Cell, Vol. 53, 219–228, April 22, 1988, Copyright © 1988 by Cell Press

The Complete Sequence of Dystrophin Predicts a Rod-Shaped Cytoskeletal Protein

M. Koenig,*† A. P. Monaco,*‡ and L. M. Kunkel*†‖
*Division of Genetics
Department of Pediatrics
Harvard Medical School
The Children's Hospital
Boston, Massachusetts 02115
†The Howard Hughes Medical Institute
The Children's Hospital
Boston, Massachusetts 02115
‡The Program in Neuroscience
Harvard University
Cambridge, Massachusetts 02138

Summary

The complete sequence of the human Duchenne muscular dystrophy (DMD) cDNA has been determined. The 3685 encoded amino acids of the protein product, dystrophin, can be separated into four domains. The 240 amino acid N-terminal domain has been shown to be conserved with the actin-binding domain of α-actinin. A large second domain is predicted to be rod-shaped and formed by the succession of 25 triple-helical segments similar to the repeat domains of spectrin. The repeat segment is followed by a cysteine-rich segment that is similar in part to the entire COOH domain of the Dictyostelium α-actinin, while the 420 amino acid C-terminal domain of dystrophin does not show any similarity to previously reported proteins. The functional significance of some of the domains is addressed relative to the phenotypic characteristics of some Becker muscular dystrophy patients. Dystrophin shares many features with the cytoskeletal protein spectrin and α-actinin and is a large structural protein that is likely to adopt a rod shape about 150 nm in length.

Introduction

Duchenne muscular dystrophy (DMD) and Becker muscular dystrophy (BMD) are both human X-linked degenerative disorders of muscle. DMD affects about 1 in 3500 live born males and results in a fatal evolution before the end of the third decade of life, while BMD is less severe and less frequent (Moser, 1984). Until recently, the molecular and biochemical basis of DMD and BMD was unknown. Cytogenetically detectable alterations of the DMD/BMD locus (Verellen-Dumoulin et al., 1984; Francke et al., 1985) initiated the isolation of the gene which, when defective, results in Duchenne or Becker muscular dystrophy (Monaco et al., 1986; Koenig et al., 1987; Burghes et al., 1987). Portions of the coding sequence of the gene (cDNA) have been used to establish the muscle specificity of expression of the gene (Hoffman et al., 1987a; Lev et al., 1987) and to produce polyclonal antisera directed against the encoded protein product (Hoffman et al., 1987b). The antisera detect a protein in muscle, called dystrophin, that is about 400 kd in size and which has been shown to be associated the transverse tubules of the muscle triadic structures (Hoffman et al., 1987c; Knudson et al., 1988). The absence of dystrophin in DMD muscle and its altered size in BMD muscle (Hoffman et al., 1988) establishes the disruption of dystrophin as the primary biochemical defect in Duchenne/Becker muscular dystrophy. From partial sequence information on dystrophin (Koenig et al., 1987; Hoffman et al., 1987a), the N-terminal 240 amino acids of dystrophin have been shown to be homologous to the N-terminus of α-actinin (Hammond, 1987) which corresponds to a presumed actin-binding domain (Baron et al., 1987; Noegel et al., 1987), and the subsequent 1200 amino acids have been shown to exhibit weak repeats similar to those found in spectrin and α-actinin (Davison and Critchley, 1988).

In this paper, we report the entire DMD cDNA sequence and the deduced primary structure of dystrophin. We use the amino acid sequence to define different domains of the protein and to predict a preliminary structure by comparing dystrophin with spectrin and α-actinin.

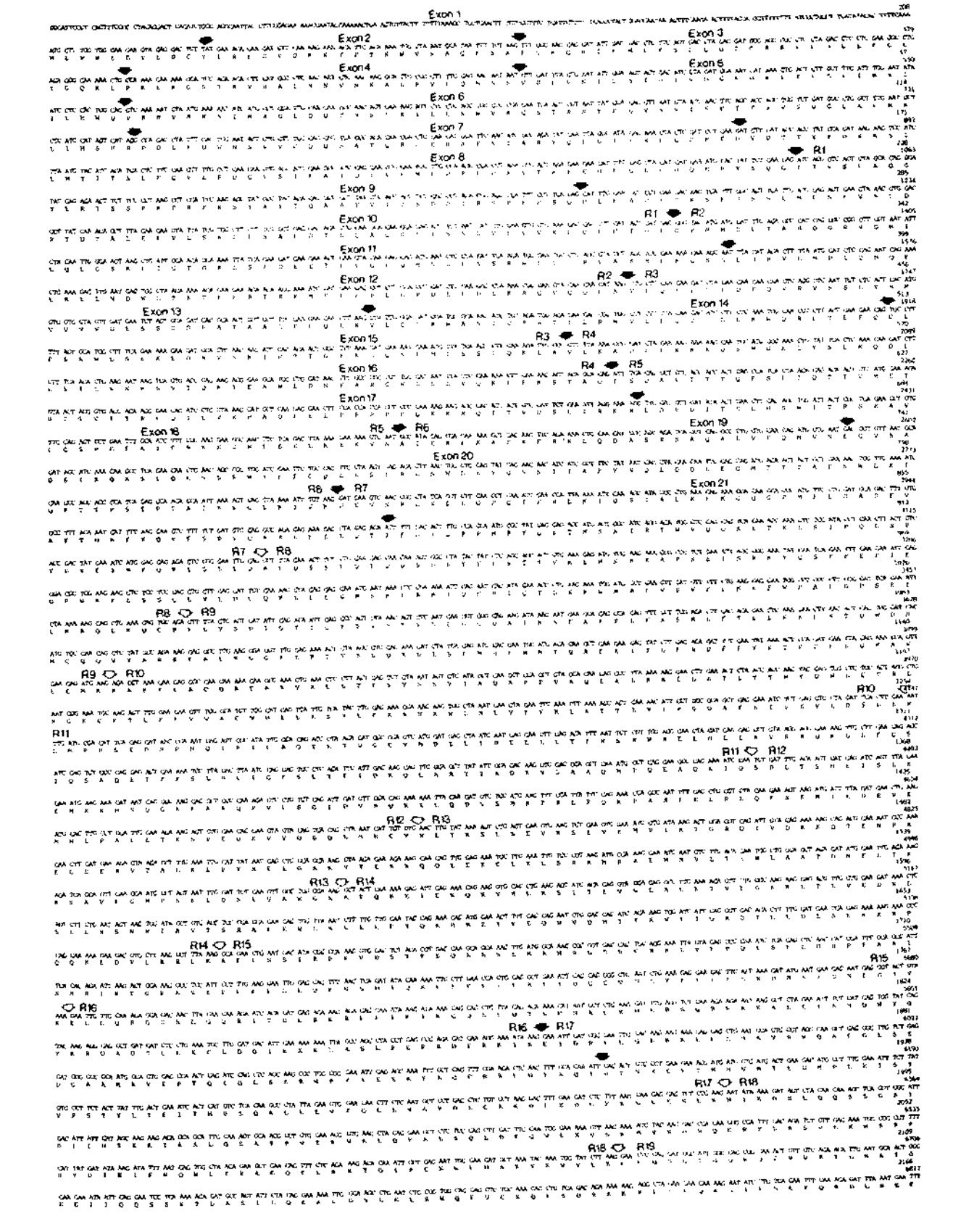
Results

Complete Nucleotide Sequence of the DMD cDNA and Preliminary Exon Organization

The entire sequence of the human DMD cDNA is presented in Figure 1. The 14 kb sequence is derived from 22 independent overlapping cDNA clones (Koenig et al., 1987). For the majority of the sequence, the sequences of the two complementary strands were determined from two different clones. Four cDNA clones appeared to differ from the other cDNA sequence by restriction mapping and/or sequence determination. In two cases, sequence divergence occurred at sites perfectly matching the consensus splice sites (Mount, 1982). The diverging sequences have stop codons in all three reading frames and are presumed to correspond to intron sequences. The presumed intron sequences were found in both cases on the 3′ end of the cDNA clones, and in one case a full 1 kb intron was included in the cDNA clone followed by an exon and the beginning of the next intron. These two cDNA clones probably represent hnRNA forms of the DMD transcript with the first introns of the gene already excised. The finding of cDNAs representing hnRNA forms may be correlated to the complexity of the DMD gene, which is formed by over 60 exons (Koenig et al., 1987). The two other cDNA clones showed minor structural differences that seemed to be artifacts of cloning rather than alternative splicing.

In addition to completing the cDNA sequence, the sequence of the intron–exon boundaries were determined for exons mapping from the 5′ end of the gene to the end of the DSX164 locus (see Experimental Procedures). The limits of the first 21 exons were determined, and they are presented in Figure 1. The sequence of the cloned

Cell
220


Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Exon 6
Exon 7
Exon 8
Exon 9
Exon 10
Exon 11
Exon 12
Exon 13
Exon 14
Exon 15
Exon 16
Exon 17
Exon 18
Exon 19
Exon 20
Exon 21
R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19

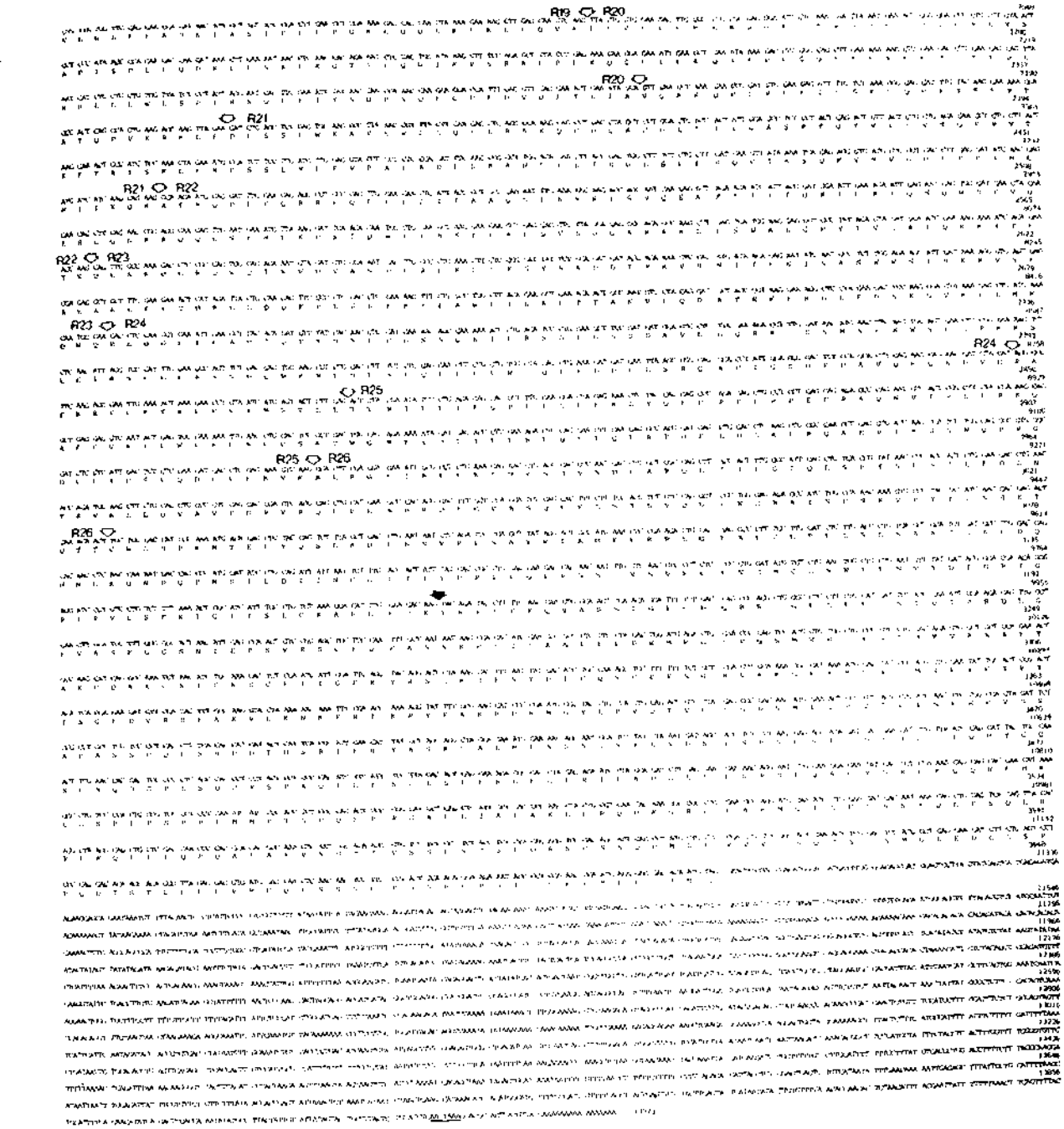


Figure 1. Nucleotide Sequence of the DMD Transcript and Deduced Primary Structure of the Encoded Protein

The DNA sequence was determined as described in Experimental Procedures and used to predict the protein sequence. The exon borders, where known, are indicated by black arrows, and the predicted repeat borders are indicated by light arrows. The three exon borders that are beyond nucleotide 5000 have been determined from the sequence of abnormal cDNA (see Results). The exons are numbered in the middle of the exon (Exon #), and the repeats are numbered at the beginning and at the end of the repeats (R #). The polyadenylation consensus sequence (Proudfoot and Brownlee, 1976) found at the end of the transcript is underlined. The GenBank accession number for this sequence is M18533.

genomic exons matched that found for the cDNA sequence with one exception. The published 5′ end of the cDNA GGGATTCCC (Koenig et al., 1987) was found to be GGGATCAC in the corresponding genomic sequence.

The Amino Acid Sequence of the DMD Gene Product

The DNA sequence was examined for presence of open reading frames consistent with the partial open reading frame published in Hoffman et al. (1987a) and Koenig et al.

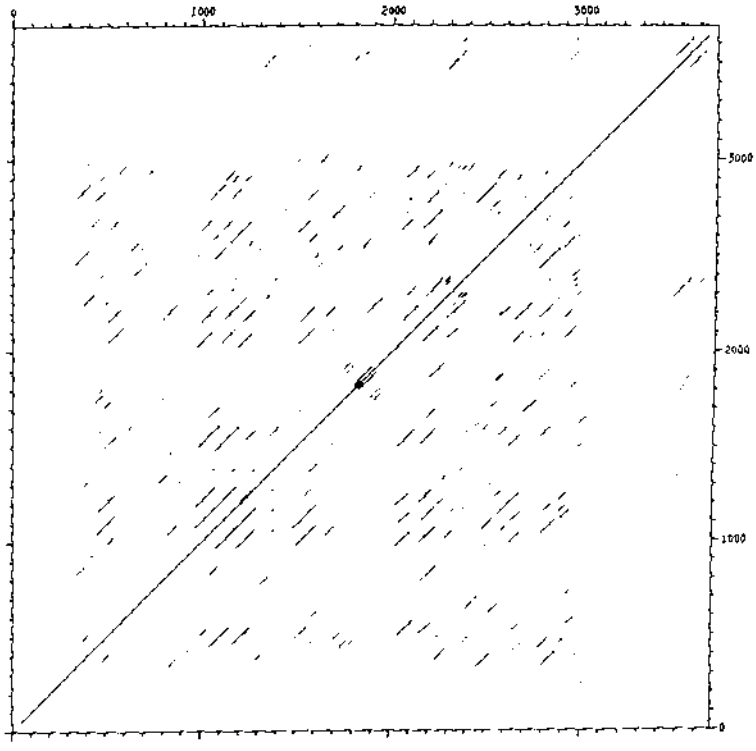


Figure 2. Dot Plot of the Dystrophin Amino Acid Sequence

The predicted amino acid sequence for the long open reading frame was analyzed for similar sequences within itself by the computer program COMPARE with a stringency of 39 and a window length of 100 (see Experimental Procedures). The output of this comparison was plotted by DOTPLOT (see Experimental Procedures) and is shown. The short lines parallel to the central diagonal represent internal similarities. The axes are labeled in residue numbers.

(1987). There is a single very large open reading frame (Figure 1) that follows the ATG initiation codon (Koenig et al., 1987) and extends over 11 kb of the cDNA. The stop codon of this open reading frame is followed by a 2.7 kb segment with no apparent coding potential. The open reading frame encodes a protein 3685 amino acids long, with a predicted molecular weight of 427 kd, in accordance with the reported 400 kd size of dystrophin, the protein product of the DMD gene (Hoffman et al., 1987b). The hydropathicity index calculated according to Kyte and Doolittle (1982) reveals that dystrophin is predominantly hydrophilic throughout the entire molecule. Thirty-one percent of amino acids are charged (Asp, Glu, His, Lys, and Arg). No substantial hydrophobic stretch of amino acids was found that could account for a signal peptide segment or a membrane spanning segment. The Chou and Fasman (1978) prediction of secondary structure reveals a very high potential for α-helical formation over the majority of the sequence. These α-helices, however, are frequently interrupted by prolines. The analysis of the cysteine distribution along the molecule reveals that the segment from amino acid positions 2200 to 3080 is completely devoid of cysteines; this segment is immediately followed by a relatively cysteine-rich segment (15 cysteines in 280 amino acids). The latter segment will be referred to as the "cysteine-rich" segment in the text that follows. The entire DMD cDNA and protein sequences were used to search for similarity with other sequences available from the GenBank nucleic acid data base and the PIR protein data base. Only the previously reported strong similarity to the nonmuscle α-actinin of chicken (Hammond, 1987) was found to be statistically significant. Dystrophin and α-actinin share the greatest similarity over a putative actin-binding N-terminal region that corresponds to about amino acids 14 to 240 of dystrophin (Hammond, 1987, and data not shown). This domain covers exons 2 to 8 while the first exon comprises the entire 5′ untranslated sequence and encodes for the first 11 amino acids of the protein.

Repeated Domains in Dystrophin

Recently, Davison and Critchley (1987) described the occurrence of weak repeats similar to those found in spectrin and α-actinin in the published partial DMD sequences. With the complete protein sequence, we further analyzed the extent of these repeats along the molecule and further characterized their nature. The dot matrix plot of the entire amino acid sequence against itself revealed many small diagonals spaced by about 100 amino acids and extending approximately from positions 300 to 3000 (Figure 2). The dot matrix gave a rough determination of the limits of the repeats, which were then aligned 2 by 2 with BESTFIT (UWGCG). The similarity between most repeats ranged from 10% to 25% identical matches. Positioning the determined exon borders on the aligned repeats shows that the 3′ splice site of exons 10, 12, 15, 18, and 20 correspond to the same position within the repeat. In addition, all five splice sites occur at the end of the coding triplet. We assumed that these exon borders corresponded to the limits of a genetic unit that was duplicated, and we used them to define the limits of the first six repeats. With these limits, repeats 1, 2, 5, and 6 are formed by two exons while repeat 3 is formed by three exons. Repeat 4 is a truncated repeat

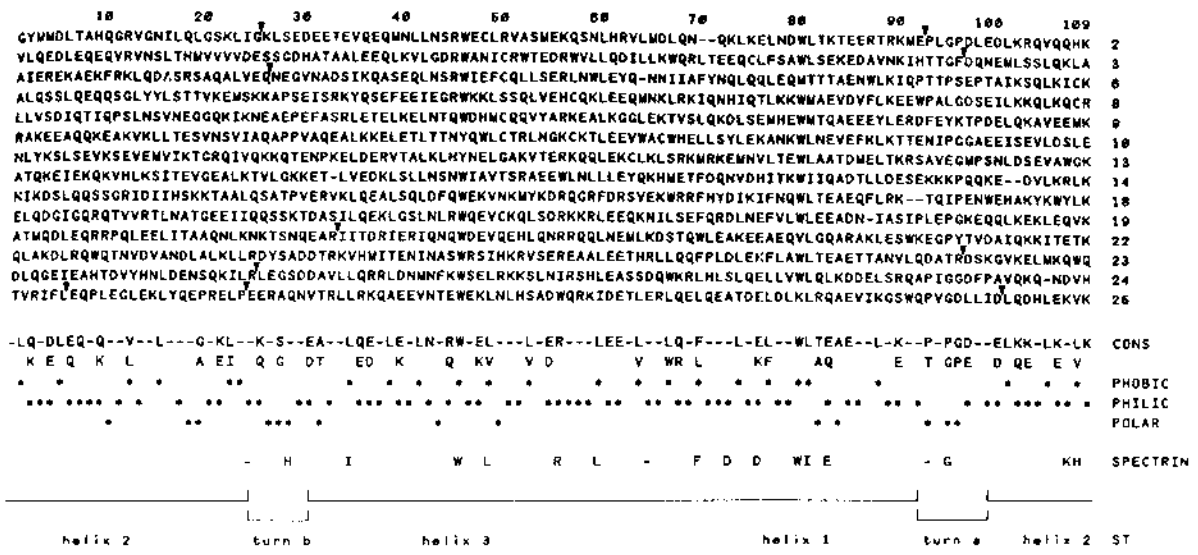


Figure 3. Table of 14 Dystrophin Repeats and Alignment with the Spectrin Repeat Consensus

The optimal alignment of the 14 repeats that exhibit the most internal similarities on the dot plot of Figure 2 is shown. The repeat number is given to the right. Dashes and dark arrowheads indicate insertions and deletions, respectively, which allow optimized alignment. Arrowheads represent insertions of one or two residues with the exception of the six-residue insertion of repeat 23. The beginning and the end of repeats 2, 3, and 6 correspond to established exon borders. CONS: the consensus sequence derived from the 14 repeats is based on the presence of one or two amino acids at a given position within a repeat ≥50% of the time (≥7/14). PHOBIC, PHILIC, and POLAR: a broader consensus was derived (as in CONS) with the amino acids grouped into three classes: PHOBIC, the large hydrophobic and aromatic residues (Leu, Ile, Val, Met, Phe, Tyr, and Trp); PHILIC, the charged and highly hydrophilic residues (Asp, Glu, Asn, Gln, His, Lys, and Arg); POLAR, the small polar residues (Gly, Ala, Ser, Thr, Cys, and Pro). SPECTRIN, the 15 best conserved positions of the spectrin consensus (according to Speicher and Marchesi, 1984) have been aligned with the dystrophin consensus. The three dashes indicate the approximate position of the gaps that should be inserted to align the spectrin consensus. ST: predicted secondary structure of the repeats. It is interesting to notice that the gaps introduced to align the dystrophin repeats as well as the spectrin repeat consensus occur predominantly in or near both turn regions.

(60 amino acids) and is formed by only one exon. The exon borders that fall in the middle of a repeat are not found at similar positions within the repeat. The limits of the remaining repeats can be predicted from their alignment with repeats 1 to 6, and their start and end points are indicated upon the sequence presented in Figure 1. The alignment for the less conserved repeats is still tentative.

The 14 most similar repeats were used to derive a consensus sequence (Figure 3). The consensus sequence is 109 amino acids long with one or two predominantly occurring amino acids (present in 7 or more repeats out of 14) assignable for 60 positions. The most striking feature of the consensus is the conservation of the tryptophan residues at positions 46 and 80. There is a clustering of proline and glycine between positions 28–29 and positions 92–98, indicative of the occurrence of turns in these two parts of the repeat. A less stringent consensus could also be derived by grouping the amino acids into three classes: the hydrophobic amino acids, the small polar amino acids, and the charged and related amino acids (see legend to Figure 3). An amino acid class could be assigned for 99 positions. This grouping reveals the alternation of one hydrophobic position with two or three hydrophilic positions, indicative of heptad repeats (McLachlan and Karn, 1982), extending between the two potential turn positions (see bottom of Figure 3). The occurrence of multiple heptad repeats (noticed in Hoffman et al., 1987a) strongly predicts that these regions adopt an α-helical conformation and indicates that the α-helices may be involved in coiled-coil interactions (McLachlan and Karn, 1982).

The entire dystrophin coding sequence contains 26 repeat domains organized in tandem. They extend from the end of the NH2 "actin-binding" domain to the beginning of the cysteine-rich domain (a schematic of the organization of the domains and repeats of dystrophin is presented in Figure 4). The length of the repeats ranges between 88 and 126 amino acids. Two segments (between repeats 20–21, 33 amino acids, and between repeats 25–26, 19 amino acids) were too short to be aligned with other repeats. The degree of similarity is also variable from one repeat to the other, as can be seen from the dot matrix comparison.

Analysis of the C-Terminal Domain of Dystrophin

Beyond the repeat domains, dystrophin is formed by a relatively cysteine-rich segment extending from amino acid position 3080 to 3360 followed by 325 additional amino acids with only one cysteine residue. To further investigate the extent of similarity between dystrophin and the α-actinins, the C-terminus of both proteins were compared (the C-terminal sequence of spectrin was not available for comparison). A significant similarity is found between the entire C-terminal domain of the slime mold (Dictyostelium discoideum) α-actinin, which contains two potential Ca^{2+}-binding sites (Noegel et al., 1987), and 150 amino acids

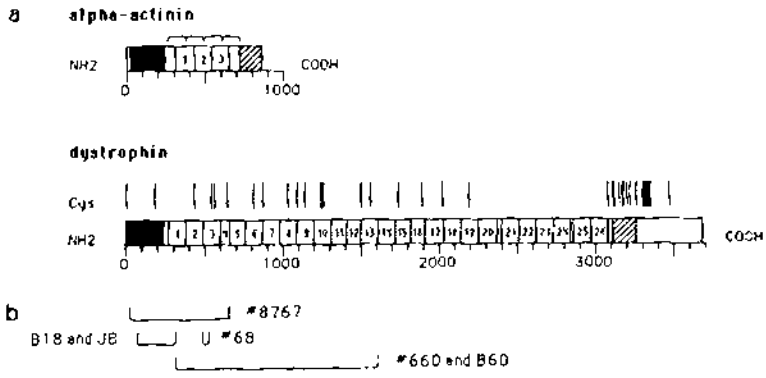


Figure 4 Domain Organization of Dystrophin and α-Actinin

(a) Landmarks of the predicted amino acid structure of dystrophin and the slime mold α-actinin are diagrammatically represented. Black box: putative actin-binding domain. Numbered open boxes: repeat units. The organization of the four triple helical domains (see Discussion) of α-actinin is represented by the brackets on top of the repeat boxes. Streaked box: putative EF hand Ca^{2+}-binding sites of α-actinin and homologous segment of dystrophin. The position of the cysteine residues of dystrophin is indicated by vertical bars on top of the dystrophin diagram. The scale in amino acid residues is given for both proteins.

(b) The extent of the internal protein deletions of the six Becker patients discussed in the text is shown along the dystrophin diagram. Patient #8767 is missing exons 3 to 16, patients JB and B18 are missing exons 5 to 9, patient #68 is missing exon 13, and the deletions of patients #660 and B60 extend from exon 10 (included) to beyond exon 21.

of the dystrophin cysteine-rich domain (Figures 4 and 5). Two gaps were introduced to optimize alignment, and there are 24% identical matches over 142 aligned amino acids (Figure 5). For comparison, the similarity between the C-terminal domains of the Dictyostelium α-actinin and the chicken fibroblastic α-actinin (Baron et al., 1987) is 31% of identical matches (Figure 5). In contrast, dystrophin and the chicken α-actinin are much less conserved over that segment (13% identical matches), suggesting that the Dictyostelium α-actinin segment may represent an ancestral sequence from which both dystrophin and the chicken α-actinin have evolved. The dystrophin segment does not match well with the EF-hand Ca^{2+}-binding site consensus (Tufty and Kretsinger, 1975), but might still have conserved the loop structure of the EF-hand.

The last 420 amino acids of dystrophin did not exhibit any similarity to α-actinin. This domain was used again to search for similarities in the GenBank (release 54) and PIR (release 14) data bases. The use of a limited section of dystrophin lowers the level of nonspecific matches and should allow the detection of less conserved sequences. Lowering the level of nonspecific matches did not reveal any new similarity.

Comparison of Domain Deletions with the Dystrophic Phenotype

Analysis of deletions of limited portions of dystrophin may give some insight into the functional significance of the various domains of the protein. Ideally, the high incidence of deletions of the DMD gene in independent patients should facilitate this study of domain function. However, only the deletions that do not result in a frame shift of the coding sequence represent true internal protein deletions. Some, if not most, of these internal protein deletions result in the less severe Becker muscular dystrophy phenotype, while deletions that disrupt the reading frame usually result in the Duchenne muscular dystrophy phenotype (Monaco et al., 1988). We discuss here the five Becker patients with genomic deletions that have been described in Monaco et al. (1988) with the addition of another Becker patient, #8767, whose deletion was first detected by Dr. B. deMartinville (University of Minnesota, Minneapolis) and further analyzed in our laboratory (see Experimental Procedures). The extent of the deletions is shown along the dystrophin diagram in Figure 4. The deletions of four of these patients have been precisely defined and the two exons flanking each of the deletions were shown to maintain an open reading frame. Because the reading frame is uninterrupted, one would predict a lower molecular weight, presumably semifunctional, dystrophin protein to be produced.

Three Becker patients have deletions that remove only sequences coding for repeat domains. The first patient (#68) has little functional deficits at age 17 and is missing only the first third of repeat 3. The two other Becker patients (#660 and B60) have deletions that remove repeats 1 to 12 or 13. The clinical progression of patients 660 and B60 is very different; one has been confined to a wheelchair since age 16 and the other can still walk at age 25. A more precise definition of the distal breakpoint of the two deletions will allow us to address the molecular basis of the phenotypic differences. Three Becker patients have deletions involving the N-terminal "actin-binding" domain of dystrophin. Two unrelated patients (B18 and JB) have similar deletions, with the same exons deleted, resulting in a loss of the last two-thirds of the "actin-binding" domain, and they both showed significant muscle weakness before age 16 (see Experimental Procedures). Finally, the dystrophin of patient #8767 is missing repeats 1, 2, 3, and 4 but also almost the complete "actin-binding" domain (with the exception of the first 18 amino acids of that domain). The deletion of patient #8767 encompasses the deletion of patients B18 and JB, yet his clinical evolution is significantly milder, since he can still walk at age 29.

```
DMD  3116 LQKALCLDLLSLSAACDALDQHNLKQNDQPQDILQIINCLTTIYDRLEQEHNNLVNVPLCVDMCLNWLLNVYDTGRTGRIRVLSFKTGIISLCKA-HL 3211
          * * *       .   *   *  *.**. .. *..  ** .* * * .*                  ** ..   **   * * .  *   ..*  *.
ADD  720 LAKQLTGVTAEELSEFKACFSHFDKDNDNKLNRLEFSSCLKSIGDELTEE-------------QLNQVISKIDIDGNGTISFEEFIDYMVSSRKGTDS 804
          *...  *.. *..  **.*.* ***.      ..   * .** *.* ...*                .. ..*  *..  * ..*  ***.*      **.
ACHK 738 LTRDAKGISQEQMNEFRASFNHFDRKKTGMMDCEDFRACLISMGYNMGEA-------------EFARIMSIVDPNRMGVVTFQAFIDFMSRETADTDT 828

                          EL' LL' LD'D'DG ID' EL' LL' L'                          EL' LL' LD'D'DG ID' EL' LL' L'
                              X Y Z-Y-X -Z                                            X Y Z-Y-X -Z


DMD  3212 EDKYRYLFKQVASSTGFCDQRRLGLLLHDSIQIPRQLGEVA-----SFGGSNIEPSVRSCFQF      3289
          .  .  ** .*    *  .  .  . ** **   *. ..     .. *.   *.  .
ADD  805 VESTKAAFKVMAEQKQFITEAQIRAAISDSKQIDYLLASMP-----AVEGGFDYNSFAEKLYQ      862  COOH END
          .  *.**..* **..**  ..*  ..       .    *    ** *..** **.  **
ACHK 821 ADQVMASFKILAGDKNYITVDELRRELPPDQAEYCIARMAPYNGRDAVPGALDYMSFSTALYGESDL 887  COOH END
```

Figure 5. Similarity between the Two Potential EF-Hand Ca^{2+}-Binding Sites of the Slime Mold α-Actinin and a Corresponding Segment of Dystrophin

The sequences were aligned with the aid of the computer program BESTFIT (UWGCG, see Experimental Procedures). The alignment of the slime mold α-actinin (Noegel et al., 1987) with the chicken α-actinin (Baron et al., 1987) is shown for comparison. Stars indicate identical matches and points indicate conservative matches. DMD, dystrophin; ADD, α-actinin of Dictyostelium discoideum; ACHK, chicken α-actinin. Below the chicken α-actinin sequence are the EF-hand consensus (Tufty and Kretsinger, 1975; D', oxygen-containing side chain; L', hydrophobic side chain), and the positions of the Ca^{2+}-chelating side chains (X, Y, Z,-X,-Z).

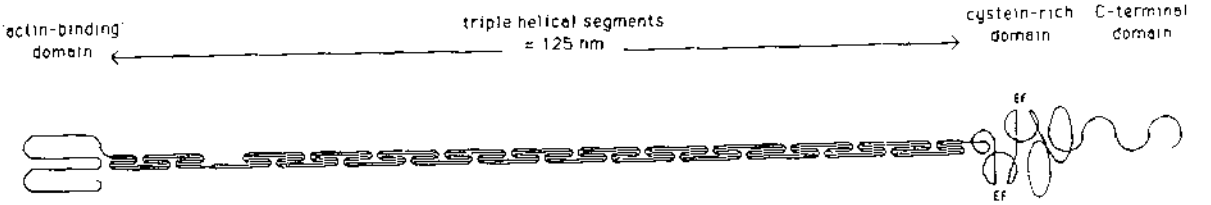


Figure 6. A Model of the Dystrophin Molecule

Four domains of dystrophin are represented on this tentative schematic. The conformation of the "actin-binding," cysteine-rich, and C-terminal domains is purely speculative, and they are represented as broad loops simply to illustrate that they may adopt a more globular structure than the rod section. The rod section is formed by 25 triple helical segments. A triple helical segment is formed by the last half of a repeat and the first half of the following repeat. The fourth triple helical segment is only formed by the first half of repeat 5 since repeat 4 is truncated (see text). The 20th and 24th triple helical segments are represented longer simply to indicate the existence of short segments between repeats 20 and 21 and between repeats 24 and 25. EF indicates two loops that might have conserved an EF-hand conformation (see text).

Discussion

Predicted Structure of Dystrophin

Dystrophin appears to be formed by four different domains: an N-terminal "actin-binding" domain (Hammond, 1987), a middle domain formed by 26 repeats in tandem, a cystein-rich domain, and a less characterized COOH domain. The first three domains show significant similarity to the three domains of α-actinin while the C-terminal domain did not reveal any similarity to other known proteins. The repeats of the second domain were shown to be similar to the repeats found in spectrin and α-actinin, and in particular, the two characteristic tryptophan positions are found in all three types of repeat (Davison and Critchley, 1988). The 109 amino acid repeat consensus of dystrophin can be aligned with the 106 amino acid repeat consensus of spectrin (Speicher and Marchesi, 1984) by the insertion of three single amino acid gaps (see bottom of Figure 3), while the 120 amino acid repeat consensus of α-actinin (Baron et al., 1987) seems more distantly related. We found that the two clusters of proline and glycine and nearly all hydrophobic positions of the dystrophin consensus are also conserved in the consensus of spectrin and α-actinin (data not shown). The predicted secondary structure of the dystrophin repeats is similar to the predicted structure of the spectrin repeats (Speicher and Marchesi, 1984), with the exception of the α-helix extending from position 31 to 92 that was predicted to be interrupted in spectrin by a median "connecting region." A triple helical model has been proposed for the spectrin repeats (Speicher et al., 1983; Speicher and Marchesi, 1984) based on sequence data. In addition, the length of a repeat unit (about 5 nm, as deduced from electron microscope measurement) is in good agreement with the triple helical model since this length is a third of the length of a single α-helix with the same number of amino acids (Speicher et al., 1983). This model is likely to be valid for the dystrophin and α-actinin repeats. The model predicts that a triple helical segment is formed by the juxtaposition of the three α-helices 1, 2, and 3 that are connected by turns a and b (see bottom of Figure 3). The dystrophin triple helical segment would then be formed by the last half of a repeat and the first half of the following repeat. In agreement with the model, the half-repeats that do not interact with another half-repeat (first half of repeat 1, last half of repeat 26, and the half-repeat following the truncated repeat 4) show very poor similarity to the repeat consensus.

The middle domain of dystrophin probably adopts a rod shape, as does the middle domain of α-actinin and spec-

trin. The length of a single spectrin triple helical segment has been calculated to be approximately 5 nm (Speicher et al., 1983). Given that the middle domain of dystrophin is formed by 25 triple helical segments, the predicted length of this domain is around 125 nm. A model emphasizing important structural features of the dystrophin molecule is given in Figure 6. The similarity of dystrophin with two distinct proteins that self-assemble as antiparallel dimers (homodimers for α-actinin and heterodimers for spectrin; Wallraff et al., 1986; Speicher et al., 1983) opens the question of the possible dimerization of dystrophin.

Dystrophin was recently demonstrated to be associated with membrane structures (Hoffman et al., 1987c). The primary structure of the protein does not pinpoint the domain of dystrophin that mediates the attachment to membranes. Dystrophin is probably not directly embedded through the membrane, but possibly binds to membranes via an ankyrin-like or protein 4.1-like molecule, in a way similar to the spectrin membrane binding (Bennett and Stenbuck, 1979; Morrow et al., 1980). The C-terminus of dystrophin may be the domain that mediates the attachment to membranes, and it will be interesting to see if this domain shares any similarity with the corresponding domain of β-spectrin, known to contain the ankyrin-binding site (Morrow et al., 1980), or with other cytoskeletal proteins as these protein sequences are entered into the data bases.

The subcellular localization of dystrophin with the transverse tubules of the triadic structure of muscle cells (Hoffman et al., 1987c; Knudson et al., 1988) and the primary structure of the protein presented here suggests that dystrophin is a component of the membrane cytoskeleton. Dystrophin might interact with the myofibrillar actin filaments through the N-terminal actin-binding domain and would then be a link that could maintain the coupling between the triadic structures and the myofibrillar system during muscle contraction (Hoffman et al., 1987c). Alternatively, dystrophin might play a role in the morphogenesis of the T-tubules. Both hypotheses could account for the occurrence of disorganized T-tubule system networks in the muscle of affected patients (Oguchi et al., 1982; Watkins and Cullen, 1987).

Evolution of the Dystrophin Protein

The extensive similarity between dystrophin and α-actinin over three major domains (see Figure 4) strongly suggests that these proteins have evolved from a common ancestor. Both proteins form a new gene family that may include spectrin as well. The sequence of the C-terminus of the β-spectrin is needed to define more precisely the structural and functional relationship between dystrophin and spectrin. The many repeat domains within dystrophin probably arose by multiple steps of domain duplication. It is tempting to speculate that the unit of duplication corresponds to the block of exon whose splice points have been conserved at the beginning and end of a repeat. The unit of duplication does not seem be the triple helical segment itself, but the last half of a segment plus the first half of the following segment. This unit allows the duplication of the predicted long α-helix 1 plus 3 (see Figure 3) as a single block. The low level of similarity between individual repeats hampers the finding of a higher order of repetition in correlation with the various steps of duplication. Details about the steps of duplication may arise when more about the exon structure of the gene is known.

Functional Significance of Internal Protein Deletions in Becker Patients

Deletions of internal domains represented by a truncated protein in some if not most Becker patients should help decipher the role of each domain. It is particularly intriguing that a deletion of the majority of the putative actin-binding domain could result in a very mild Becker phenotype. This raises two questions: First, does the NH2 domain of dystrophin really have an actin-binding activity? Second, if it does, is the actin-binding activity necessary for dystrophin function? Alternatively, the dystrophic phenotype may be modulated by other genetic factors and/or environmental factors, as suggested in Forrest et al. (1987). Phenotypic differences between Becker patients with identical or similar internal domain deletions should help address the question of the influence of such factors in the progression of the muscular dystrophy. In addition, the direct analysis of dystrophin in the muscle of affected patients (Hoffman et al., 1988) would help define the influence exerted by the level of expression and protein stability on the phenotype. The functional significance of each domain can also be tested in vitro by the expression of altered sequences of the DMD cDNA in prokaryotic or eukaryotic expression vectors. The combination of both the in vivo and in vitro approaches will give new insights into the role of dystrophin in the muscle cytoskeletal network.

Experimental Procedures

cDNA Sequencing

The cDNA sequences were determined by the dideoxynucleotide method (Sanger et al., 1977). The sequence of the first 3830 bp of the human DMD cDNA was previously reported (Koenig et al., 1987; Hoffman et al., 1987a; Cross et al., 1987). The sequence of the remaining 10 kb was first determined on one strand by using the various restriction sites and cDNA clone ends indicated on the cDNA map in Koenig et al. (1987). The sequencing was performed on double-stranded plasmid DNA purified on cesium gradient. The universal primers recommended by Stratagene for the Bluescript vectors were used for sequencing. The cDNA sequence was checked on the second strand with the use of specific oligonucleotide primers synthesized by the Howard Hughes Medical Institute Biopolymers Laboratory (Harvard Medical School, Boston) according to the first strand sequence. The primers were spaced by about 200–300 bp such that the second strand sequence was determined across all the restriction sites used in the first strand sequence determination.

Computer Analysis

The nucleotide sequence and predicted amino acid sequence were analyzed by the programs of the University of Wisconsin Genetics Computer Group (UWGCG, Devereux et al., 1984) and BIONET, Intelligenetics. The amino acid sequence was searched for internal repeats by the method of Maizel and Lenk (1981) using COMPARE and DOTPLOT of the UWGCG program. Sequences exhibiting similarities were further compared with the alignment method (BESTFIT program) of Smith and Waterman (1981). Secondary structure and hydropathicity were predicted by PEPPLOT (UWGCG), using the algorithms of Chou and Fasman (1978) and Kyte and Doolittle (1982), respectively. The PIR (National Biomedical Research Foundation) protein data base and the

GenBank nucleic acids data base were searched with the predicted dystrophin amino acid sequence and the corresponding nucleotide sequence, respectively, by the programs IFIND (BIONET) and WORDSEARCH (UWGCG). Both programs were adapted from the alogrithm of Wilbur and Lipman (1985).

Exon Border Determination

The cloning of the genomic segment of all the exons from the 5' end of the gene to the DXS164 locus has been completed with the exception of three exons: the two exons of the DXS206 locus (Burghes et al., 1987) and the exon contained in the 3.25 kb HindIII genomic fragment at the 5' end of the gene (Koenig et al., 1987). The exon flanking sequences were determined by sequencing the genomic clones with oligonucleotide primers synthesized by the Howard Hughes Medical Institute Biopolymers Laboratory according to the cDNA sequences. The exon boundaries were deduced by comparing the genomic and cDNA sequences and by comparing the genomic sequences with the splice sites consensus sequences (Mount, 1982). The boundaries of one of the two exons of the DXS206 locus has been published (Heilig et al., 1987), and the boundaries of the remaining two exons for which genomic clones were not available have been inferred from the flanking exons. We cannot exclude the possibility that one of these two inferred exons corresponds to two exons. The 21 exons are present in the following HindIII genomic fragments (size in kb; Koenig et al., 1987): exon 1, 3.2; exon 2, 3.25; exon 3, 4.2; exon 4, 8.5; exon 5, 3.1; exon 6, 8.0; exon 7, 4.6; exons 8 and 9, 7.5; exons 10 and 11, 10.5; exon 12, 4.2; exon 13, 6.6; exons 14 and 15, 2.7; exon 16, 6.0; exon 17, 1.7; exon 18, 12; exon 19, 3.0; exon 20, 7.3; exon 21, 11.

Clinical Information of BMD Patients and Deletion Analysis

Patient #8767 was initially diagnosed as DMD due to an early onset of muscle weakness. The fact that he still walks at age 29 and had an uncle diagnosed as BMD who became wheelchair bound at age 49 indicates that he, too, has BMD. The five other BMD patients (JB, B18, #68, #660, and B60) are described in detail in Monaco et al. (1988). Their clinical history is briefly summarized here. JB has been using a wheelchair since age 13; at age 19, he can, with difficulty, maneuver from the chair. JB had an uncle with BMD who died at age 38. B18 is not confined to a wheelchair at age 16, but does require assistance climbing stairs and raising from the floor. #68 has few functional deficits at age 17, although he shows mild proximal muscle weakness. #660 has been confined to a wheelchair since age 16 and is stable at age 21. B60 still walks at age 25 but he has difficulty climbing stairs. The extent of the genomic deletions was analyzed by Southern blot hybridization as detailed in Aldridge et al. (1984) with the DMD cDNA probes (Koenig et al., 1987).

Acknowledgments

We would like to thank B. deMartinville for sharing a DNA sample of a BMD patient, E. P. Hoffman and S. Harrison for discussions and critical reading of the manuscript, and C. J. Bertelson and C. Feener for technical assistance. L. M. K. is an associate investigator of the Howard Hughes Medical Institute; A. P. M. is supported in part by Public Health Service grant NRSA (2T 32 GM07753-07) from the National Institute for General Medical Sciences; M. K. and A. P. M. received support from the Muscular Dystrophy Association of America. This work was in part supported by the National Institutes of Health (#HD 18658 and NS23740) and by the Muscular Dystrophy Association of America.

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

Received March 4, 1988.

References

Baron, M. D., Davison, M. D., Jones, P., and Critchley, D. R. (1987). The sequence of chick alpha-actinin reveals homologies to spectrin and calmodulin. J. Biol. Chem. *36*, 17623–17629.

Bennett, V., and Stenbuck, P. (1979). Identification and partial purification of ankyrin, the high affinity membrane attachment site for human erythrocyte spectrin. J. Biol. Chem. *254*, 2533–2541.

Burghes, A. H. M., Logan, C., Hu, X., Belfall, B., Worton, R., and Ray, P. N. (1987). Isolation of a cDNA clone from the region of an X:21 translocation that breaks within the Duchenne/Becker muscular dystrophy gene. Nature *328*, 434–436.

Chou, P. W., and Fasman, G. D. (1978). Empirical prediction of protein conformation. Annu. Rev. Biochem. *47*, 251–276.

Cross, G. S., Speer, A., Rosenthal, A., Forrest, S. M., Smith, T. J., Edwards, Y., Flint, T., Hill, D., and Davies, K. E. (1987). Deletions of fetal and adult muscle cDNA in Duchenne and Becker muscular dystrophy patients. EMBO J. *6*, 3277–3283.

Davison, M. D., and Critchley, D. R. (1988). Alpha-actinins and the DMD protein contain spectrin-like repeats. Cell *52*, 159–160.

Devereux, J., Haeberli, P., and Smithies, O. (1984). A comprehensive set of sequence programs for the VAX. Nucl. Acids Res. *12*, 387–395.

Forrest, S. M., Cross, G. S., Speer, A., Gardner-Medwin, D., Burn, J., and Davies, K. E. (1987). Preferential deletion of exons in Duchenne and Becker muscular dystrophies. Nature *329*, 638–640.

Francke, U., Ochs, H. D., de Martinville, B., Giacalone, J., Lindgren, V., Distèche, C., Pagon, R. A., Hofker, M. H., van Ommen, G. J. B., Pearson, P. L., and Wedgwood, R. J. (1985). Minor Xp21 chromosome deletion in a male associated with expression of Duchenne muscular dystrophy, chronic granulomatous disease, retinitis pigmentosa and McLeod syndrome. Am. J. Hum. Genet. *37*, 250–267.

Hammond, R. G., Jr. (1987). Protein sequence of DMD gene is related to actin-binding domain of α-actinin. Cell *51*, 1.

Heilig, R., Lemaire, C., and Mandel, J. L. (1987). A 230 kb cosmid walk in the Duchenne muscular dystrophy gene: detection of a conserved sequence and of a possible deletion prone region. Nucl. Acids Res. *15*, 9129–9142.

Hoffman, E. P., Monaco, A. P., Feener, C. A., and Kunkel, L. M. (1987a). Conservation of the Duchenne muscular dystrophy gene in mice and humans. Science *238*, 347–350.

Hoffman, E. P., Brown, R. H., Jr., and Kunkel, L. M. (1987b). Dystrophin: the protein product of the Duchenne muscular dystrophy locus. Cell *51*, 919–928.

Hoffman, E. P., Knudson, C. M., Campbell, K. P., and Kunkel, L. M. (1987c). Subcellular fractionation of dystrophin to the triads of skeletal muscle. Nature *330*, 754–757.

Hoffman, E. P., et al. (1988). Dystrophin characterization in muscle biopsies from Duchenne and Becker muscular dystrophy patients. New Engl. J. Med., in press.

Knudson, C. M., Hoffman, E. P., Kahl, S. D., Kunkel, L. M., and Campbell, K. P. (1988). Characterization of dystrophin in skeletal muscle triads. J. Biol. Chem., in press.

Koenig, M., Hoffman, E. P., Bertelson, C. J., Monaco, A. P., Feener, C., and Kunkel, L. M. (1987). Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. Cell *50*, 509–517.

Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.

Lev, A., Feener, C., Kunkel, L. M., and Brown, R. H. (1987). Expression of the Duchenne muscular dystrophy gene in cultured muscle cells. J. Biol. Chem. *262*, 15817–15819.

Maizel, J. V., and Lenk, R. P. (1981). Enhanced graphic matrix analysis of nucleic acid and protein sequences. Proc. Natl. Acad. Sci. USA *78*, 7665–7669.

McLachlan, A. D., and Karn, J. (1982). Periodic charge distributions in the myosin rod amino acid sequence match cross-bridge spacing in muscle. Nature *299*, 226–233.

Monaco, A. P., Neve, R. L., Colletti-Feener, C., Bertelson, C. J., Kurnit, D. M., and Kunkel, L. M. (1986). Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. Nature *323*, 646–650.

Monaco, A. P., Bertelson, C. J., Liechti-Gallati, S., Moser, H., and Kunkel, L. M. (1988). An explanation for the phenotypic differences between patients bearing partial deletions of the DMD locus. Genomics, in press.

Morrow, J. S., Speicher, D. W., Knowles, W. K., Hsu, J. S., and Marchesi, V. T. (1980). Identification of functional domains in human erythrocyte spectrin. Proc. Natl. Acad. Sci. USA *77*, 6592–6596.

Moser, H. (1984). Duchenne muscular dystrophy: pathogenic aspects and genetic prevention. Hum. Genet. *66*, 17–40.

Mount, S. M. (1983). A catalogue of splice junction sequences. Nucl. Acids Res. *10*, 459–472.

Noegel, A., Witke, W., and Schleicher, M. (1987). Calcium-sensitive non-muscle alpha-actinin EF-hand structures and highly conserved regions. FEBS Lett. *221*, 391–396.

Oguchi, K., Yanagisawa, N., and Tsukagoshi, H. (1982). The structure of the T-system in human muscular dystrophy. J. Neurol. Sci. *57*, 333–341.

Proudfoot, N. J., and Brownlee, G. G. (1976). 3′ non-coding region sequences in eukaryotic mRNA. Nature *263*, 211–214.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA *74*, 5463–5467.

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. J. Mol. Biol. *147*, 195–197.

Speicher, D. W., and Marchesi, V. T. (1984). Erythrocyte spectrin is comprised of many homologous triple helical segments. Nature *311*, 177–180.

Speicher, D. W., David, G., Yurchenco, P. D., and Marchesi, V. T. (1983). Structure of human erythrocyte spectrin. J. Biol. Chem. *258*, 14931–14947.

Tufty, R. M., and Kretsinger, R. H. (1975). Troponin and parvalbumin calcium binding regions predicted in myosin light chain and T4 lysozyme. Science *185*, 167–169.

Verellen-Dumoulin, C., Freund, M., de Meyer, R., Laterre, C., Frederic, J., Thompson, M. W., Markovic, V. D., and Worton, R. G. (1984). Expression of an X-linked muscular dystrophy in a female due to translocation involving Xp21 and non-random X-inactivation. Hum. Genet. *67*, 115–119.

Wallraff, E., Schleicher, M., Modersitzki, M., Rieger, D., Isenberg, G., and Gerisch, G. (1986). Selection of *Dictyostelium* mutants defective in cytoskeletal proteins; use of an antibody that binds to the ends of alpha-actinin rods. EMBO J. *5*, 61–67.

Watkins, S. C., and Cullen, M. J. (1987). A qualitative and quantitative study of the ultrastructure of regenerating muscle fibres in Duchenne muscular dystrophy and polymyositis. J. Neurol. Sci. *82*, 181–192.

Wilbur, W. J., and Lipman, D. J. (1985). Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. USA *80*, 726–730.