The EMBO Journal vol.7 no.13 pp.4157–4162, 1988

The chicken dystrophin cDNA: striking conservation of the C-terminal coding and 3′ untranslated regions between man and chicken

C.Lemaire, R.Heilig and J.L.Mandel

Laboratoire de Génétique Moléculaire des Eucaryotes du CNRS, Unité 184 de Biologie Moléculaire et de Génie Génétique de l'INSERM, Faculté de Médecine, 11, rue Humann, 67085 Strasbourg Cédex, France

Communicated by J.L.Mandel

Dystrophin is a very large muscle protein (~400 kd) the deficiency of which is responsible for Duchenne muscular dystrophy. Its function is unknown at present. In order to know whether different domains of the protein are differentially conserved during evolution, we have cloned and sequenced the chicken dystrophin cDNA. The protein coding sequence has almost the same size as in man. The N-terminal region that resembles the actin binding domain of α actinin, as well as the large spectrin like domain show 80% and 75% conservation respectively between chicken and man. In contrast, the C-terminal region shows 95% identity over 627 aa suggesting that it is an important region of interaction with other proteins. Comparison of the amino acid sequence of this C-terminal region to other protein sequences shows only marginally significant similarities. Finally we have found a striking conservation of three segments of the 3′ untranslated sequence (85% homology over a total of 920 nt) between chicken and man. These also appear to be conserved in other mammals. This high conservation is not linked to open reading frames.

Key words: cDNA/chicken/dystrophin/evolution/3′-untranslated

Introduction

The Duchenne muscular dystrophy gene, which codes for a protein recently named dystrophin, has many exceptional features. It is by far the largest gene known, with a very high mutation rate in man and its protein product is also very large (400 kd) (Hoffman *et al.*, 1987a; Koenig *et al.*, 1987, 1988; Burmeister *et al.*, 1988). The protein appears localized within the muscle cell, at the internal face of the plasma membrane (Watkins and Cullen, 1987; Arahata *et al.*, 1988; Zubrzycka-Gaarn *et al.*, 1988). It contains a domain of homology to the actin binding domain of α actinin, and a very large domain (2796 aa) that shows evolutionary remnants of tandem repeats of a 109 aa sequence, with some homology to spectrin (Hammond, 1987; Davison and Critchley, 1988; Koenig *et al.*, 1988). These observations suggest that dystrophin may play a role in anchoring the cytoskeleton to the plasma membrane. The determination of structural and functional domains of such a long and complex protein is an important goal. Analysis of the conservation of the protein sequence during evolution may be useful for directing attention to regions of potential functional importance. It has been reported previously that the N-terminal region of the protein shows 87% homology over 700 aa between man and mouse (Hoffman *et al.*, 1987b). Comparison over a longer evolutionary period might be better to detect differential selective pressure operating of various regions of the protein coding sequence. We report here the cloning and sequencing of chicken dystrophin cDNA and show that the C-terminal region is much better conserved than the actinin-like N-terminal region or the spectrin-like central region. Furthermore a striking conservation of a large part of the 3′ untranslated region has been found.

Results

Cloning and sequencing of the chicken dystrophin cDNA

A chicken skeletal muscle cDNA library constructed in λgt10 according to Gubler and Hoffman (1983) was initially screened with a chicken genomic probe (CCR-P) isolated previously (Heilig *et al.*, 1987), that is homologous to exon 8 of the human dystrophin gene. Several walking steps allowed us to clone a 13 575 bp sequence that corresponds to 97% of the length of the human dystrophin cDNA sequence (Figure 1). For most of the length, the sequence was obtained from at least two independent clones, in order to avoid both sequencing and cloning artefacts. Two cDNA clones appear to correspond to incompletely spliced RNAs, and contained typical exon–intron junctions. This allows us to place exon limits at position 5310 (between aa 1720/1721) and 8013 (aa 2621/2622) of the chicken sequence, corresponding to 5362 (aa 1718/1719) and 8080 (aa 2624/2625) of the human sequence. It is likely that these positions also correspond to exon limits in the human gene, which has not yet been mapped in this region. They are exactly positioned at the limits of spectrin-like repeats predicted by Koenig *et al.* (1988) (between repeats 14 and 15, and 22 and 23 respectively), which further substantiate that these repeats arose through exon duplications.

An open reading frame (ORF) of 3660 codons can be aligned with the human sequence. The initiation codon in chicken is 4 codons upstream from its position in the human gene, while the position of the termination codon is conserved.

Conservation of the protein coding sequence

The comparison of the human and chicken protein sequence is shown in Figures 2 and 3. The first 3033 amino acids show 76% homology (80% over the actinin-like N-terminal end, and 75% over the large spectrin-like region), with only four insertions or deletions of 1 to 7 amino acids. Three of these appear clustered in a 350 aa region which shows the least conservation (57% homology). In contrast the 627 C-terminal amino acids, which include a cysteine rich region, show a very striking 95% conservation (and most of the

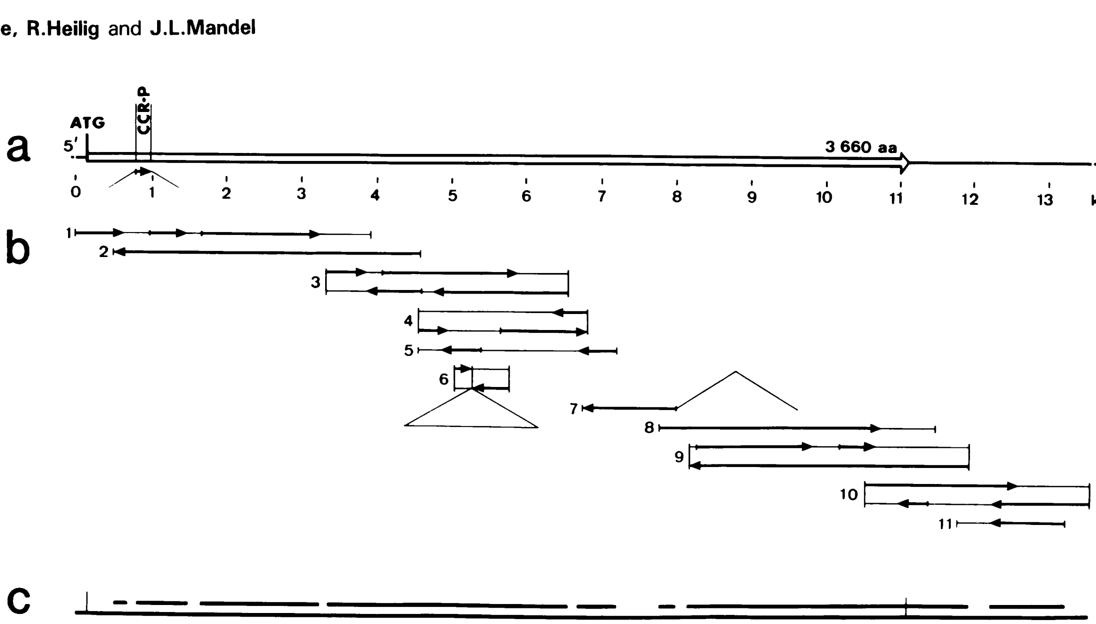


Fig. 1. Cloning and sequencing strategy. The probe that was used for initial screening is indicated (CCR). The open arrow in (**a**) corresponds to the protein coding sequence. The clones that were used for sequencing are designated 1–11. Clones 8–10 were screened with the human cDNA clone 63.1 (probe 9–14) (Koenig *et al.*, 1987) while all the others were derived by walking steps with the chicken clones. (**b**) Heavy arrows indicate extent and direction of sequencing. Position of intron sequences are indicated in clones 6 and 7. (**c**) The bottom lines summarize the extent of sequence that was obtained on at least two independent cDNA clones.

```
                                                                                                   100
MSAHVLWYEEVEDDYEREDVQKKTFTKWINAQFAKCGRRCIEDLFNDFRDGRKLLELLECLTGQKIAKEKGSTRVHALNNVNKALQILQRNNVDLVNIGSSDIVDGNHKLTLGLIWNIILHWQVKDVMKNIMAGLQQTNSEKILLSWVRQSTRNYPQVNV
----M  W     C              V    S F KQH  N  S LQ   R  D   G     LP                  RV   N         T                        N
                                      200                                                                                                 300
INFTSSWSDGLAFNALLHSHRPDLFDWNAVASQQSPVQRLDHAFNIARQHLGIEKLLDPEDVATACPDKKSILMYVTSLFQVLPQQVTMEAIREVEMLPRHSRVTTEEHIQVHHQQHFSQEITVNIPQ--RPSPSPKPRFKSYAYAQTAYVIPPDQKRRQ
    T       L   I           S  VC   AT  E        YQ           D TY         I            SI   Q       PPK K   F L   M Y  Q    SLA   GYERTS        T A    TTS PT SP
                                                                                400
VPPQFLETVEKRTYTTTVMRSEMDLDSYQTALEEVLTWLLSAEDALQAQGDISSDVEVVKEQFHTHEGFMMELTAHQGRVGNVLQVGSQLLAMGKLSDDEENEIQEQMNLLNSRWESLRVASMEKQSNLHKILMDLQNQQLAQLADWLTKTEERTKKIDS
F S H  AP DKSFGSSL E  VN  R         S       T     E  N      D       Y  D          I  L  K IGT    E   T V            C             RV       K KE N          R MEE
                    500                                                                                                 600
EPLGPDLEDLKRQVEEHKAFQDDLEQEQVKVNSLTHMVVVVDENSGDKATAALEEQLQHFGSRWAAICRWTEDRWVLLQDILRKWQHFAEEQCLFDAWLTEKEGSLSKIQTSDFKDENEMLTSLRKLAILKGDIEMKKQMMSKLKSLSRDLLVAVKNKAV
              QQ  VL E       R             S   H         KVL D   N                L   RLT      S   S   DAVN  H TG   Q    S  Q   V  A L K   S G  Y  KQ   STL   S
                                                            700
AQKLESRLENFAQRWDSLVQKLESDSKQVSQAVTTTQTSLTQTTVMETVTMVTTREQILVKHAKEELPPPPHKKRQLLVDSEIRKRFDSDTTELHSWMTRSEAVLQSPEFAIYRKEGNLSDLRERVNAIQREKPEKYRKLQDASRSAEALVEQMVNEGL
T  T AW D    RC N      KSTA I        P            T            Q        Q    IT        L V I      I              F     F   K K    E   A  F          Q         V
800                                                                                                 900
NADNIRQASEQLKSRWIEFCQLLSERLVWLEYQNSIIDFYSQLQRLEQTAITAENWLKAQPTPATDPATVKIQLEKCKDEIIRMSTLQPQIERLKAQSQALKEKEQCPVFLDADLAAFTSHFKQILADMHTREKQLQTIFDSLPPARYKDTVTTILSWIQ
   SK       N              N      N  A  N   Q   MTT       I   TPSE TAI S  KI    VN L  G        I  I     G G M      FV  N    VFS VQA   E      T   M  QE MSA RT V
                                        1000                                                                                               1100
QSETKVSIPPVAVAEYEIMEQRLGELKALQSSLQEQQKGLKYLNTTVEDLSRKAPAEVSQKYRSEVELIVGRWKKLSSQLVEHCQKLEDLMTKLQRFQNDTKTLKKWMAEVDVFLKEEWPALGDSEALEKQLEQCTALVNDIQTIQPSLNSVNEIGKKMK
     L   QLS TD           Q          S  Y  S   KEM K   S I R  Q F E E                   EQ N  RKI  HIQ                        I K   K  RL  S              G Q I
                                                                               1200
REAEPEFASRIATELKDLNAQWEHICQQAHAKKAALKGGLDKTVSLRKDLSEMHEWITQAEEEYLERDFEYKTPEELQKAVEELKRAKEDAMQKEVKVKLITDSVNNFIAKAPPAANEALKKELDVLITSYQRLCSRLNGKCKTLEEVWACWHELLSYLD
N         LE    E  T  D M   VY R E      E     Q         M                 D        M     EQ    A    L E   SV  Q   V Q       ET T N  W  T                      E
                    1300                                                                                                1400
AENKWLNEVELKLKATENIQGGAEEISESLDSLERLMRHPEDNRNQIRELAQTLTDGGILDELINEKLEKFNTRWEELQQEAVRRQKSLEQSIQSAQETDKTLRLIQESLAAIDKQLTAYTADRVDAAQVPQEAQKIQSELTSHEISLEEMKKRNRGKES
KA        F   T    P        V     N    S   P    I         VM      E  T  S  R  HE       L           E S H      TF     A  I  K     M         D            H Q   A
                                                            1500
AKRVLSQIDVAQKKLQDVSMKFRLFQKPANFEQRLQECKRILDEVKLQVPKLETKSVEQEVVQSHLDHCMKLYKSLSEVKSEVETVIKTGRQIVQKQQTENPKELDERLTALKLQYNELGAKVTEKKQELEKCLKLSRKLRKEINSLTEWLAATDVELTK
 Q                              L    S M      MHL A             Q N  VN             M           K           V     H          R  Q          M   M V        M
.600                                                                                                1700                  ▼
RSAVQGMPSNLDAEIAWGKATRKEIEKRQVQLKNICDLGENLKTVLKGKESLVEDKLSLLNSNWIAVTSRAEEWLNLLMEYQKHMEAFDQKVANVTTWIYRAEILLDESDKQKPQQKEETLKRLKAELNDMHPKVDSVRDQAVDLMTNRGDHCRKVIEPK
    E       SV       Q     QK H  S TEV  A     GK  T                           L       T   N DHI K   IQ DT    E K      DV          IR     T    AN A         LV Q
                                        1800                                                                                               1900
LSELNHRFAAISQRIKSGKPFIPLKELEQFDFDIQKLLEPLEVEIQQGVNLKEEDFNKDMSEDDESTVKELLQRGDTLQKRITDERKREEIKIKQQLLQTKHNALKDLRSQRRKKALEISHQWYQYKRQADDLMTWLDDIEKKLASLPDHKDEQKLKEIG
I           H   T  AS         NS          A                 N  N G          N Q                                                     LKC            EPR  R I   D
                                                                               2000
GELEKKKEDLNAVNRQAERLSKDGAAKAVEPTLVQLSKRWRDFESKFAQFRRLNYAQIQTVLEDTTFVMTESMTVETTYVPSTYLAEILQLLQALSEVEERLNSPVLQAKDCEDLLKQEECLKNIKDCLGRLQGHIDIIHSKKTPALQSATPRETANIQD
R  Q    E    R    G  E    M     QI        EI          F   H  R E MM    D PL IS       T  THVS   L   QL  A D C   F   F    S      S QQSS R         A       V RVKL E
                    2100                                                                                                2200
KLTQLNSQWEKVNKMYRDRQARFDKSKEKWRLFHCEMKSFNEWLTETEEKLSRAQIEAGDVGHVKTKQFLQELQDGIGRQQTVVKTLNVTGEEIIEQSSAADANVLKEQLGNLNTRWQEICRQLVEKRKRIEEEKNILSEFQEDLNKLILWLEETENVIA
A S  DF         K   G   R V    R  YDI I Q     A QF RKT  PENWE- A Y WY K       QR    R   A      Q   KT  SI Q K  S  L   V K   SDRK  L Q         R   EFV    AD IAS
                                                            2300
IPLEPGNEDQLRDCLGKVKLRVEELLPHKGILKRLNETGGTTLGSASLNPERKHKLESTLKEASRRLLKVSRDLPEKQKEIEILLKDFIELNQ-------QINQLTLWITPVKNQLELYNQVGQPGAFDIKETEAAVQAKQPNVEEVLSKGCHLYKEKPA
      K Q  KEK EQ   L    PLRQ     Q   PV  V   PIS   EQD    NK QTNLQWI  A       G    AQI  LGQ EKKLEDLEE L H L LS   IR    I   PN E P  VQ  I       D   I    Q
2400                                                                                                                       2500
THPVKKKLEDLNADWKAINHLILQLKEKPTFGEPALTSPGVLTSGQTVAVDTQARVTKETTSFTPEMPSSVLLEVPALADFNKAWAELTDWLSRLDREIKAQRVTVGDLDDINDMIIKQKANMQDLEQRRPQLDELITAAQNLKNKTSNQEARTIITDRI
Q   R      SSE    V R LQE RA  QPDLA G TI  ASPT- TLV  PV    AISKL    LM          R T         L  QV  S    M   E  E     T                E
                                               2600                   ▼                                                                           2700
EKIQSQWDDVHGYLQNRRQQLHEMQKDSTQWLEAKQEAEQVLEQAKAKLESWKEISYTVEALKKQNSELKQFSKEIRQWQMNIEGVNDVALKPVRDYSADDTRKVELMTDNINATWATINKRVSEREAALESALLMLQEFYLDLEKFLAWLTEAETTANV
 R  N   E QEH         N  L          E    G  R        GP   D IQ KIT T LA DL    T VDVA  L  LL               HMI  E    S RS H              ETHRL  Q P
                                                                                                2800
LQDATHKEKTLEDPQMVRELMKQWQDLQAEIDAHTDIFHNLDENGQKILRSLEGSEDAVLLQRRLDNMNFRWSELRKKSLNIRSHLEASTDQWKRLHLSLQELLAWLQLKEDELKQQAPIGGDIPTVQKQNDVHRTFKRELKTKEPVIMNALETVRLFLA
    R   RL   SKG   K       G   E   VY      S       D         K                            S                              V   D   SR        F   A        ST  I T
                         2900                                                                                          3000
DQPVEGLEKVYPEPRDLSPEERAQNVTKVLRRQADDVRTEWDKLNLRSADWQKKIDDALERLQGLQEAMDELDLKLRQAEAFKGSWQPVGDLLIDSLQDHLEKVKVYRAEMVPLKEKVHQVNELAHRFAPPDIQLSPYTLSCLEDLNTRWKVLQVAIDER
 E   L    LQ  E   P       RL K  EE  N     E     H    R     ET   E  T            VI                              AL  G IA   N SH  D  RQLTTLG        N   T  L  VED
                                                                                          3100
IRQLHEAHRDFGPTSQHFLTTSVQGPWERAISPNKVPYYINHETQTTCWDHPKMTELYQSLADLNNVRFSAYRTAMKLRRLQKALCLDLLNLSAACDALDQHNLKQNDQPMDILQIINCLTTIYDRLEQEHNNLVNVPLCVDMCLNWLLNVYDTGRTGRI
V      A    S                                                                                  S
3200                                                                                                    3300
RVLSFKTGVVSLCKAHLEDKYRYLFKQVASSTGFCDQRRLGLLLHDSIQIPRQLGEVASFGGSNIEPSVRSCFQFANNKPEIEAALFLDWMRLEPQSMVWLPVLHRVAAAETAKHQAKCNICKECPIIGFRYRSLKHFNYDICQSCFFSGRVAKGHKMHY
  II
                                           3400
PMVEYCTPTTSGEDVRDFAKVLKNKFRTKRYFAKHPRMGYLPVQTVLEGDNMETPVTLINFWPVDSA----------------------LAEMENSNGSYLNDSISPNESIDDEHLLIQHYCQSLNQESPLSQPRSPAQILISLESEERGELERILADLE
                                                                   PASSPQLSHDDTHSRIEHYASR                                  D
3500                                                                                                          3600
EENRNLQAEYDRLKQQHDHKGLSPLPSPPEMMPVSPQSPRDAELIAEAKLLRQHKGRLEARMQILEDHNKQLESQLHRLRQLLEHAQADAKVNGTTLSSPSTSLQRSDSSQPMLLRVVGSQTSETMGEDDLLSPPQDTSTGLEEVMEQLNNSFPSSRGRN
                                      E              T                                                QP E    V                               DS   E
APGKPVREATM
T    M  D
```

Fig. 2. Amino acid sequence of chicken dystrophin. Amino acids that differ in the human sequence are indicated below the chicken sequence. Positions of two exon limits are indicated by arrows.

changes are conservative). In particular there are only two amino acid changes between positions 3123 and 3418. The chicken lacks a 22 aa sequence found in humans. As was found for the human sequence, the chicken protein shows evidence of internal repeats (Figure 4). The dot matrix pattern has a striking resemblance with that obtained by

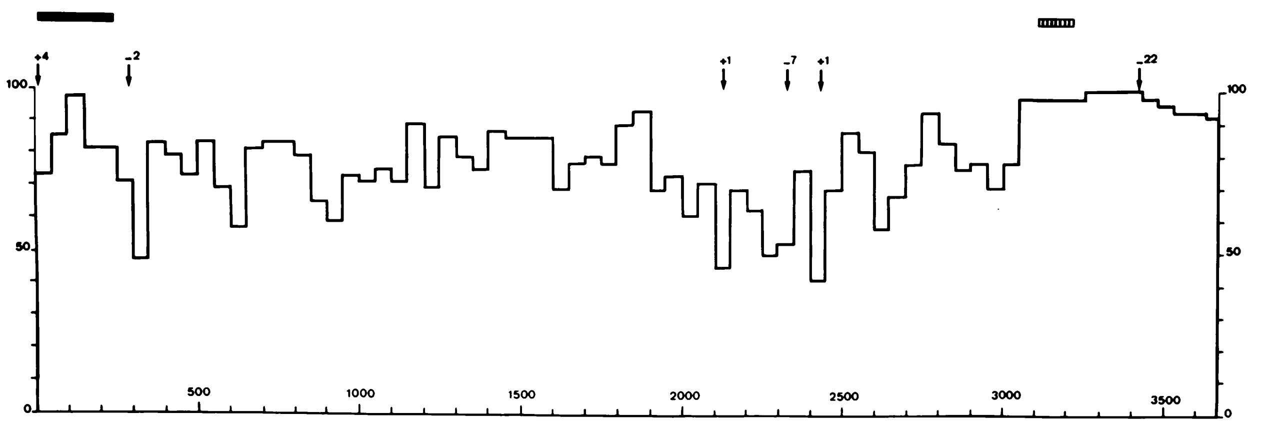


Fig. 3. Comparison of the chicken and human dystrophin sequence. The percent homology was calculated for stretches of 50 amino acids. The black and hatched bars indicate the regions homologous to the actin binding region and EF hand region of α actinin, respectively. Arrows and numbers indicate position and extent (in the number of amino acids) of insertions (+) or deletions (−) with respect to the human sequence.

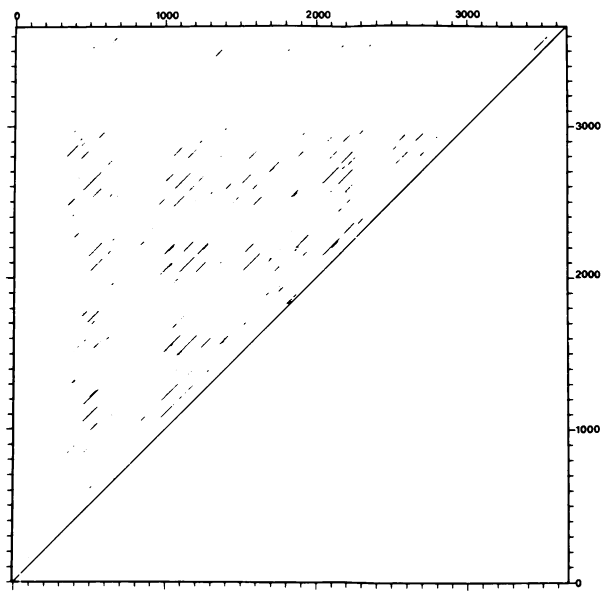


Fig. 4. Detection of internal repeats in the dystrophin sequence. The compare program (UWGCG) was used with the same parameters as in Koenig *et al.* (1988) (match in at least 39 aa out of 100).

Koenig *et al.* (1988) showing that the repeat pattern is best conserved in the regions corresponding to repeat 1,2,8,9,10,18,19,22,23,24,25.

The very high conservation of the C-terminal region indicates that it must play an important functional role. Using a program based on Garnier *et al.* (1978) predictions for secondary structure features were obtained. Among the characteristics, one can note alternating clusters of hydrophobic and hydrophilic amino acids between residues 3100 and 3220, and a relative lack of α helices, compared to the internally repeated region (not shown).

We compared the dystrophin sequence with protein sequences in the SWISS-PROT database and with some recently published sequences (corresponding to 12 calcium binding proteins and to fodrin, a non-erythroid spectrin). Homology to calcium binding proteins appeared non-significant and we did not find additional support to the hypothesis that dystrophin could contain calcium binding EF hand-like structures in the C-terminal region (Koenig *et al.*, 1988). We detected no homologies as convincing as that found between α actinin and the N-terminal end of dystrophin (Hammond, 1987; similarity score 480). Some marginally significant alignments were obtained (with similarity scores between 100 and 180, while scores were always inferior to 80 with the randomized dystrophin sequence). Apart from weak similarities with protein containing rod-like structures (tropomyosins and myosins, lamin A and keratin), already noted by Hoffman *et al.* (1987b) and Cross *et al.* (1987) we found similarities to chicken desmin (also in the rod-like portion) and mouse laminin. Quite surprisingly the best scores (130 to 180) were obtained between the spectrin-like region and human apolipoprotein A4.

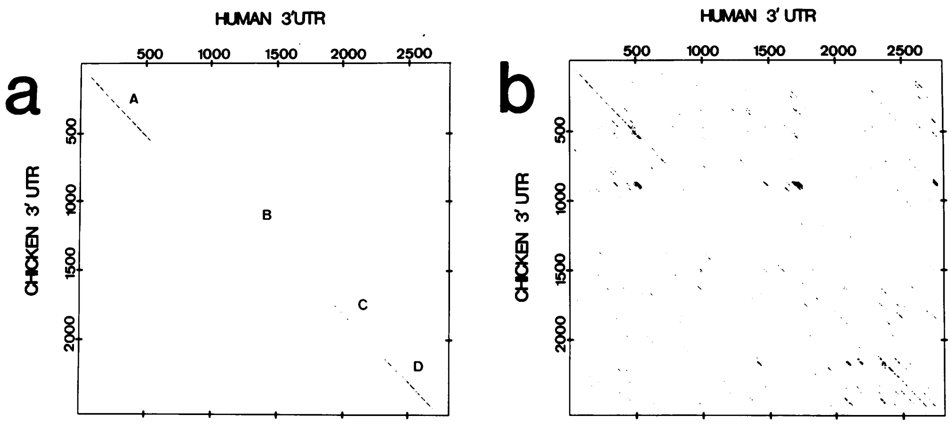
Comparison of untranslated sequences

The comparison of the 5′ untranslated sequences (150 nt in chicken were compared to the 208 nt human sequence) did not show any significant homology, even right upstream from the initiation codon. However, both sequences are rich in runs of Ts. In contrast, very high homology was obtained when the 3′ untranslated sequences were compared. There is 90% homology in the 473 nt that immediately follow the termination codon (Regions A in Figure 5). (By comparison, an 87% homology is found for the 1500 nt preceding the stop codon, i.e. the region which is best conserved at the protein coding level.) Two other regions of 82% homology were found over 113 and 343 nt (Regions C and D), close to the human polyadenylation signal. There is no significant homology in the internal 3′ untranslated region B which is however of similar size in both species (Figure 5).

The conservation of a large part of the 3′ UTR in mammals was confirmed by Southern blot hybridization (Figure 6). A human probe which corresponds to most of region A but which contains 3′ adjacent sequences is conserved on the X in lemurs, but contains a region of homology to highly repetitive sequences in rodents. A probe covering regions C and D showed strong cross-hybridization to lemur, mouse, rat and bovine genomic DNAs. This sequence is not repeated in these genomes or in chicken and X linkage was confirmed for lemur and mouse. No cross-hybridization was found for region B (not shown).

Discussion

The function of dystrophin is unknown at present and the only clues come from the subcellular localization on the



		homology %	n° of gap
A	1 to 473	90	22
C	1703 to 1814	80	1
D	2099 to 2441	83	29
	(2320 to 2441)	(95)	(1)

Fig. 5. Comparison of human and chicken 3' untranslated sequences. The matrix comparison program (Microgenie) was used under two different sets of parameters. The extent of highly homologous sequences **A**, **C** and **D** is best shown under high stringency parameters (match in at least 50 nt out of 83, **panel a**), while the absence of conservation of the internal region of the 3' UTR is apparent under the less stringent conditions (match in at least 20 nt out of 33, **panel b**). Numbering starts at the stop codon.

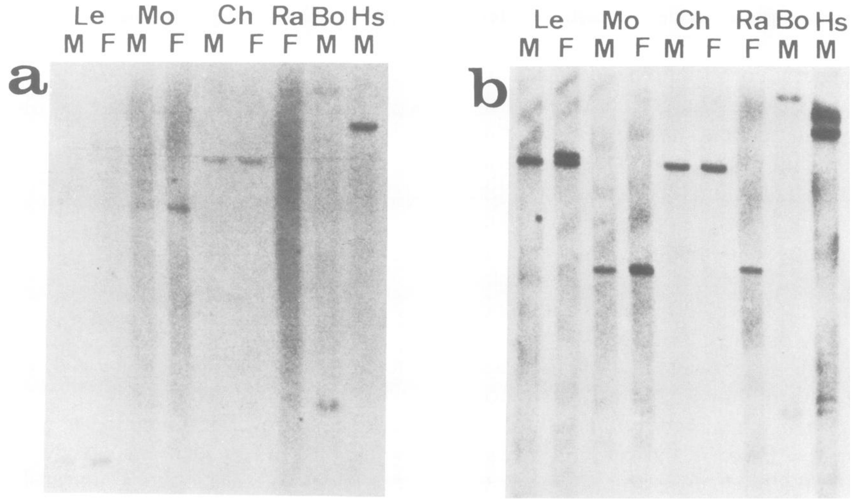


Fig. 6. Conservation of the 3' UTR in various species by Southern blot analysis. DNAs were digested with *Eco*RI and correspond to male (M) and female (F) Lemur (Le), Mouse (Mo), Chicken (Ch), to female Rat (Ra) and to male Bovine (Bo) and Human (Hs). Sex dependent dosage is found in lemurs (note an *Eco*RI polymorphism with probe C–D) and mouse, but not in chicken. **a**, hybridization to human probe A (see text and Figure 5) washing in 1 × SSC at 55°C. **b**, hybridization to human probe C–D, washing in 0.5 × SSC at 65°C.

plasma membrane (Watkins and Cullen, 1987; Arahata *et al.*, 1988; Zubrzycka-Gaarn *et al.*, 1988), or from comparison to the sequence of other proteins, of more or less known function. Strong homology has been found between the N-terminal regions of dystrophin and α actinin (Hammond, 1987), while the region from 282–3078 shows a repeat pattern similar to that of spectrin, with weak homologies of repeat units between dystrophin and spectrin (Davison and Critchley, 1988; Koenig *et al.*, 1988). A region of weak homology has also been described between residues 3112 and 3208 of dystrophin, and a potential calcium binding region of α actinin (Koenig *et al.*, 1988). By comparing the chicken and human dystrophin sequence we have found that the most conserved region is the 627 aa C-terminal region which starts at the end of the spectrin-like domain, and includes a cystein rich region (which contains the weak homology to the Ca^{2+} binding sites of actinin), and a region which shows no strong homology to known proteins. The very high selective pressure on this C-terminal region suggests that it plays an important role of interaction with other proteins, since similar conservation between chicken and man is found for proteins that are part of highly structured elements of the cell (actins, histones).

The rest of the protein shows 75% homology, which is less than, for instance, the 84% homology found between human and chicken myosin heavy chain (Molina *et al.*, 1987).

The second striking feature is the extremely high conservation of a large part of the 3′ untranslated region in several mammalian species and in chicken. Search for ORFs in both orientations and in both human and chicken sequences shows that this conservation cannot be due to the presence of protein coding sequences from a neighbouring or overlapping gene (not shown). Other cases of conservation of all or part of 3′ UTR over the same evolutionary period have only been reported for the histone H3.3 gene (85% homology over 520 bp), the β actin gene (60% homology over 590 bp, including shorter regions of 90% homology) or α skeletal actin ($\sim$90% over $\sim$100 bp) and the c-*fos* gene (92% over 300 bp). It is difficult at present to explain these features, since protein binding regulatory sequences are much shorter. In c-*fos* mRNA the highly conserved sequence upstream of the polyadenylation signal is thought to be implicated in the short half-life of the mRNA (Rahmsdorf *et al.*, 1987) and similar effects of AU rich sequences have been described for a lymphokine (GM-CSF) mRNA (Shaw and Kamen, 1986). However, one would not expect that a very large mRNA which codes for a structural protein, should be subjected to such a regulatory mechanism (especially since transcription of the DMD gene should take $\sim$20 h). It should also be noted that conservation between chicken and man of an intronic region of the dystrophin gene has been found previously (Heilig *et al.*, 1987).

Although deletions are scarce in the 3′ end of the dystrophin gene, it should be interesting to analyse the phenotype of such deletions in the C-terminal coding region and to search for deletions limited to the 3′ non-coding region in patients. The identification of the proteins that interact with the highly conserved C-terminal region of dystrophin appears as an important goal.

Materials and methods

Construction of the cDNA library

Poly(A)$^+$ RNA was prepared from leg muscle of 13-day-old chicken by the guanidium hydrochloride method (Munnich *et al.*, 1982), and tested for the presence of the specific transcript by dot hybridization to the chicken CCR-P probe (Heilig *et al.*, 1987). Two libraries were established in λgt10 following the Gubler and Hoffman (1983) procedure, the first using oligo(dT) priming, the second using random oligonucleotides (13 mer) for priming the first strand. Double-stranded cDNAs >3 kb were selected by sedimentation on a sucrose gradient. They were ligated to the λgt10 *Eco*RI digested vector, *in vitro* packaged and plated. Enzymes and reagents were from Amersham cDNA synthesis kit. 0.8 × 10^6 recombinants were obtained for the oligo(dT) primed library and 2.5 × 10^6 for the random primed library, starting from 5 μg and 2 μg of poly(A)$^+$RNA, respectively.

Screening of the cDNA libraries

The libraries were plated at high density ($\sim$25 000 plaques per ϕ 14-cm dish) for a total of 3.3 × 10^6 clones. Replicas were made on Pall (Biodyne) nylon membrane. Screening with probe CCR-P yielded 3.5 kb of cDNA sequence (clones 1 and 2 in Figure 1). Four additional walking steps were performed using the 3′ side of the preceding clone as probe, giving clones number 3–7 and clone 11 (Figure 1). Clones 8, 9 and 10 were obtained by screening the libraries with a human cDNA probe (9–14) corresponding to the last 6.3 kb (Koenig *et al.*, 1987). Several clones ended on one side at *Eco*RI sites present within the dystrophin cDNA sequence, due to imperfect methylation during the construction of the libraries.

Sequencing strategy

The plasmid vector pEMBL18$^+$ which produces ssDNA upon superinfection with the phage M13K07 (Pharmacia) was used for subcloning in *Escherichia coli* TG2. cDNA inserts containing one or more internal *Eco*RI site(s) were subcloned after isolation on LMP agarose of the upper fragment of a partial digest of the corresponding λ clone DNA.

Single-stranded DNA was sequenced using the dideoxy chain terminator method (Sanger *et al.*, 1977) and sequenase DNA sequencing kit (USB Kit no 70 700) with [^{35}S]dATP (NEN). An ordered deletion strategy was used to sequence clones 1 and 2 (DNaseI–Mn^{2+} method of Lin *et al.*, 1985).

For clones 3–11, sites present both in the polylinker and in the cDNA insert were used to generate clones with partial deletions and to obtain sequence data using the universal primer. 60–70% of the cDNA sequence was obtained in this way. Specific oligonucleotides (20 mers) were used as primer to complete the sequence, most of which was determined on at least two independent clones. 60-cm long polyacrylamide-urea gels were used in general (acrylamide to bisacrylamide ratio of 30:1). 8% gels were run at 40 W (27 000 V/h) and 5% gels at 35 W (33 000 V/h). Up to 750 nt could be read from the primer, under these conditions. cDNA sequence data will be submitted to *Nucleic Acids Research*, For the Record.

Computer analysis

The nucleotide sequence and deduced amino acid sequence were analysed for internal repeats and for secondary structure predictions using programs of the University of Wisconsin Genetics Computer Group (UWGCG, Devereux *et al.*, 1984), Micro Genie software package (Beckman) and ANALYSIS software package (LGME, Strasbourg, France).

The search for similarities between the dystrophin sequence and sequences in the SWISS-PROT database and EMBL nucleic acid database was performed using programs based on the algorithm of Wilbur and Lipmann (1983).

Acknowledgements

We would like to thank C.Kretz and C.Weber for efficient technical assistance; M.Koenig and L.Kunkel for providing the human cDNA clones, A.Staub and F.Ruffenach for oligonucleotide synthesis, G.De Marcillac and A.Eichwald for computer help and D.Massel for typing the manuscript. This work was supported by grants from the Association des Myopathes de France, the Fondation pour la Recherche Médicale Française to J.L.M.

References

Arahata,K., Ishiura,S., Ishiguro,T., Tsukahara,T., Suhara,Y., Eguchi,C., Ishihara,T., Nonaka,I., Ozawa,E. and Sugita,H. (1988) *Nature*, **333**, 861–863.

Burmeister,M., Monaco,A.P., Gillard,E.F., Van Ommen,G.J.B., Affara,N.A., Ferguson-Smith,M.A., Kunkel,L.M. and Lehrach,H. (1988) *Genomics*, **2**, 189–202.

Cross,G.S., Speer,A., Rosenthal,A., Forrest,S.M., Smith,T.J., Edwards,Y., Flint,T., Hill,D. and Davies,K.E. (1987) *EMBO J.*, **6**, 3277–3283.

Davison,M.D. and Critchley,D.R. (1988) *Cell*, **52**, 159–160.

Devereux,J., Haekerli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.

Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) *J. Mol. Biol.*, **120**, 97–120.

Gubler,U. and Hoffman,B. (1983) *Gene*, **25**, 263–269.

Hammond,R.G.,Jr (1987) *Cell*, **51**, 1.

Heilig,R., Lemaire,C. and Mandel,J.L. (1987) *Nucleic Acids Res.*, **15**, 9129–9142.

Hoffman,E.P., Brown,R.H.,Jr and Kunkel,L.M. (1987a) *Cell*, **51**, 919–928.

Hoffman,E.P., Monaco,A.P., Feener,C.A. and Kunkel,L.M. (1987b) *Science*, **238**, 347–350.

Koenig,M., Hoffman,E.P., Bertelson,C.J., Monaco,A.P., Feener,C. and Kunkel,L.M. (1987) *Cell*, **50**, 509–517.

Koenig,M., Monaco,A.P. and Kunkel,L.M. (1988) *Cell*, **53**, 219–228.

Lin,H.C., Lei,S. and Wilcox,G. (1985) *Anal. Biochem.*, **147**, 114–119.

Molina,M.I., Kropp,K.E., Gulick,J. and Robbins,J. (1987) *J. Biol. Chem.*, **262**, 6478–6488.

Munnich,A., Daegelen,D., Besmond,C., Marie,J., Dreyfus,J.C. and

Kahn,A. (1982) *Pediatr. Res.*, **16**, 335–339.

Rahmsdorf,H.J., Schönthal,A., Angel,P., Liftin,M., Rüther,U. and Herrlich,P. (1987) *Nucleic Acids Res.*, **15**, 1643–1659.

Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA.*, **74**, 5463–5467.

Shaw,G. and Kamen,R. (1986) *Cell*, **46**, 659–667.

Watkins,S.C. and Cullen,M.J. (1987) *J. Neurol. Sci.*, **82**, 181–192.

Wilbur,W.J. and Lipman,D.J. (1983) *Proc. Natl. Acad. Sci. USA.*, **80**, 726–730.

Zubrzycka-Gaarn,E.E., Bulman,D.E., Karpati,G., Burghes,A.H.M., Belfall,B., Klamut,H.J., Talbot,J., Hodges,R.S., Ray,P.N. and Worton,R.G. (1988) *Nature*, **333**, 466–469.

Received on August 15, 1988; revised on September 30, 1988