins suggests that smooth muscle and striated muscle tissues evolved independently from nonmuscle tissues and that any similarities between these types of muscle are the result of convergent evolution.

Previous phylogenetic analyses of myosin sequences have concentrated on the relationships between vertebrate striated muscle myosin tail domains, taking particular interest in determining the relationships between developmentally specific isoforms (16–19). Results obtained here are generally consistent with the results of these studies. One exception to this agreement is that phylogenetic analysis of cardiac myosin tails implies that human $\alpha$ cardiac myosin is more related to human $\beta$ cardiac myosin than it is to rat $\alpha$ cardiac myosin (19). Our results disagree, indicating that the divergence of $\alpha$ and $\beta$ cardiac myosins predates the divergence of humans and rats. Matsuoka *et al.* (14) have noticed that the human $\alpha$ cardiac tail does have a region containing sequence normally specific to $\beta$ isoforms. This sequence similarity may have been a response to pressure for the human $\alpha$ protein, normally the "fast" isoform, to acquire properties of the slower, $\beta$ protein as the primate heart grew larger through evolution (14). Such an interpretation is hard to reconcile with a simple

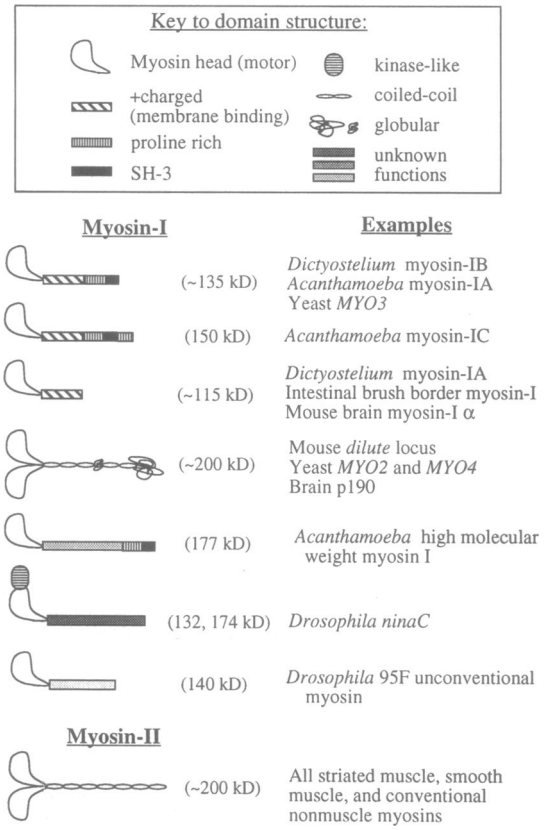


FIG. 3. Myosin motor superfamily: Classification by tail domains (adapted from refs. 4 and 5). Lengths are not to scale. kD, kilodaltons.

model of myosin function in which all determinants of speed and force reside in the head. It will be interesting to examine cardiac proteins from other large animals to see whether similar conversions have occurred.

As mentioned above, neither distance matrix nor maximum parsimony is able to unambiguously place *Acanthamoeba* high molecular weight myosin I, *Drosophila* ninaC, or *Drosophila* 95F unconventional myosin into the context of the three main classes, possibly implying the existence of additional myosin classes. However, maximum parsimony bootstrap analysis does usually group together ninaC and *Drosophila* 95F unconventional myosin, depending on the particular data-set sample being analyzed. Distance matrix clearly gives different results. It is possible that this grouping is an artifact of the maximum parsimony method, as theoretical simulation has shown that long branches tend to attract each other in this type of analysis (10). However, one should not dismiss this result without considering the possibility that it reflects biological reality. If constraints upon an amino acid sequence were altered by changes in the environment in which a protein functions, one might expect that the sequence with altered constraints would accumulate mutations in otherwise conserved areas much faster than sequences with constant constraints. A sequence like this might appear to be very divergent by distance matrix calculations. However, if a sufficiently small amount of time had passed since the divergence of this protein from its ancestors, the shadow of its ancestral amino acids would still be present in its nucleotide sequence and would thus be accessible to maximum parsimony analysis. The large number of nonconservative mutations which ninaC has in otherwise conserved regions of the myosin head does indeed suggest that ninaC is under constraints different from those of other myosins (see alignment in ref. 4).

## GENERAL CONCLUSIONS

The primary conclusion to draw from this work is that there are at least three classes of myosin likely to be present in all eukaryotes. The two classes previously grouped as myosin I are as divergent from each other as each is from myosin II. It is possible that three additional classes of myosin are represented in this data by *Acanthamoeba* high molecular weight myosin I, *Drosophila* ninaC, and *Drosophila* 95F unconventional myosin. A seventh class may be represented by another *Drosophila* myosin recently identified by D. Kiehart (personal communication). Though clearly sufficiently divergent in both head and tail domains to be categorized by themselves, we have hesitated to call these proteins separate classes, because homologous proteins from other organisms have yet to be found. The observation that half of the possible classes are represented by only one sequence suggests that many more types of myosin are yet to be discovered.

We also find that the evolution of tail sequences appears to be tightly correlated with the evolution of head sequences. As can be seen in comparing Figs. 1–3, myosins with similar tails have head sequences which are more closely related to each other than to myosins with unrelated tails. This apparent tight coupling between evolution of myosin heads and tails challenges traditional ideas about the relative functions of these domains. Are myosin heads interchangeable units, or is the unique functionality of a head coupled to that of a tail? While obviously myosin heads have been attached to new tails across evolutionary time spans, results presented here imply that such swapping events are very rare and that heads and tails, once connected, do not interchange. The existence of $\beta$ cardiac sequence characteristics in the tail of the human cardiac $\alpha$ isoform seems to contradict this conclusion but instead may support it by demonstrating that such interchanges can occur but are functional enough to be retained in only very specific cases. Perhaps myosin heads need certain characteristics to function correctly when attached to certain types of tails. This idea may make sense when one considers the different biophysical constraints upon the myosin motor in different roles—e.g., driving the contraction of a myofibril vs. translocating an organelle along an actin filament.

The results of the analysis presented here illustrate several dangers present when interpreting results from phylogenetic analysis of sequence families. First, one sees that different methods can give different results, especially when looking at the relationships between long branches. Second, we see how important it is to remember that any tree of this type intermixes divergence due to gene duplication with divergence due to speciation. This consideration is especially important when one is constructing species phylogenies based on sequence derived from cDNA libraries or PCR. One could get a very skewed view of metazoan evolution by using data from mixed striated, smooth, and nonmuscle myosin isoforms to produce a tree depicting points of species divergence. We also see how the pace of the "molecular clock" varies drastically between different types of myosin. The brush-border myosins I have undergone >10 times more change since the divergence of mammals and chickens than have the nonmuscle myosins (see Fig. 1). Of course, one must remember that these two brush-border proteins may represent different brush-border isoforms which have been separate longer than mammals and chickens. Alterations of the clock can confound attempts at determining evolutionary relationships because, if severe enough, they can cause artifactual results (10). They can also be useful, however, because they suggest that changes in function have altered constraints on the sequences. Correlation of altered function with regions of sequence under different constraints in dif-

ferent types of myosin may help us to dissect the relationship between particular structures and functions.

A question that cannot be answered by this type of analysis is which type of myosin is the most ancestral form. While our data imply that striated muscle is "older" than smooth muscle and suggest that these types of muscle evolved independently from nonmuscle tissue, the information available does not allow us to assign the root of our trees. However, some hints do exist. For example, when one counts amino acid positions conserved between any two classes but divergent in the third, one finds that there are many more positions conserved between myosin II and the dilute class than between either classic myosin I and the dilute class or between classic myosin I and myosin II (data not shown). This result suggests that myosin II and the dilute class are the most closely related. Examination of branch lengths in Fig. 1 suggests a similar conclusion: node U marks the point at which lines leading to all three major myosin classes split. Nodes A, B, and C mark approximately equivalent species divergences in the three main classes of myosin. There is twice as much distance from node U to node A (the divergence of yeast and amoeboid classic myosin I) as there is from this point to node B (the divergence of yeast, vertebrate, and amoeboid myosin II) or node C (the divergence of yeast and vertebrate dilute-class myosin I). The sum of this information suggests that the root of the myosin tree lies somewhere between the node U and node A, possibly near the node marking the divergence of *Acanthamoeba* high molecular weight myosin I (the midpoint between nodes A, B, and C lies there). Of course, the inconstancy of the molecular clock as described above means that any conclusions derived from these data must be treated with caution.

A more interesting hint is provided by recent analysis of the evolution of myosin II light chains, which shows that they are derived from calmodulin (or its immediate ancestor) (20, 21). This information suggests that the primordial myosin gene had calmodulin for light chains, and at some point a gene duplication occurred, allowing one of the calmodulin genes to become specific to the myosin II protein. Dilute-class myosins, as well as the brush-border myosins I, still have calmodulin for light chains. It is simpler to imagine that these myosin proteins started out with calmodulin light chains and never lost them than to suggest that they started out with conventional myosin light chains (or no light chains) and regained the ability to bind calmodulin. These lines of reasoning suggest that both of these classes of myosin I are closer to the ancestral state of myosin than is myosin II. Further testing of this hypothesis awaits the characterization of additional myosin I light chains and the identification of new classes of myosin I.

**Note Added in Proof.** Espreafico *et al.* (23) have recently performed a phylogenetic analysis of myosin head-domain sequences, using a related distance matrix method.

We gratefully acknowledge the computational expertise of Tod Klingler and Gur Hoshen and are indebted to Hans Warrick for helpful discussions and careful reading of the manuscript. We thank Susan Brown, Elliott Sherr, Kathy Kellerman, Kathy Miller, Weidong Sun, and Dan Kiehart for sharing sequence data before publication. This work was supported by National Institutes of Health Grant GM46551 to J.A.S.

1. Emerson, P. & Bernstein, S. I. (1987) *Annu. Rev. Biochem.* **56,** 695–726.
2. Warrick, H. M. & Spudich, J. A. (1987) *Annu. Rev. Cell Biol.* **3,** 379–421.
3. Korn, E. D. & Hammer, J. A. III. (1990) *Curr. Op. Cell Biol.* **2,** 57–61.
4. Pollard, T. D., Doberstein, S. K. & Zot, H. G. (1991) *Annu. Rev. Physiol.* **53,** 653–681.
5. Cheney, R. E. & Mooseker, M. S. (1992) *Curr. Op. Cell Biol.* **4,** 27–35.
6. Feng, D. F. & Doolittle, R. F. (1990) in *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, ed. Doolittle, R. F. (Methods in Enzymology), Vol. 183, pp. 375–387.
7. Felsenstein, J. (1989) *Cladistics* **5,** 164–166.
8. Fitch, W. M. & Margoliash, E. (1967) *Science* **155,** 279–284.
9. Klotz, L. C. & Blanken, R. L. (1981) *J. Theor. Biol.* **91,** 261–272.
10. Felsenstein, J. (1988) *Annu. Rev. Genet.* **22,** 521–565.
11. Higgins, D. G., Bleasby, A. J. & Sharp, P. M. (1992) *CABIOS* **8,** 189–191.
12. Felsenstein, J. (1985) *Evolution* **39,** 783–791.
13. George, E. L., Ober, M. B. & Emerson, C. P. Jr. (1989) *Mol. Cell. Biol.* **9,** 2957–2974.
14. Matsuoka, R., Beisel, K. W., Furutani, M., Arai, S. & Takao, A. (1991) *Am. J. Med. Genet.* **41,** 537–547.
15. Fukui, Y., Lynch, T. J., Brzeska, H. & Korn, E. D. (1989) *Nature (London)* **341,** 328–331.
16. Stedman, H. H., Eller, M., Jullian, E. H., Fertels, S. H., Sarkar, S., Sylvester, J. E., Kelly, A. M. & Rubinstein, N. A. (1990) *J. Biol. Chem.* **265,** 3568–3576.
17. Stewart, A. F., Camoretti-Mercado, B., Perlman, D., Gupta, M., Jakovcic, S. & Zak, R. (1991) *J. Mol. Evol.* **33,** 357–366.
18. Moore, L. A., Tidyman, W. E., Arrizubieta, M. J. & Bandman, E. (1992) *J. Mol. Biol.* **223,** 383–387.
19. Moore, L. A., Tidyman, W. E., Arrizubieta, M. J. & Bandman, E. (1992) *J. Mol. Evol.*, in press.
20. Moncrief, N. D., Kretsinger, R. H. & Goodman, M. (1990) *J. Mol. Evol.* **30,** 522–562.
21. Collins, J. H. (1991) *J. Mus. Res. Cell Motil.* **12,** 3–25.
22. Katsuragawa, Y., Yanagisawa, M., Inoue, A. & Masaki, T. (1989) *Eur. J. Biochem.* **184,** 611–616.
23. Espreafico, E. M., Cheney, R. E., Matteoli, M., Nascimento, A. A. C., DeCamilli, P. V., Larson, R. E. & Moosekes, M. S. (1993) *J. Cell Biol.*, in press.