The Human Embryonic Myosin Heavy Chain

COMPLETE PRIMARY STRUCTURE REVEALS EVOLUTIONARY RELATIONSHIPS WITH OTHER DEVELOPMENTAL ISOFORMS*

(Received for publication, November 1, 1989)

Hannel H. Stedman†, Mark Eller‡, Eric H. Julian‡, Scott H. Ferteig‡, Satyapriya Sarkar‡, James E. Sylvester†, Alan M. Kelly‡, and Neal A. Rubinstein‡

From the IDepartment of Anatomy, School of Medicine and the IIDepartment of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, the IIIpartment of Anatomy and Cellular Biology, Tufts University, School of Medicine, Boston, Massachusetts 02111, and the IVpartment of Pathology, Hahnemann University, Philadelphia, Pennsylvania 19102

We have isolated a single 6021-nucleotide cDNA fragment encoding the full length of the myosin heavy chain (MHC) isoform initially expressed in developing human limb muscle. The corresponding transcript is expressed in fetal, but not adult, human muscle, and the corresponding gene maps to human chromosome 17. Comparison of the full length nucleotide sequence with that of the orthologous rat gene transcript reveals 74, 90, and 80% similarities in the 5'-untranslated, coding, and 3'-untranslated regions, respectively. To precisely quantify the degree of nucleotide sequence divergence between the human embryonic and other developmentally regulated MHC gene transcripts, we utilize the algorithm of Perler et al. (Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) Cell 20, 555-566) and make use of the codon-for-codon register attainable in alignments of the MHC rod encoding cDNA fragments. The results allow reconstruction of the order and relative timing of certain gene duplication events involved in the evolution of the multimembered mammalian MHC loci. By this analysis, the principal sarcomeric MHC gene expressed in the 14-day chick embryo is shown to be more distantly related to the mammalian embryonic MHC genes than to those expressed peri- and postnataly. Attention is focused on regional patterns of MHC sequence conservation, ordered with reference to the topology of our phylogenetic tree. We present a composite map depicting the deduced evolutionary age of various primary structural subdomains of the human embryonic MHC.

The myosin heavy chain (MHC) gene family encodes a series of protein isoforms whose developmental regulation plays a crucial role in normal muscle morphogenesis and specialization (Hoh and Yeoh, 1979; Whalen et al., 1979, 1981; Bandman et al., 1982; Nadal-Ginard et al., 1982; Robbins et al., 1986; Crow and Stockdale, 1986a, 1986b; Weydert et al., 1987; Rubinstein et al., 1988). Primary structural differences among the MHC isoforms are thought to relate directly to the differences in ATPase rate, force generation, and thermodynamic efficiency observed among individual muscle fibers and within the same fiber at different stages of development (Reiser et al., 1985a, 1985b, 1988; Alpert and Mulieri, 1986; Sweeney et al., 1986; Park et al., 1987; Acker et al., 1987). The range of electrophoretic mobilities exhibited by myosins or their proteolytic fragments provided early evidence for the existence of differentially expressed genes (Whalen et al., 1979; Hoh and Yeoh, 1979). Presently, monoclonal antibodies provide the best probes for the resolution of MHC gene expression at the single fiber level (Miller et al., 1985; Narusawa et al., 1985; 1987; Taylor and Bandman, 1989), while cDNA probes for specific mRNA transcripts are useful in the study of tissue extracts (Nadal-Ginard et al., 1982; Kropp et al., 1987). Further refinement of these techniques and their extension to other species will require improved primary structural representations of the various MHC isoforms.

Recombinant DNA-based research strategies have recently contributed an appreciable amount of primary structural information about both sarcomeric MHCs and their distantly related nonsarcomeric counterparts (Karn et al., 1983; Periasamy et al., 1984; Strehler, et al., 1986; Saenz and Leinwand, 1986; Warrick et al., 1986; Hammer et al., 1987; Yanagisawa et al., 1987; Casimir et al., 1988). Our laboratories are interested in the relationship between developmentally regulated isoform transitions and the corresponding diversification of myosin structure and function. To extend the scope of our general investigations, we have recently isolated and characterized cDNA fragments encoding a MHC isoform expressed early in the development of human skeletal muscle (Eller et al., 1989). Here we present the complete primary structure of the encoded polypeptide deduced from a full length cDNA sequence. These data allow the quantitation of divergence times among a large variety of previously characterized MHC gene fragments. Comparisons between aligned amino acid sequences of distantly related MHCs yield information about regional differences in functional constraint. With this information at hand, we construct a detailed map of amino acid sequence conservation as it relates to predicted higher order structural regions in the MHC S-1 domain.

MATERIALS AND METHODS

Isolation and Characterization of Cloned cDNA—Monoclonal antibody 2B9 (Gambke and Rubinstein, 1984), which specifically binds the embryonic MHC, was used to screen a plating of 5 × 10^6 plaques from our Xgtll 18-week human fetal muscle library. Following im-
Human Embryonic Myosin Heavy Chain: Structure and Evolution

RESULTS

Isolation of Cloned cDNA—Fig. 1 depicts cDNA fragment HEMHC-1, isolated from the human fetal muscle Agt11 library following immunodetection of the expressed 6-galactosidase fusion polypeptide. EcoRI restriction of DNA mini-preparations identified the largest cloned fragment. This fragment, termed HEMHC-1 (see Fig. 1), was subcloned into the B-H lactamase encoding plasmid vector Bluescript (Stratagene) and restricted-mapped. Sequence analysis of both ends of HEMHC-1 allowed placement of the restriction-mapped cDNA relative to the published rat embryonic MHC coding sequence. Subfragment A (Fig. 1), the product of a HindIII/EcoRI double digest, was 32P-labeled by the random primer method (Feinberg and Vogelstein, 1984) and utilized to screen our Agt10 human fetal muscle library (Stedman et al., 1988). HEMHC-2 was the largest cDNA insert recovered on this screen (107 plaques) from the amplified library. This EcoRI-linkeled cDNA was subcloned and found to share the HEMHC-1 restriction map over 3.5 of its 6 kb in length (Fig. 1). HindIII, PstI, and SstI subfragments of HEMHC-2 were subcloned at quasirandom into the same vector. The nucleotide sequence of the double-stranded DNA was determined by the dyeoxie chain termination method (Sanger et al., 1977).

Northern Blots—Northern blots were prepared and hybridized as previously described (Stedman et al., 1988). A synthetic oligonucleotide corresponding to 25 nucleotides from the 3' untranslated region of HEMHC-2 was synthesized by the University of Pennsylvania DNA synthesis service (photolithographic solid-phase method) and was sequentially 5'-end-labeled with [γ-32P]ATP (Maxam and Gilbert, 1980) and tagged with [γ-32P]dCTP (Deng and Wu, 1981) to achieve high specific activity. This fragment was termed HEMHC-01. Subfragment D (see Fig. 1) was labeled by the "random primer" technique with hexadeoxynucleotide primers and the Klenow fragment of DNA polymerase (Feinberg and Vogelstein, 1984).

Southern Blots—Genomic DNA fragments hybridizing the cDNA subfragments B and C of HEMHC-1 (see Fig. 1) were localized as previously described (Nussbaum et al., 1985; Stedman et al., 1988). A secondary subpanel which discriminates human chromosomes 5 and 17 was also utilized (data available on request).

Sequence Analysis—Nucleotide and deduced amino acid sequences were analyzed with programs from three software packages: those of PCGENE (IntelliGenetics), SEQUENCE ANALYSIS SOFTWARE PACKAGE (University of Wisconsin Genetics Computer Group [GCG]), and the CALTECH PACKAGE (Molecular Biology Computer Research Resource). The PCGENE programs ASCALE, DIA PRO, GARNIER, NAOUT, NMATPUS, NOYVOTNY, RESTRI, SE QUIN, and TRANSIL were used to reconstruct the full length HEMHC-2 nucleotide sequence, deduce the encoded amino acid sequence, and conduct secondary structure predictions by the methods of Garnier et al. (1978) and Chou and Fasman (1978) as modified by Novotny and Auffray (1984). The decision constant used in the implementation of the GARNIER program were those recommended for proteins with 20-50% secondary structure (~75, ~85, 0, and 0 centinats for helical, extended, turn, and coil configurations, respectively). Alteration of these settings to those recommended for proteins of lesser or greater secondary structure generally changed the length but not the position of domains predicted in consensus with the Chou and Fasman algorithms. The GCG programs GAP, LINEUP, CO DON FREQUENCY, and DIVERGE were used to align the amino acid and nucleotide sequences, calculate codon usage, and calculate the pairwise nucleotide sequence divergence. Matrices for the MHC rod sequences were generated with the CalTech MATRIX program. The additional nucleotide sequences analyzed were obtained from the June 1988 release of the GenBank data base (56.0) or entered by hand from the original manuscripts.

Expression of the Human Embryonic MHC Transcrite—A 25-base oligomer (HEMHC-01) complementary to the HEMHC-1 sense strand in the 3'-untranslated region, an area of maximal isorform heterology, was synthesized and end-labeled to high specific activity. Following hybridization and wash at 52 °C in 0.9 M NaCl (approximately 5 °C below the empirically determined denaturation temperature), HEMHC-01 detects an approximately 6-kb mRNA species unique to human fetal muscle (Fig. 2A). The autoradiogram in Fig. 2B was exposed from the same Northern blot following rehybridization with the 3.0-kb subfragment D of HEMHC-1. Here, the probe spans regions of sufficient intersform homology to detect MHC mRNA in preparations from human adult skeletal (lane 2), ventricular (lane 3), and atrial (lane 4) muscle, as well as from human fetal muscle (lane 1). This further documents the specificity of the oligomer HEMHC-01.

Physical Linkage among Human MHC Gene Family Members—Chromosomal assignment for homologous human genomic sequences detected by subfragments B and C of HEMHC-1, at moderate and high stringency, respectively, was accomplished utilizing a Southern blot of DNA preparations from a panel of hamster-human hybrid fibroblast cell lines (Fig. 3). Subfragment B hybridized at least six BgIII fragments localizing to human chromosome 17, while subfrag-
ment C detected only one of these BglII fragments. The hybrid panel illustrated does not allow discrimination between chromosome 5 and chromosome 17 sequences. Reutilization of the probes on a discriminatory subpanel (data not shown) allowed assignment to chromosome 17, in agreement with previously published results for skeletal MHC cDNA-detected sequences (Leinwand et al., 1983; Edwards et al., 1985). It has been recently reported that two tightly linked cardiac MHC genes map to human chromosome 14 (Saez and Leinwand, 1987). The localization of all BglII fragments detected by subfragment B of HEMHC-1 to human chromosome 17 suggests that this skeletal muscle cDNA probe retains specificity for those MHC genes primarily expressed in skeletal muscle. Since the additive molecular weight of the fragments detected with the 1.3-kb subfragment B is approximately 30 kb, and published vertebrate sarcomeric MHC gene sequences indicate an approximate 3:1 intron-to-exon size ratio, it is most likely that this result represents sequencing homology between several linked members of this MHC gene subfamily.

Nucleotide Sequence Determination—The total length of cloned fragment HEMHC-2 is 8021 nucleotides: 74 nucleotides in the 5'-untranslated region, 5820 in the coding region, and 127 in the 3'-untranslated region. The nucleotide sequence has been previously recorded (Eller et al., 1989). In the present work we provide the derived amino acid sequence of HEMHC-2 (Figs. 5 and 6) as the basis for evolutionary comparisons with other published MHC sequences. Simple comparison of the aligned human and rat (embryonic MHC) S1 coding sequences reveals homologies of 90.2 and 90.8% at the nucleotide and amino acid levels, respectively, while the corresponding values for the rod domain are 89.6 and 97.7%. The calculated amino acid sequence divergence for the full length MHCs from these species is only 1.86% or about one-fifth of the 16.5% previously reported for a comparison of the rat embryonic MHC sequence with an MHC from 14-day embryonic chick (presumed to represent an orthologous gene comparison) (Molina et al., 1987). Divergence distances similarly calculated for cytochromes and globins isolated from primate, rodent, and avian species yield ratios of approximately 1:3, consistent with paleontological evidence dating the avian/mammalian split and the mammalian radiation at 300 and 85 million years before the present, respectively (Romero-Herrera et al., 1973; Perler et al., 1980 (and references therein)). To gain further insight as to the source of the apparent discrepancy in divergence ratios (1:3 versus 1:9), we undertook the quantitation of nucleotide sequence divergence rates among available MHC cDNA sequences. Many of the published sequences correspond to only the carboxyl-terminal portions of the MHC rods, precisely the domain in which codon register is best preserved. Among the MHC rod-coding sequences that can be unambiguously aligned codon-for-codon (a list which includes, with minor exceptions, all muscle sequences thus far published) it is possible to apply the pairwise comparison algorithm of Perler et al. (1980) without introducing artifact in the process of gap element manipulation (Feng and Doolittle, 1987).

Table IA shows the percentage corrected divergence at replacement sites among 10 sets of paired MHC rod cDNA sequences. The GCG GAP program and dot matrix comparisons between paired sarcomeric and nonsarcomeric MHC amino acid sequences were used to identify the best rod cDNA sequence alignments. Based on these data (not shown, available on request) our aligned full length rod sequences all begin just carboxyl-terminal to the highly conserved S1-terminating proline residues and retain this register to their respective carboxyl termini. These alignments required the insertion of gap elements to accommodate only the previously noted differences in "skip residue" positions among the nonsarcomeric MHCs (McLachlan and Karn, 1983; Warrick et al., 1986; Yanagisawa et al., 1987). Because of the nonequivalent lengths of the various MHC rod sequences, all were truncated at

Table I

|    | HE | RE | CE | NS | CSm |
|----|----|----|----|----|-----|
| A. Rod domain |    |    |    |    |     |
| RE | 1.14 |     |    |    |     |
| CE | 11.2 | 11.4 |    |    |     |
| NS | 52.7 | 52.2 | 52.4 |    |     |
| CSm | 69.2 | 68.7 | 72.5 | 76.2 |     |
| DC | 98.4 | 97.8 | 100.6 | 111.4 | 97.6 |
| HA | 13.3 | 13.3 | 7.47 | 51.5 |     |
| HZ | 18.6 | 19.0 | 15.9 | 50.6 | 18.3 |
| B. S1 domain |    |    |    |    |     |
| RE | 0.706 |     |    |    |     |
| CE | 10.6 | 10.5 |    |    |     |
| NS | 50.3 | 50.2 | 48.7 |    |     |
| CSm | 52.9 | 51.9 | 51.4 | 52.5 |     |
| DC | 56.6 | 67.1 | 58.0 | 58.9 | 51.9 |

Results of quantitative comparisons, by the method of Perler et al. (1980), among coding regions of published MHC gene sequences

The data in A represent the result of comparisons among coding sequences corresponding to the HEMHC-2 rod sequence depicted, those in B the HEMHC-2 S1 domain (see Eller et al., 1989). The data presented in the last four rows of A are based on partial length nucleotide sequence information for the cDNAs labeled HA, HZ, RN, and SC. The rightmost number in row HZ depicts the comparison between the HA and HZ sequences. Values shown are the percentage (corrected) of nucleotide sequence divergence. HE, human embryonic sarcomeric; RE, rat embryonic sarcomeric (Strehler et al., 1986); CE, chicken embryonic sarcomeric (Molina et al., 1987); RN, rat neonatal sarcomeric (Ferriasmay et al., 1984); HA, human adult ("fast") sarcomeric (Saez and Leinwand, 1986); HZ, human slow ("fast") sarcomeric (Saez and Leinwand, 1986); NS, nematode (unc 54) sarcomeric (Karn et al., 1983); CSm, chicken smooth muscle (Yanagisawa et al., 1987); DC, slime mold cytoplasmic (Warrick et al., 1986); SC, newt "cardio-skeletal" (Casimir et al., 1988).

Fig. 3. MHC gene copy number and chromosomal localization as revealed by Southern analysis. Lanes correspond to BglII-digested DNA from 10 different human-hamster hybrid cell lines (A-J), one hamster (K), and three human cell lines (M-O). Lanes A-O are from an autoradiogram after hybridization with 32P-labeled subfragment C, then washed at 65°C in 15 mM NaCl. Lanes D', K', and O' correspond to lanes D, K, and O after elution of subfragment C, rehybridization with 32P-labeled subfragment B, and washing at 65°C in 15 mM NaCl. For details of the human chromosome content of each hybrid cell line, see Nussbaum et al. (1985) and Stedman et al. (1988). The human specific BglII fragments detected by both subfragments B and C segregate with perfect concordance only with human chromosome 17.
HEMHC-2 nucleotide 5883 (corresponding to rod amino acid 1086 in Fig. 5A).

cDNA sequences encoding the S1 domains of six of these MHCs have been similarly compared in Table II. The nucleotide sequence alignments utilized to generate this data set are based codon-for-codon on the amino acid sequence alignments shown in Fig. 7. Although each gap element in the sequence alignments corresponds to at least one insertion/deletion event in the divergence of the genes represented, these events are not considered in our quantitative comparison of the nucleotide sequence divergences. We estimate that the error resulting from this simplification is less than 1% with this set of nucleotide sequences (i.e. we see evidence of greater than 100 nucleotide substitutions per insertion/deletion event).

The phenogram in Fig. 4 illustrates the possible evolutionary relationships among these MHC genes based on the data in Table IA and the tree generation algorithm of Fitch and Margoliash (1967), as modified by Doolittle (1986). It is important to note that this approach is based entirely on the quantification of pairwise differences. This simplification facilitates the incorporation of data from partial length cDNAs (the four rows at the bottom of Table IA). The meaningful application of alternative algorithms, especially those that mathematically optimize evolutionary parsimony (Fitch, 1977; Penny and Hardy, 1986), must await the sequence analysis of full length cDNAs for the various MHCs. At present we regard the overall tree topology as reliable, but the distal branch arm lengths as tentative (especially among the partial fragments of genes HA, HZ, RN, and SC).

In Fig. 4, a vertical stripe corresponding to ~300 million years has been placed in the immediate vicinity of nodes G and H, offering one provisional normalization of the distal part of the phenogram. This assumes orthology of the HZ and SC gene pair, as suggested by Casimir et al. (1988). In addition, a second vertical stripe at ~85 million years has been placed through node F, corresponding to the rat-human ancestral split.

**Patterns of Sequence Conservation in the Rod Domain—** Differences in the rates of replacement site sequence divergence between S1 and rod domains are apparent from Table I, A and B, and may reflect different levels of functional constraint within different portions of the encoded polypeptide. In an effort to further localize MHC-coding regions subject to extraordinary selective pressure, we focused further attention on our MHC amino acid sequence alignments.

Fig. 5A shows the entire human embryonic rod sequence, arranged in the manner of McLachlan and Karn (1983) to highlight the 7- and 28-amino acid repeats that characterize this largely α-helical portion of the polypeptide. The numbers of amino acid residues found to differ in five pairwise comparisons with other published vertebrate rod sequences are tabulated to the right of the sequence. This analysis reveals the extraordinarily high level of sequence conservation in three regions of the vertebrate rod, centered around repeats 20, 34, and 35. The uniqueness of the structure of these subdomains is apparent in Fig. 5B, where the individual amino acids are replaced by symbols representing hydrophobic or charged side groups. When charge or hydrophobicity alone is considered, even the comparison between the human embryonic and the nematode sarcomeric sequences shows extensive conservation in these subdomains. Repeats 20 and 34 are both dramatically different from the consensus structure shown at the bottom of the diagram.

The conservation of charge distribution among the rod sequences is extraordinary, as demonstrated by the numerical columns to the right of the sequence diagrams. Although 42.5% of the residues in the HEMHC rod are charged, fewer than 8% of the substitutions in a comparison with the chicken embryonic sarcomeric (CE) rod result in changes in the charge distribution (not shown). Furthermore, in a similar comparison involving the more distantly related nematode (unc 54) (NS) rod sequence, the mutation rate among charged residues is approximately three times as high in the S2 region (top 10 lines of column A) as in the remainder of the rod. This effect disappears in comparison between sarcomeric and nonsarcomeric rods, as shown for the human embryonic sarcomeric/chicken smooth muscle (HE/CSm) comparison in column B of Fig. 5B.

**Secondary Structure Predictions for the S1 Domain—** In contrast, substitution site mutations appear at dramatically different position-dependent rates throughout the S1 domain. The deduced amino acid sequence of the human embryonic MHC S1 domain, its simplification by amino acid subclass (by the algorithm shown), and its further simplification based on the hydropathicity scale of Kyte and Doolittle (1982) are aligned in Fig. 6, rows a, b, and c, respectively. Colored backgrounds are used to highlight subdomainal patterns of differential sequence conservation. In the first row there are seven independent regions where primary structure is rigidly conserved among all MHCs compared (blocks of eight or greater contiguous residues against a white background). The patterns of conservation depicted in the second and third rows buttress five of these regions, drawing attention to the conservative nature of most of the flanking amino acid substitutions. Amino acids shown against a blue background on the second row depict positions of nonconservative substitution among even the closely related vertebrate MHCs. Secondary structural domains predicted by the independent methods of Chou and Fasman (1978) and Garnier et al. (1978) are shown in rows d and e, respectively. An intervening line appears in the HEMHC-2 nucleotide 5883 (corresponding to rod amino acid 1086 in Fig. 5A).

cDNA sequences encoding the S1 domains of six of these MHCs have been similarly compared in Table II. The nucleotide sequence alignments utilized to generate this data set are based codon-for-codon on the amino acid sequence alignments shown in Fig. 7. Although each gap element in the sequence alignments corresponds to at least one insertion/deletion event in the divergence of the genes represented, these events are not considered in our quantitative comparison of the nucleotide sequence divergences. We estimate that the error resulting from this simplification is less than 1% with this set of nucleotide sequences (i.e. we see evidence of greater than 100 nucleotide substitutions per insertion/deletion event).

The phenogram in Fig. 4 illustrates the possible evolutionary relationships among these MHC genes based on the data in Table IA and the tree generation algorithm of Fitch and Margoliash (1967), as modified by Doolittle (1986). It is important to note that this approach is based entirely on the quantification of pairwise differences. This simplification facilitates the incorporation of data from partial length cDNAs (the four rows at the bottom of Table IA). The meaningful application of alternative algorithms, especially those that mathematically optimize evolutionary parsimony (Fitch, 1977; Penny and Hardy, 1986), must await the sequence analysis of full length cDNAs for the various MHCs. At present we regard the overall tree topology as reliable, but the distal branch arm lengths as tentative (especially among the partial fragments of genes HA, HZ, RN, and SC).

In Fig. 4, a vertical stripe corresponding to ~300 million years has been placed in the immediate vicinity of nodes G and H, offering one provisional normalization of the distal part of the phenogram. This assumes orthology of the HZ and SC gene pair, as suggested by Casimir et al. (1988). In addition, a second vertical stripe at ~85 million years has been placed through node F, corresponding to the rat-human ancestral split.

**Patterns of Sequence Conservation in the Rod Domain—** Differences in the rates of replacement site sequence divergence between S1 and rod domains are apparent from Table I, A and B, and may reflect different levels of functional constraint within different portions of the encoded polypeptide. In an effort to further localize MHC-coding regions subject to extraordinary selective pressure, we focused further attention on our MHC amino acid sequence alignments.

Fig. 5A shows the entire human embryonic rod sequence, arranged in the manner of McLachlan and Karn (1983) to highlight the 7- and 28-amino acid repeats that characterize this largely α-helical portion of the polypeptide. The numbers of amino acid residues found to differ in five pairwise comparisons with other published vertebrate rod sequences are tabulated to the right of the sequence. This analysis reveals the extraordinarily high level of sequence conservation in three regions of the vertebrate rod, centered around repeats 20, 34, and 35. The uniqueness of the structure of these subdomains is apparent in Fig. 5B, where the individual amino acids are replaced by symbols representing hydrophobic or charged side groups. When charge or hydrophobicity alone is considered, even the comparison between the human embryonic and the nematode sarcomeric sequences shows extensive conservation in these subdomains. Repeats 20 and 34 are both dramatically different from the consensus structure shown at the bottom of the diagram.

The conservation of charge distribution among the rod sequences is extraordinary, as demonstrated by the numerical columns to the right of the sequence diagrams. Although 42.5% of the residues in the HEMHC rod are charged, fewer than 8% of the substitutions in a comparison with the chicken embryonic sarcomeric (CE) rod result in changes in the charge distribution (not shown). Furthermore, in a similar comparison involving the more distantly related nematode (unc 54) (NS) rod sequence, the mutation rate among charged residues is approximately three times as high in the S2 region (top 10 lines of column A) as in the remainder of the rod. This effect disappears in comparison between sarcomeric and nonsarcomeric rods, as shown for the human embryonic sarcomeric/chicken smooth muscle (HE/CSm) comparison in column B of Fig. 5B.

**Secondary Structure Predictions for the S1 Domain—** In contrast, substitution site mutations appear at dramatically different position-dependent rates throughout the S1 domain. The deduced amino acid sequence of the human embryonic MHC S1 domain, its simplification by amino acid subclass (by the algorithm shown), and its further simplification based on the hydropathicity scale of Kyte and Doolittle (1982) are aligned in Fig. 6, rows a, b, and c, respectively. Colored backgrounds are used to highlight subdomainal patterns of differential sequence conservation. In the first row there are seven independent regions where primary structure is rigidly conserved among all MHCs compared (blocks of eight or greater contiguous residues against a white background). The patterns of conservation depicted in the second and third rows buttress five of these regions, drawing attention to the conservative nature of most of the flanking amino acid substitutions. Amino acids shown against a blue background on the second row depict positions of nonconservative substitution among even the closely related vertebrate MHCs. Secondary structural domains predicted by the independent methods of Chou and Fasman (1978) and Garnier et al. (1978) are shown in rows d and e, respectively. An intervening line appears in

![Fig. 4. Phenogram depicting approximate MHC gene divergence chronology, based on the data in Table I. Portions of the diagram left of and below point I are not shown to scale with respect to the remainder. The shaded vertical band at left corresponds to the estimated period during which the three-way amphibian-avian-mammalian split occurred, while that at right represents the mammalian radiation, both as reconstructed on the basis of the fossil record. Letters A–I mark the points at which putative ancestral genes diverged to yield the modern day genes represented at extreme right. Abbreviations are defined in the legend to Table I.](http://www.jbc.org/Downloaded from)
to the right of the diagram in five numerical columns. See legend to Table I for abbreviation definitions. B, distribution of hydrophobications, as identified by comparison with other MHC sequences, appears of amino acid substitutions resulting in changes in the net charge at

Based on S2, S2 hinge, and LMM regions of the MHC rod are demarcated at positions where the two structural predictions are in agreement.

A notable feature of the MHC amino acid sequence alignment on which Fig. 6 is based (see Fig. 7) is the concentration of insertion/deletion mutations in the two regions corresponding to the protease-sensitive sites of rabbit skeletal MHC S1. Elsewhere in the aligned S1 domain sequences, the scarcity of such "gap elements" implies the existence of selective pressure to conserve the relative positions of the highly conserved subdomains backlighted in white in Fig. 6, row a. The largest region of low level primary sequence conservation corresponds to HEMHC residues 290-455. In our search for interspecific similarities over this region, it was noted that both hydropathicity and secondary structural propensities were conserved among distantly related MHC amino acid sequences.

In Fig. 7, we have aligned the amino acid sequences of the S1 regions of five MHCs (human embryonic skeletal, chicken skeletal, nematode sarcomeric, chicken smooth muscle (nonsarcomeric), and slime mold (cytoplasmic)), and we have subjected all five to secondary structural analysis by the methods of Chou and Fasman (1978) and Garnier et al. (1978). The following discussion only applies to regions where both the Chou and Fasman and the Garnier analyses showed a consensus. In pairwise comparisons between these aligned MHCs, many regions of agreement in predicted secondary structure correspond to regions of relatively little primary structural homology (e.g. α-helices A1, A4, A5, and A6). In total, 16 α-helical and 9 β-sheet domain positions are in agreement among all the sequences analyzed.

**DISCUSSION**

**Relationship to Other Cloned DNAs**

Genomic and complementary DNA fragments encoding MHCs from a wide range of cellular sources have recently been characterized (Hammer et al., 1987; Karn et al., 1983; Molina et al., 1987; Strehler et al., 1986; Warrick et al., 1986; Watts et al., 1987). The cytoplasmic MHCs synthesized by two unicellular organisms, Saccharomyces cerevisiae (Watts et al., 1987) and Dictyostelium discoideum (Warrick et al., 1986), are encoded by single copy intron-lacking genes. To our knowledge, clone HEMHC-2 represents the first solitary cDNA fragment encoding the full length of a MHC of muscle origin. The importance of this fact lies in the desirability of expressing sarcomeric MHC coding sequences in nonmuscle cells (de Vos et al., 1988; Maruyama and MacLennan, 1988).

**MHC Genes**—The similarity in replacement site substitution rates for the S1 and rod domains in the comparison of human embryonic to the chicken embryonic and nematode (MHC B) sarcomeric MHC sequences suggests that following sarcomeric MHC gene divergence, mutations are initially fixed at similar overall rates in the two domains. However, differences between S1 and rod divergence rates emerge in comparisons involving smooth muscle or cytoplasmic MHCs. We offer two hypothetical explanations. In the first, the overall domainal divergence rates represent composites of subdomainal rates that vary with the functional constraints right. Flanking this are columns tabulating the number of amino acid substitutions (per 28-residue repeat) that result in differences in charge distribution, A for the comparison HE/NS and B for HE/CSm. At the foot of each column in B appears a symbol depicting the most prevalent amino acid type in that column. Solid, I, L, M, V; light shading, D, E; dark shading, H, K, R, %, hydrophilic.
FIG. 6. Structure of the human embryonic MHC S1 domain: a study in differential conservation. a, complete amino acid sequence of the HEMHC S1 domain deduced from the nucleotide sequence in Fig. 2. Simplification of sequence in a by side chain type, where: A = A,G,P,S,T; C = C; D = D,E,N,Q; F = F,W,Y; H = H,K,R; and I = I,L,M,V. c, unidimensional depiction of regional hydropathicity based on the sequence in a, scale of Kyte and Doolittle (1982) normalized to I = 1, and local averaging over a window of five amino acids. Hydropathicity: (h) > 0.55 = *; 0.54 > (h) > 0.46 = -; 0.45 < (h) = %. d, secondary structure as predicted by the method of Chou and Fasman (1978). e, secondary structure as predicted by the method of Garnier et al. (1978). The broken line between rows d and e highlights regions of complete agreement. In a and b the level of conservation observed in sequence alignments (see Fig. 7) corresponds to the background color used as per the following relationship: white, conserved among all MHCs; green, all sarcomeric MHCs; orange, all vertebrate sarcomeric MHCs; blue, variable among all vertebrate sarcomeric MHCs. Above row a appear lower case letters depicting the least conservative substitutions thus far found at corresponding positions in other published vertebrate sarcomeric MHC sequences.

on various parts of the protein. If a proportion (on the order of 40%) of the amino acids in S1 are rigidly constrained, while the remaining positions can tolerate varying degrees of conservative amino acid substitution, the overall divergence rates might appear to plateau (as a corrected divergence on the order of 60% is approached). Our alternative hypothesis posits that approximately the same proportions of highly and poorly conserved amino acids exist in the S1 and rod domains of all sarcomeric MHCs but that the nature of thick and thin filament register in the myofibril imposes a further functional constraint on the rod sequences of only the sarcomeric MHCs. Thus, only among sarcomeric MHC gene sequences will overall divergence be found to occur at similar rates in S1 and rod-encoding domains. In any event, inconsistency in domainal mutation rates will influence calculations of relative divergence times principally among the most distantly related MHC genes compared.

Fig. 4 suggests that the mammalian slow skeletal (cardiac β) gene diverged from a common ancestor of the mammalian embryonic, neonatal, and adult fast genes more than 500 million years ago (node D). These data further suggest that the genomic reorganization events that brought about the physical segregation of the cardiac/slow and skeletal/fast MHC gene clusters (Saez et al., 1987) occurred between nodes D and E in the phenogram, approximately 400–600 million years ago. The data summarized in Fig. 4 indicate that the avian MHC gene expressed at the highest level in 12–14-day chick embryonic (Kropp et al., 1987) muscle is most likely to be orthologous to the human adult IIa MHC gene (HA). Furthermore, the human embryonic/IIa gene divergence long predated the separation of avian and mammalian species. At least seven differentially regulated MHC genes have been identified in the chicken genome (Kropp et al., 1987), all of which exhibit mutual sequence divergence levels (over the first three coding exons) markedly less than that found in the human-chicken embryonic MHC cDNA sequence comparison. (The coding sequence of one of these genes corresponds to the CE sequence in Table I.) This suggests that the chicken MHC gene family has expanded in number since the divergence depicted at point E in Fig. 4. There is insufficient data at present to correlate the timing of these gene duplication events with that of the mammalian IIa (HA) and neonatal (RN) MHC gene divergence (point H or I in Fig. 4). Moreover, independent evidence for a localized gene correction event...
FIG. 7. Consensus secondary structure predictions for five MHCs, aligned on the basis of primary structure. A consensus between the Chou and Fasman (1978) and the Gamier et al. (1978) predictions for each MHC was obtained, exactly as illustrated in Fig. 5 for HEMHC. Numbering is based on the HEMHC amino acid sequence. Two marker residues labeled by the ATP analogs N-(4-azido-2-nitrophenyl)-2-aminoethyl diphosphate (NANDP) and [5,6-H]UTP are highlighted by the letters N and U, respectively (Okamoto and Yount, 1985; Atkinson et al., 1986). The reactive sulfhydryls SH1 and SH2 are labeled accordingly (Yanagisawa et al., 1987). The consensus α-helical and β-sheet domains referred to in Fig. 7 are enumerated below the sequence alignments. Color code: green, α-helix; orange, β-strand; and purple, β-turn.

involving approximately one-third of the reported MHC coding sequence (Kropp et al., 1987) complicates the reconstruction of the duplication events in the evolutionary history of the chicken locus. However, events of the latter sort may help explain why quantitative comparisons involving the chicken embryonic cDNA sequence generated the anomalous length of the distal branch leading to point CE in Fig. 4 (as if the mutation rate had slowed along this branch).

One question raised by these findings is of importance to myologists concerned with developmentally regulated iso-myosin transition. If paralogous MHC genes have been selected for initial expression in avian and mammalian embryogenesis, what has become of the avian orthologue to the mammalian embryonic MHC gene? Divergent evolution of the regulatory elements associated with this coding sequence would indicate the operation of dramatically different selective pressures in different vertebrate lineages. Alternatively, a MHC gene orthologous to the mammalian embryonic MHC gene could be down-regulated sufficiently early in avian myogenesis that its transcript has thus far escaped detection. The appearance of pseudogenes at any of the multimember MHC gene loci is needed to complement the more detailed study of evolutionary relationships among members of this gene family. Furthermore, orthologous gene comparisons for additional MHCs will be required to determine whether the substitution rates are truly identical among all of the members of this gene family.

Rod Domain—The existence of 28- and 196-residue repeat patterns in the rod domains of distantly related MHCs (McLachlan, 1983; Warrick et al., 1986) has led to the hypothesis that a primordial MHC gene sequence arose through serial duplications of a 28-codon block, followed at some later point by at least four duplications of a 197-codon block ((28 x 7) + 1 “skip residue”). Furthermore, this higher order periodic distribution of charged residues within the rod domain has been taken as evidence for a 98-residue stagger between interacting myosin rods in the thick filament (McLachlan and Karn, 1983). Finally, the higher level of repetition seen on comparison matrices in a quadrant corresponding to the S2 portion of the nematode MHC B amino acid sequence has been taken as evidence for a comparatively greater level of sequence divergence (from the “primordial” 28-residue repeat) in the LMM region (McLachlan, 1983).

The data shown in Fig. 5B demonstrate greater positional and numerical conservation among charged residues in the
LMM and hinge than in the short S2 subdomains of distantly related sarcomeric MHCs. This finding suggests the emergence of a further functional constraint upon only the LMM and hinge subdomains after the divergence of genes encoding sarcomeric and nonsarcomeric MHCs. Sequence conservation among these residues would operate to preserve both the 197-residue periodicity and the potential for a 98-residue optimal stagger. Optimal stagger distance represents a candidate sarcomeric MHC-specific functional constraint. (The necessity for register between positively charged residues in one myosin rod and negatively charged residues 98 amino acids displaced in a second should manifest as selective pressure against evolutionary drift of charged residue distribution toward that of a consensus 20-residue repeat.)

A similar argument should apply for regions with critical roles in the initiation of myofibrillogenesis. Repeats numbered 20 and 34 in Fig. 5a are found in this comparison of five vertebrate sarcomeric MHC sequences to undergo amino acid substitutions at one-fourth to one-sixth the average rate among repeats, suggesting that these subdomains have unique roles in the overall function of the corresponding myosin rods.

SI Domain—Evolutionary comparisons of this sort yield evidence of a different series of functional constraints upon the primary structure of the MHC SI domain. Although regional differences in amino acid sequence conservation are readily appreciated in the MHC SI domain, it has been shown that there is comparatively little conservation of charged amino acid numbers and positions among distantly related MHC sequences (Warrick and Spudich, 1987). The color scheme used in Fig. 6 draws attention to primary structural domains that were established at various time points in the evolution of the human embryonic MHC sequence. For instance, all of the letters which appear on a white background correspond to primary structural domains common to all functional descendants of the ancestral MHC encoded by gene A in Fig. 4, while letters which appear against a brown background represent domains that were established as MHC gene C evolved to MHC gene D.

Most of the regions of rigid primary structural conservation are flanked by blocks of amino acids that have undergone only conservative substitutions. These moderately conserved flanking domains generally correspond to regions of rigid conservation among the closely related vertebrate sarcomeric (vs) MHCs, suggesting that selective pressure against nonconservative substitutions dramatically slows the rate at which any substitutions are fixed. There are regions where primary structural conservation exists among vsMHCs, while no detectable propensity for conservative substitution exists among more distantly related sequences (e.g. amino acids, 13-29 and 799-737). Such regions represent protein structural domains with roles possibly unique to the function/regulation of the vsMHCs. Among those residues at positions where variability is seen even among vsMHCs, Fig. 6 allows the identification of that subset for which the least conservative substitutions have been documented (residues against a blue background in Fig. 6, row b). In the absence of any detailed tertiary structural information regarding a native MHC globular head, these nonconservatively substituted sites represent the best candidates for roles in fiber type-specific (acto)myosin ATPase rate control.

Computer-assisted methodologies for extrapolating from protein secondary to tertiary structural predictions are presently restricted to helix-rich (Cohen et al., 1986) and parallel α/β (Cohen et al., 1983) domains. Based on the assumption of topographic similarity among functionally similar but distantly related proteins, evolutionary comparisons have been utilized to strengthen secondary structural predictions, usually in support of tertiary structural models (Crawford et al., 1987; Yanagisawa et al., 1987). Fig. 7 shows the consensus among secondary structural predictions for sequences as distantly related as those of the human embryonic sarcomeric and the slime mold cytoplasmic MHCs. All of these MHCs stoichiometrically bind light chains, and the resulting myosins all possess actin-activated ATPase activities. However, contractile activity is differently regulated among each of the three classes of myosin represented in these comparisons. Furthermore, there is evidence that the topography of light and heavy chain interactions differs among these myosins (Okamoto et al., 1986). Therefore, some of the predicted secondary structural differences shown in Fig. 7 may correspond to tertiary structural regions that differentially mediate regulatory signals in cytoplasmic, smooth, and skeletal muscle MHCs. That proportion of the SI domain for which the predictions in Fig. 7 are in complete agreement can be taken as the most conservative data base for use in the formulation of MHC tertiary structural models.

Finally, it is of interest to note that among vertebrate sarcomeric MHCs for which complete amino acid sequences of this region are available (see Fig. 5), only residues 136-138 exhibit variability within the immediate primary structural proximity of residues bound by ATP analogs (Okamoto and Yount, 1985; Atkinson et al., 1986; Sutoh, 1987). As additional full length MHC cDNA sequences become available, this map should be of use in the correlation of domainal sequence microheterogeneity with the actin-activated ATPase activities of the various tissue and developmental stage-specific myosins.

Acknowledgments—We would like to thank Drs. Britton Chance, Mildred Cohn, Yale Goldman, Frank Pepe, Avrii Somlo, and Andrew Somlo for their helpful suggestions. We acknowledge Dr. Robert L. Nussbaum for the generous gift of cell lines and DNA for use in chromosomal localization.

REFERENCES

Acker, M., Anderson, Q. A., Hammond, R. L., DiMeo, F., McCullem, J., Staun, M., Velchik, M., Brown, W. E., Gale, D., Salmons, S. & Stevenson, L. (1987) J. Thorac. Cardiovasc. Surg. 94, 702-709

Alpert, N. R. & Mulieri, L. A. (1986) Med. Sci. Sports Exercise 18, 309-313

Atkinson, M. A. L., Robinson, E. A., Appella, E. & Korn, E. D. (1986) J. Biol. Chem. 261, 1844-1848

Bandman, E. R., Matsuda, R. & Stroehman, R. C. (1982) Dev. Biol. 93, 508-518

Casimir, C. M., Gates, P. B., Ross-Macdonald, P. B., Jackson, J. F., Patient, R. K. & Brockes, J. P. (1988) J. Mol. Biol. 202, 287-296

Chou, F. Y. & Fasman, G. D. (1978) Adv. Enzymol Relat. Areas Mol. Biol. 47, 45-65

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. & Plettenberg, R. J. (1983) Biochemistry 22, 4894-4904

Cohen, F. E., Kosen, P. A., Kuntz, I. D., Epstein, L. B., Ciardelli, T. L. & Smith, K. A. (1986) Nature 324, 349-352

Crawford, I. P., Niermann, T. & Kirschner, K. (1987) Proteins Struct. Funct. Genet. 2, 118-129

Crow, M. T. & Stockdale, F. E. (1986a) Dev. Biol. 119, 238-254

Crow, M. T. & Stockdale, F. E. (1986b) Dev. Biol. 118, 333-342

Deng, G. & Wu, R. (1981) Nucleic Acids Res. 9, 4173-4179

de Vos, A. M., Tong, L., Milburn, M. V., Matias, P. M., Jancarik, J., Noguchi, S., Nishimura, M., Miura, K., Ohlsuka, E. & Kim, S. (1989) Science 240, 888-893

Doolittle, R. F. (1986) Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences, University Science Books, Mill Valley, CA

Edwards, Y. H., Parker, M., Popvey, S., West, L. F., Parrington, J. M. & Solomon, E. (1986) Ann. Hum. Genet. 49, 101-109

Eller, M., Stedman, H. H., Sylveste, J. E., Fertels, S. H., Rubinstein, N. A., Kelly, A. M. & Sarkar, S. (1989) Nucleic Acids Res. 17, 3591-3592
The human embryonic myosin heavy chain. Complete primary structure reveals evolutionary relationships with other developmental isoforms.
H H Stedman, M Eller, E H Jullian, S H Fertels, S Sarkar, J E Sylvester, A M Kelly and N A Rubinstein

*J. Biol. Chem.* 1990, 265:3568-3576.

Access the most updated version of this article at [http://www.jbc.org/content/265/6/3568](http://www.jbc.org/content/265/6/3568)

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 0 references, 0 of which can be accessed free at [http://www.jbc.org/content/265/6/3568.full.html#ref-list-1](http://www.jbc.org/content/265/6/3568.full.html#ref-list-1)