Dystrophin

Gene, protein and cell biology

Edited by

SUSAN C. BROWN

School of Biological Sciences, Royal Holloway, University of London, UK

and

JACK A. LUCY

Department of Cellular Physiology,
The Babraham Institute, Cambridge, UK

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

© Cambridge University Press 1997

This book is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 1997

Printed in the United Kingdom at the University Press, Cambridge

Typeset in Times 11/14

A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication data

Dystrophin: gene, protein, and cell biology / edited by Susan S.
Brown and Jack A. Lucy.
p. cm.
Includes bibliographical references and index.
ISBN 0 521 55033 5 (hardbound)
1. Dystrophin. 2. Dystrophin genes. 3. Duchenne muscular
dystrophy – Molecular aspects. I. Brown, Susan S., 1958–
II. Lucy, Jack A., 1929–
QP552.D95D97 1997
612'.01575–dc21 96-37800 CIP

ISBN 0 521 55033 5 hardback

Contents

List of contributors xiii
Foreword xvii

1 **Structure and mutation of the dystrophin gene** 1
Andrea Amalfitano, Jill A. Rafael and Jeffrey S. Chamberlain
Introduction 1
Clinical considerations 1
General genetics 2
The dystrophin gene 3
Dystrophin domains 10
Mutations leading to DMD/BMD 13
Exceptions and phenotype/genotype correlations 16
N-terminal mutations 17
The central rod domain 18
The C-terminus 19
References 20

2 **Protein structure** 27
S. J. Winder, A. E. Knight and J. Kendrick-Jones
Overall structure 27
Actin-binding domain 30
Central triple-helical repeat region 36
COOH-terminal domains 42
Regulation 48
Summary 51
References 52

3 **Dystrophin, utrophin and their associated proteins** 56
J. M. Tinsley, D. J. Blake and K. E. Davies
Introduction 56
Transcripts from the DMD locus 56
Dystrophin-associated proteins 59
Mutations in the components of the DPC 62
Transcripts from the UTRN locus 64
Similarities and differences between dystrophin and utrophin 65
References 70

4 **Localization of dystrophin in skeletal, cardiac and smooth muscle** 79
Simon C. Watkins, Darl R. Swartz and Timothy J. Byers
Background: the early work 79
The discovery of dystrophin 81
The differential distribution of dystrophin within muscle: electron microscopy and high resolution light microscopic observations 82
Immuno-electron microscopy 83
High-resolution light microscopy 88
Dystrophin, the basal lamina and periodicity 92
Studies of washed, isolated myofibrils 93
Specialized distributions of dystrophin at the plasma membrane 95
The neuromuscular junction 95
Myotendinous junction 96
Cardiac muscle 98
Smooth muscle 99
Summary 100
References 100

5 **Expression of the dystrophin complex in the brain** 105
Dariusz C. Górecki and Eric A. Barnard
Introduction 105
Psychological studies 105
Neurological and neuropathological studies 106
Expression of specific dystrophin transcripts in the brain 107

Truncated dystrophin isoforms: apodystrophins 115
Dystrophin isoforms in retina 118
Alternative splicing at the 3′-end of dystrophin transcripts in the brain and in the retina 118
Diversity of brain dystrophins 120
Genotype–phenotype correlation in DMD-associated mental retardation 121
Immunolocalization of dystrophin in the brain 123
Co-expression of dystroglycan and dystrophin in specific brain areas 125
Functional implications of the dystrophin complex in the brain 128
References 131

6 **The dystrophin family, agrin and construction of the neuromuscular junction** 139
Robert Sealock and Stanley Froehner
Introduction 139
Agrin 141
The 43 kDa protein (rapsyn) 149
Syntrophins 152
Utrophin 155
References 156

7 **Functions of dystrophin** 163
Susan C. Brown and Jack A. Lucy
Introduction 163
Structure of dystrophin and its association with other proteins 165
The DAGC and other forms of muscular dystrophy 166
Dystrophin and its relationships to other cytoskeletal components 167
Functional diversity of dystrophin 168
Dystrophin isoforms and autosomal homologues in striated muscle 169
Dystrophin deficiency and membrane integrity 170
Dystrophin and the regulation of intracellular calcium 175
Dystrophin and signal transduction 184
Summary 191
References 192

8 Pathophysiology of dystrophin deficiency: a clinical and biological enigma 201
J. Rafael M. Gorospe, Bronwen K. Nishikawa and Eric P. Hoffman
Problems presented by dystrophin deficiency 201
Consequences of dystrophin deficiency 207
A pathophysiological model 211
Conclusion 221
References 223

9 Muscle regeneration 233
M.J. Cullen
Introduction 233
Regeneration in normal muscle 235
Regeneration in dystrophic muscle 253
Conclusions 263
References 264

10 Gene therapy of Duchenne muscular dystrophy 274
Gaynor Turner, Matthew G. Dunckley and George Dickson
Introduction 274
Problems associated with gene transfer to skeletal muscle 277
Adenoviral vectors for gene therapy of DMD 283
Retroviral vector systems 290
Future perspectives 298
References 301

11 Models of dystrophinopathy, pathological mechanisms and assessment of therapies 310
T.A. Partridge
Introduction 310
Animal models of muscular dystrophy 311
The *mdx* mouse 312
Dystrophic dogs 313
Dystrophic cat 313
The need for other large animal models of muscular dystrophy 314
Assessment of damage *in vivo* 314

Stages of disease in dystrophinopathies 316
Tissue culture models 320
Prospects for therapy 322
Genetic therapies 322
References 326

Index 333

1

Structure and mutation of the dystrophin gene

Andrea Amalfitano, Jill A. Rafael and Jeffrey S. Chamberlain

Introduction

Duchenne muscular dystrophy (DMD) is the most frequent of the lethal, X-linked recessive disorders, occurring once per 3500 male births (Emery, 1993). The gene mutated in DMD and the milder Becker muscular dystrophy (BMD) encodes the protein dystrophin. BMD is less common than DMD, and has an estimated incidence of approximately 1 in 18 500 births (Bushby, Thambyayah & Gardner-Medwin, 1991). The original clinical descriptions of DMD are quite lucid and readily convey the poor prognosis of a DMD diagnosis. This chapter will concentrate on the differences between DMD, BMD, and intermediate phenotypes, and how they correlate with mutations in various portions of the dystrophin gene.

Clinical considerations

DMD patients are clinically normal at birth, although serum levels of the muscle isoform of creatine kinase (M-CK) are quite elevated. Elevated enzyme levels taper off with increasing age, likely reflecting decreasing overall muscle mass, due to increasing dystrophic involvement. One of the first clinical signs of DMD is delayed achievement of motor milestones, including a delayed ability to walk (56% of DMD patients do not walk until after the age of 18 months), excessive clumsiness, and difficulty running. These symptoms alone can make it difficult to arrive at a diagnosis; therefore, a high index of suspicion has to be maintained in any male infant with delayed achievement of motor milestones in regard to a possible diagnosis of DMD. Eventually, the onset of symmetric, bilateral, muscular pseudohypertrophy, proximal limb muscle weakness, and the Gower's sign strongly suggest a DMD diagnosis. With time, decreased muscle strength and joint contractures result in wheelchair dependence. Finally, involvement of the diaphragm and cardiac muscles results in most of

Table 1.1 *Distinguishing DMD from BMD (described by Emery, 1993)*

	Becker muscular dystrophy	Duchenne muscular dystrophy
Onset	>5.3 years in 88%	<5.3 years in 88%
Wheelchair dependence	>12.4 years in 96%	<12.4 years in 96%
Death	>21.1 years in 96%	<21.1 years in 93.7%
Dystrophin protein expression	Decreased, but present*	Absent, or extremely low*

*With exceptions, see text.

the clinical mortality. While most patients succumb due to the repeated pulmonary infections secondary to respiratory muscle involvement, 10–20% of the mortality can be attributed to cardiomyopathy and/or cardiac conduction abnormalities. Furthermore, one-third of DMD patients demonstrate varying degrees of cognitive impairment. In BMD affected individuals, the clinical course is similar to that of DMD patients; however, the onset of symptoms and the rate of progression are delayed. Table 1.1 summarizes several key diagnostic features used to distinguish between DMD and BMD (Emery, 1993). An intermediate phenotype has also been described for those individuals who can ambulate after the age of 12, but begin using a wheelchair before the age of 16.

General genetics

The pattern of male-limited expression due to X-linked inheritance of DMD was understood even with the earliest descriptions of the disorder, and has a number of important clinical implications. Haldane's hypothesis states that for any X-linked recessive disorder, the mutation rate $= 1/3(I)(1 + f)$, where f equals the reproductive fitness, and disease incidence is I (Haldane, 1935). Thus, in a lethal disorder such as DMD, $f = 0$, and the mutation rate becomes $1/3(I)$. This rate indicates that one-third of all cases of DMD should arise from a new, sporadic mutation, a prediction that has been largely confirmed by direct mutation detection studies in families (Ward *et al.*, 1989; Müller *et al.*, 1992). Conversely, the likelihood that an affected individual inherited the mutation from a carrier female is two-thirds, a fact that has serious ramifications in regard to genetic counselling and involvement of other family members. A diagnosis in a single individual should always be followed by genetic counselling and evaluation for carrier status of at-risk individuals in the pedigree. New mutations often occur early during female

embryonic development, resulting in germ-line mosaicism for the mutation. Estimates for germ-line mosaicism in DMD mothers are 14%–18% when the at-risk chromosome is known (Bakker *et al.*, 1989). As a result, even 'non-carrier' females who have given birth to a DMD male should be considered at risk for having another affected male child.

The dystrophin gene mutation rate of approximately 1×10^{-4} genes per generation is one of the highest mutation rates reported, and in part reflects the enormous size of this locus (see below). With the cloning of the gene, a number of rapid diagnostic tests have been developed for both direct mutation identification and linkage analysis using multiple intragenic polymorphisms. This information can be used to ascertain whether a mutation is sporadic or inherited, and also for highly accurate prenatal diagnosis and carrier detection (see below). DNA diagnostics can detect gene mutations or polymorphisms in blood samples, amniotic fluid cells, and chorionic villus samplings (see Chamberlain, 1993, for a review) while dystrophin antisera can distinguish DMD from BMD and other muscle disorders by analysis of muscle needle biopsy samples (Hoffman *et al.*, 1988*a*). Numerous studies have now been reported regarding the various types of mutations responsible for DMD or BMD. These observations have a number of implications not only for genetic counselling issues, but also for genotype/phenotype correlations of the dystrophin protein, as well as for potential attempts at a clinical treatment for this lethal disorder.

The dystrophin gene

The dystrophin gene was the first gene isolated by positional cloning methods and has proved to be the most complex genetic locus identified (Monaco *et al.*, 1986; Burghes *et al.*, 1987; Ahn & Kunkel, 1993). It is the largest known gene, spanning at least 2.4 megabases (Mb) at Xp21.1 on the human X-chromosome and is composed of a minimum of 85 exons that have been well conserved throughout vertebrate evolution (Koenig *et al.*, 1987; Lemaire, Heilig & Mandel, 1988; Monaco *et al.*, 1992; Roberts *et al.*, 1993). Seven different promoters have been identified, with some evidence for an eighth (Nishio *et al.*, 1994), and more may remain to be uncovered (see below). These different promoters generated at least five different protein isoform size classes that are expressed in tissue-specific patterns. Alternative splicing at the 3′ end of the gene leads to the expression of an even greater variety of isoforms (Fig. 1.1) (Feener, Koenig & Kunkel 1989; Bies *et al.*, 1992*b*).

The dystrophin coding region shares clear homology with three other classes of genes. The entire dystrophin sequence is similar to that of utrophin

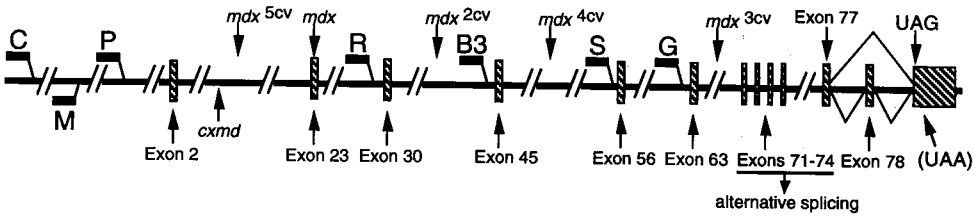

Fig. 1.1. Schematic illustration of the organization of dystrophin mRNA isoforms. Indicated are the relative position of the seven confirmed dystrophin promoters, key exons as discussed in the text, and the location of characterized mutations in animal models for DMD. Five mouse *mdx* mutations have been identified (Sicinski *et al.*, 1989; Cox *et al.*, 1993 and Im *et al.*, 1996), as has the *cxmd* golden retriever mutation (Sharp *et al.*, 1992), and all six are point mutations. Exons 71–74 are alternatively spliced in complex patterns, and exon 78 is also alternatively spliced from some transcripts, leading to the use of an alternate stop codon in exon 79. Additional details are described in the text.

(also known as the dystrophin-related protein; (Pearce *et al.*, 1993)), while the 5′ and central portions are homologous with members of the spectrin gene family, including α-actinin (Koenig, Monaco & Kunkel, 1988). The 3′ end of dystrophin is homologous with an 87 kilodalton (kD) post-synaptic protein (dystrobrevin) characterized from *Torpedo* electric organ (Wagner, Cohen & Huganir, 1993). These observations suggest that dystrophin and utrophin may have arisen during evolution by juxtaposition of ancestral spectrin and 87 kD-like genes to form a larger transcription unit.

The dystrophin cortical (C), muscle (M), and Purkinje (P) promoters express 'full-length' transcripts that consist of unique first exons spliced to a shared set of 78 exons (the names of these promoters indicate major, but not exclusive, sites of expression: Koenig *et al.*, 1987; Nudel *et al.*, 1988; Gorecki *et al.*, 1992). Full-length dystrophin isoforms are expressed in various sublocations within muscle and the central nervous system (CNS) (Chamberlain *et al.*, 1988; Lidov *et al.*, 1990; Gorecki *et al.*, 1992). The C promoter is expressed primarily in cortical neurons and portions of the hippocampus, while the P promoter is expressed in cerebellar Purkinje cells (see Chapter 5 and Lidov *et al.*, 1990; Lidov, Byers & Kunkel, 1993; Gorecki *et al.*, 1992). In contrast, the M promoter is expressed at low levels in some glial cells of the brain (Chelly *et al.*, 1990*b*), but at much higher levels in all muscles of the body (Hoffman *et al.*, 1988*b*; Nudel *et al.*, 1988). The transcripts from these promoters are approximately 14 kilobases (kb) long and generate proteins of approximately 427 kilodaltons (kD). Dystrophin has at least four internal promoters with unique first exons that splice into exons 30, 45, 56, and 63. These can be referred to, respectively, as the retinal (R), brain-3 (B_3), Schwann cell (S), and general (G)

promoters, which produce proteins of 260, 140, 116, and 71 kilodaltons (Dp260, Dp140, Dp116, and Dp71); (Rapaport *et al.*, 1992; Byers, Lidov & Kunkel, 1993; D'Souza *et al.*, 1995; Lidov, Selig & Kunkel, 1995).

The 85 dystrophin exons together span approximately 15 kb of genomic DNA, with perhaps another 10–50 kb of genomic DNA involved in promoter/enhancer functions. Thus, approximately 98% of the 2.4 Mb gene is composed of introns, and the large size of the locus appears to be conserved in mouse and chicken (Koenig *et al.*, 1988; Lemaire *et al.*, 1988; Chamberlain *et al.*, 1989). Due to the location of the various promoters, the gene contains at least 80 introns. The largest intron, other than the intron shared by the three 5′ promoters, is located between exons 44 and 45, and is approximately 180 kb (Monaco *et al.*, 1992). The largest exon (2.7 kb) is exon 79, which contains the entire 3′ untranslated region (Bies *et al.*, 1992*b*). The remaining exons average less than 200 bp (Roberts *et al.*, 1993). Fig. 1.2 displays the cDNA and predicted amino-acid sequences from the mouse dystrophin gene (Lee *et al.*, 1991; Chamberlain *et al.*, 1991), and highlights a number of features that will be discussed in this chapter. The murine amino acid sequence is 92% identical to the human dystrophin sequence, which has been described in detail elsewhere (Koenig *et al.*, 1988).

Deficiency of dystrophin in skeletal muscle leads to the major feature of DMD, progressive dystrophy of muscle tissue, found in patients and in several animal models of the disease. Striated and smooth muscle tissues express almost exclusively the full-length isoform of dystrophin, from the M-promoter, although low levels of alternatively spliced mRNAs lacking exon 71 (human), 71 and 72 (mouse), or 72 (chicken) have been observed (Feener *et al.*, 1989; Bies *et al.*, 1992*b*; Lemaire *et al.*, 1988). Additional alternatively spliced products

---

Fig. 1.2. (*Overleaf*) The cDNA and predicted amino-acid sequences for the mouse dystrophin gene (Lee *et al.*, 1991; Bies *et al.*, 1992*b*) from exon 1 (the mouse muscle first exon, 'M' in Fig. 1.1) into the beginning of the 3′ untranslated region (the full cDNA is 13 815 bp). The amino-acid sequence is shown in the one-letter code below the corresponding cDNA sequence, while features of the sequence are indicated above the cDNA sequence. The beginning of each exon and hinge region is indicated by an asterisk above the first base of the exon or hinge, followed by a number to designate order (e.g. *E2 designates the beginning of exon 2). Also indicated by an asterisk is the end of each hinge (arrows point towards inclusive regions of the hinges and other protein domains). Actin binding sites 1–3 are indicated (ABS1, etc.), and the inclusive amino acids are underlined. The WW domain begins at the 5′ end of exon 62 (WW- ->), and continues as indicated into exon 63. Also indicated are the α-syntrophin binding domain (α-syn B.D.), both leucine zippers (L-Zip 1 and 2; the leucines are underlined), and the location of the alternate stop codon used when exon 78 is spliced from the mRNA (alt C, in exon 79).

```
CCTCACTCAC TTGCCCCTTA CAGGACTCAG CTCTTGAAGG CAATAGCTTT ATAGAAAAAA CGAATAGGAA GACTTGAAGT GCTATTTTTT TTTTTTTTTT TGTCAAGGCT GCTGAAGTTT ATTGGCTTCT CATCGTACCT AAGCCTCCTG 150
                                                                                                                                  *E2                    ABS1
GAGCAATAAA ACTGGGAGAA ACTTTTACCA AGATTTTTAT CCCTGCCTTG ATATATACTT TTTCTTCCAA ATGCTTTGGT GGGAAGAAGT AGAGGACTGT TATGAAAGAG AAGATGTTCA AAAGAAAACA TTCACAAAAT GGATAAATGC 300
                                                                                M L W  W E E V  E D C  Y E R  E D V Q  K K T  F T K  W I N A
                 *E3                                                                                                  *E4
ACAATTTTCT AAGTTTGGAA AGCAACACAT AGACAACCTC TTCAGTGACC TGCAGGATGG AAAACGCCTC CTAGACCTCT TGGAAGGCCT TACAGGGCAA AAACTGCCAA AAGAAAAGGG ATCTACAAGA GTTCATGCCC TGAACAATGT 450
 Q F S  K F G  K Q H I  D N L  F S D  L Q D G  K R L  L D L  L E G L  T G Q  K L P  K E K G  S T R  V H A  L N N V
                                  *E5                                                  ABS3                                                                  *E6
CAACAAGGCA CTGCGGGTCT TACAGAAAAA TAATGTTGAT TTAGTGAATA TAGGAAGCAC TGACATAGTG GATGGAAATC ATAAACTCAC TCTTGGTTTG ATTTGGAATA TAATCCTCCA CTGGCAGGTC AAAAATGTGA TGAAAACTAT 600
 N K A  L R V  L Q K N  N V D  L V N  I G S T  D I V  D G N  H K L T  L G L  I W N  I I L H  W Q V  K N V  M K T I
                                                  ABS2
CATGGCTGGA TTGCAGCAAA CCAACAGTGA AAAGATTCTT CTGAGCTGGG TTCGACAGTC AACACGTAAT TATCCACAGG TTAACGTCAT CAACTTCACC TCTAGCTGGT CCGACGGGTT GGCTTTGAAT GCTCTTATCC ATAGTCACAG 750
 M A G  L Q Q T N S E K I L L S W V R Q S T R N Y P Q V N V I N F T S S W S D G L A L N A L I H S H R
*E7                                                                                                                                                  *E8
GCCCGACCTG TTTGATTGGA ATAGTGTGGT TTCACAGCAC TCAGCCACCC AAAGACTGGA ACATGCCTTC AACATTGCAA AATGCCAGTT AGGCATAGAA AAACTTCTTG ATCCTGAAGA TGTTGCTACC ACTTATCCAG ACAAGAAGTC 900
 P D L  F D W  N S V V  S Q H  S A T  Q R L E  H A F  N I A  K C Q L  G I E  K L L  D P E D  V A T  T Y P  D K K S
                                                                                  *H1-->
CATCTTAATG TACATCACAT CACTCTTTCA AGTTTTGCCA CAACAAGTGA GCATTGAAGC CATTCAAGAA GTGGAAATGT TGCCCAGGAC ATCTTCAAAA GTAACTAGAG AAGAACATTT TCAATTACAT CACCAGATGC ATTACTCTCA 1050
 I L M  Y I T  S L P  Q V L P  Q Q V  S I E A  I Q E  V E M  L P R T  S S K  V T R  E E H F  Q L H  H Q M  T Q S
   *E9                                                                                                                                  *E10
ACAGATCACA GTCAGTCTAG CACAGGGCTA TGAACAAACT TCTTCATCTC CTAAGCCTCG ATTCAAGAGT TATGCCTTCA CACAGGCTGC TTATGTTGCC ACCTCTGATT CCACACAGAG CCCCTATCCT TCACAGCATT TGGAAGCTCC 1200
 Q I T  V S L  A Q G Y  E Q T  S S S  P K P R  F K S  Y A F  T Q A A  Y V A  T S D  S T Q S  P Y P  S Q H  L E A P
<--H1*
CAGAGACAAG TCACTTGACA GTTCATTGAT GGAGACGGAA GTAAATCTGG ATAGTTACCA AACTGCTTTA GAAGAAGTAC TTTCATGGCT TCTTTCTGCC GAGGATACAT TGCGAGCACA AGGAGAGATT TCAAATGATG TTGAAGAAGT 1350
 R D K  S L D  S S L M  E T E  V N L  D S Y Q  T A L  E E V  L S W L  L S A  E D T  L R A Q  G E I  S N D  V E E V
                  *E11
GAAAGAACAG TTTCATGCTC ATGAGGGATT CATGATGGAT CTGACATCTC ATCAAGGACT TGTTGGTAAT GTTCTACAGT TAGGAAGTCA ACTAGTTGGA AAAGGGAAAT TATCAGAAGA TGAAGAAGCT GAAGTGCAAG AACAAATGAA 1500
 K E Q  F H A  H E G F  M M D  L T S  H Q G L  V G N  V L Q  L G S Q  L V G  K G K  L S E D  E E A  E V Q  E Q M N
                                                                  *E12
TCTCCTAAAT TCAAGATGGG AATGTCTCAG GGTAGCTAGC ATGGAAAAAC AAAGCAAATT ACACAAAGTT CTAATGGATC TCCAGAATCA GAAATTAAAA GAACTAGATG ACTGGTTAAC AAAAACTGAA GAGAGAACTA AGAAAATGGA 1650
 L L N  S R W  E C L R  V A S  M E K  Q S K L  H K V  L M D  L Q N Q  K L K  E L D  D W L T  K T E  E R T  K K M E
                                                                  *E13
GGAAGAGCCC TTTGGACCTG ATCTTGAAGA TCTAAAATGC CAAGTACAAC AACATAAGGT GCTTCAAGAA GATCTAGAAC AGGAGCAGGT CAGGGTCAAC TCGCTCACTC ACATGGTAGT AGTGGTTGAT GAATCCAGCG GTGATCATGC 1800
 E E P  F G P  D L E D  L K C  Q V Q  Q H K V  L Q E  D L E  Q E Q V  R V N  S L T  H M V V  V V D  E S S  G D H A
                  *E14                                                                                                                                  *E15
AACAGCTGCT TTGGAAGAAC AACTTAAGGT ACTGGGAGAT CGATGGGCAA ATATCTGCAG ATGGACTGAA GACCGCTGGA TTGTTTTACA AGATATTCTT CTAAAATGGC AGCATTTTAC TGAAGAACAG TGCCTTTTTA GTACATGGCT 1950
 T A A  L E E  Q L K V  L G D  R W A  N I C R  W T E  D R W  I V L Q  D I L  L K W  Q H F T  E E Q  C L F  S T W L
                                                                                  *E16
TTCAGAAAAA GAAGATGCAA TGAAGAACAT TCAGACAAGT GGCTTTAAAG ATCAAAATGA AATGATGTCA AGTCTTCACA AAATATCTAC TTTAAAAATA GATCTAGAAA AGAAAAAGCC AACCATGGAA AAACTAAGTT CACTCAATCA 2100
 S E K  E D A  M K N I  Q T S  G F K  D Q N E  M M S  S L H  K I S T  L K I  D L E  K K K P  T M E  K L S  S L N Q
                                                                                                                                  *E17      *H2-->
AGATCTACTT TCGGCACTGA AAAATAAGTC AGTGACTCAA AAGATGGAAA TCTGGATGGA AAACTTTGCA CAACGTTGGG ACAATTTAAC CCAAAAACTT GAAAAGAGTT CAGCACAAAT TTCACAGGCT GTCACCACCA CTCAACCATC 2250
 D L L  S A L  K N K S  V T Q  K M E  I W M E  N F A  Q R W  D N L T  Q K L  E K S  S A Q I  S Q A  V T T  T Q P S
                                                                                                                                  <--H2*                  *E18
CCTAACACAG ACAACTGTAA TGGAAACGGT AACTATGGTG ACCACAAGGG AACAAATCAT GGTAAAACAT GCCCAAGAGG AACTTCCACC ACCACCTCCT CAAAAGAAGA GGCAGATAAC TGTGGATTCT GAACTCAGGA AAAGGTTGGA 2400
 L T Q  T T V  M E T V  T M V  T T R  E Q I M  V K H  A Q E  E L P P  P P P  Q K K  R Q I T  V D S  E L R  K S H D
                                                                                                                                  *E19
TGTCGATATA ACTGAACTTC ACAGTTGGAT TACTCGTTCA GAAGCTGTAT TACAGAGTTC TGAATTTGCA GTCTATCGAA AAGAAGGCAA CATCTCAGAC TTGCAAGAAA AAGTCAATGC CATAGCACGA GAAAAAGCAG AGAAGTTCAG 2550
 V D I  T E L  H S W I  T R S  E A V  L Q S S  E F A  V Y R  K E G N  I S D  L Q E  K V N A  I A R  E K A  E K F R
                                                  *E20
AAAACTGCAA GATGCCAGCA GATCAGCTCA GGCCCTGGTG GAACAGATGG CAAATGAGGG TGTTAATGCT GAAAGTATCA GACAAGCTTC AGAACAACTG AACAGCCGGT GGACAGAATT CTGCCAATTG CTGAGTGAGA GAGTTAACTG 2700
 K L Q  D A S  R S A Q  A L V  E Q M  A N E G  V N A  E S I  R Q A S  E Q L  N S R  W T E F  C Q L  L S E  R V N W
                                                                                                                                                  *E21
GCTAGAGTAT CAAACCAACA TCATTACCTT TTATAATCAG CTACAACAAT TGGAACAGAT GACAACTACT GCCGAAAACT TGTTGAAAAC CCAGTCTACC ACCCTATCAG AGCCAACAGC AATTAAAAGC CAGTTAAAAA TTTGTAAGGA 2850
 L E Y  Q T N  I I T F  Y N Q  L Q Q  L E Q M  T T T  A E N  L L K T  Q S T  T L S  E P T A  I K S  Q L K  I C K D
                                                  *E22
TGAAGTCAAC AGATTGTCAG CTCTTCAGCC TCAAATTGAG CAATTAAAAA TTCAGAGTCT ACAACTGAAA GAAAAGGGAC AGGGGCCAAT GTTTCTGGAT GCAGACTTTG TGGCCTTTAC TAATCATTTT AACCACATCT TTGATGGTGT 3000
 E V N  R L S  A L Q P  Q I E  Q L K  I Q S L  Q L K  E K G  Q G P M  F L D  A D F  V A F T  N H F  N H I  F D G V
                                                  *E23
GAGGGCCAAA GAGAAAGAGC TACAGACAAT TTTTGACACT TTACCACCAA TGCGCTATCA GGAGACAATG AGTAGCATCA GGACGTGGAT CCAGCAGTCA GAAAGCAAAC TCTCTGTACC TTATCTTAGT GTTACTGAAT ATGAAATAAT 3150
 R A K  E K E  L Q T I  F D T  L P P  M R Y Q  E T M  S S I  R T W I  Q Q S  E S K  L S V P  Y L S  V T E  Y E I M
                                                  *E24
GGAGGAGAGA CTCGGGAAAT TACAGGCTCT GCAAAGTTCT TTGAAAGAGC AACAAAATGG CTTCAACTAT CTGAGTGACA CTGTGAAGGA GATGGCCAAG AAAGCACCTT CAGAAATATG CCAGAAATAT CTGTCAGAAT TTGAAGAGAT 3300
 E E R  L G K  L Q A L  Q S S  L K E  Q Q N G  F N Y  L S D  T V K E  M A K  K A P  S E I C  Q K Y  L S E  F E E I
                                                  *E25
TGAGGGGCAC TGGAAGAAAC TTTCCTCCCA GTTGGTGGAA AGCTGCCAAA AGCTAGAAGA ACATATGAAT AAACTTCGAA AATTTCAGAA TCACATAAAA ACCTTACAGA AATGGATGGC TGAAGTTGAT GTTTTCCTGA AAGAGGAATG 3450
 E G H  W K K  L S S Q  L V E  S C Q  K L E E  H M N  K L R  K F Q N  H I K  T L Q  K W M A  E V D  V F L  K E E W
                                                  *E26
GCCTGCCCTG GGGGATGCTG AAATCCTGAA AAAACAGCTC AAACAATGCA GACTTTTAGT TGGTGATATT CAAACAATTC AGCCCAGTTT AAATAGTGTT AATGAAGGTG GGCAGAAGAT AAAGAGTGAA GCTGAACTTG AGTTTGCATC 3600
 P A L  G D A  E I L K  K Q L  K Q C  R L L V  G D I  Q T I  Q P S L  N S V  N E G  G Q K I  K S E  A E L  E F A S
                                                  *E27
CAGACTGGAG ACAGAACTTA GAGAGCTTAA CACTCAGTGG GATCACATAT GCCGCCAGGT CTACACCAGA AAGGAAGCCT TAAAGGCAGG TTTGGATAAA ACCGTAAGCC TCCAAAAAGA TCTATCAGAG ATGCATGAGT GGATGACACA 3750
 R L E  T E L  R E L N  T Q W  D H I  C R Q V  Y T R  K E A  L K A G  L D K  T V S  L Q K D  L S E  M H E  W M T Q
                                                  *E28
AGCTGAAGAA GAATATCTAG AGAGAGATTT TGAATATAAA ACTCCAGATG AATTACAGAC TGCTGTTGAA GAAATGAAGA GAGCTAAAGA AGAGGCACTA CAAAAAGAAA CTAAAGTGAA ACTCCTTACT GAGACTGTAA ATAGTGTAAT 3900
 A E E  E Y L  E R D F  E Y K  T P D  E L Q T  A V E  E M K  R A K E  E A L  Q K E  T K V K  L L T  E T V  N S V I
```

```
                                                                                                                      *E28
AGCTCACGCT CCACCCTCAG CACAAGAGGC CTTAAAAAAG GAACTTGAAA CTCTGACCAC CAACTACCAA TGGCTGTGCA CCAGGCTGAA TGGAAAATGC AAAACTTTGG AAGAAGTTTG GGCATGTTGG CATGAGTTAT TGTCATATTT 4050
 A  H  A  P  P  S  A  Q  E  A  L  K  K  E  L  E  T  L  T  T  N  Y  Q  W  L  C  T  R  L  N  G  K  C  K  T  L  E  E  V  W  A  C  W  H  E  L  L  S  Y  L
                                                                                                     *E29
AGAGAAAGCA AACAAGTGGC TCAATGAAGT AGAATTGAAA CTTAAAACCA TGGAAAATGT TCCTGCAGGA CCTGAGGAAA TCACTGAAGT GCTAGAATCT CTTGAAAATC TGATGCATCA TTCAGAGGAG AACCCAAATC AGATTCGTCT 4200
 E  K  A  N  K  W  L  N  E  V  E  L  K  L  K  T  M  E  N  V  P  A  G  P  E  E  I  T  E  V  L  E  S  L  E  N  L  M  H  H  S  E  E  N  P  N  Q  I  R  L
                                                                                                     *E30
ATTGGCACAG ACTCTTACAG ATGGAGGAGT CATGGATGAA CTGATCAATG AGGAGCTTGA GACGTTTAAT TCTCGTTGGA GGGAACTACA TGAAGAGGCT GTGAGGAAAC AAAAGTTGCT TGAACAGAGT ATCCAGTCTG CCCAGGAAAT 4350
 L  A  Q  T  L  T  D  G  G  V  M  D  E  L  I  N  E  E  L  E  T  F  N  S  R  W  R  E  L  H  E  E  A  V  R  K  Q  K  L  L  E  Q  S  I  Q  S  A  Q  E  I
                                                                                                                *E31
TGAAAAGTCC TTGCACTTAA TTCAGGAGTC GCTTGAATTC ATTGACAAGC AGTTGGCAGC TTATATCACT GACAAGGTGG ATGCAGCTCA AATGCCTCAG GAAGCCCAGA AAATCCAATC AGATTTGACA AGTCATGAGA TAAGTTTAGA 4500
 E  K  S  L  H  L  I  Q  E  S  L  E  F  I  D  K  Q  L  A  A  Y  I  T  D  K  V  D  A  A  Q  M  P  Q  E  A  Q  K  I  Q  S  D  L  T  S  H  E  I  S  L  E
                                                                       *E32
AGAAATGAAG AAACATAACC AGGGGAAGGA TGCCAACCAA AGGGTTCTTT CACAAATTGA TGTTGCACAG AAAAAATTAC AAGATGTCTC CATGAAATTT CGATTATTCC AAAAACCAGC CAATTTTGAA CAACGTCTAG AGGAAAGTAA 4650
 E  M  K  K  H  N  Q  G  K  D  A  N  Q  R  V  L  S  Q  I  D  V  A  Q  K  K  L  Q  D  V  S  M  K  F  R  L  F  Q  K  P  A  N  F  E  Q  R  L  E  E  S  K
                                                                                          *E33
GATGATTTTA GATGAAGTCA AGATGCATTT GCCTGCATTG GAAACCAAGA GTGTTGAACA GGAAGTAATT CAGTCACAAC TAAGTCATTG TGTGAACTTG TATAAAAGCC TGAGTGAAGT CAAGTCTGAA GTGGAAATGG TGATTAAAAC 4800
 M  I  L  D  E  V  K  M  H  L  P  A  L  E  T  K  S  V  E  Q  E  V  I  Q  S  Q  L  S  H  C  V  N  L  Y  K  S  L  S  E  V  K  S  E  V  E  M  V  I  K  T
                                                                                                     *E34
CGGACGTCAA ATTGTACAGA AAAAGCAGAC AGAAAATCCC AAAGAGCTTG ATGAACGAGT AACAGCTTTG AAATTGCATT ACAATGAGTT GGGTGCGAAG GTAACAGAGA GAAAGCAACA GTTGGAGAAA TGCTTGAAGT TGTCCCGTAA 4950
 G  R  Q  I  V  Q  K  K  Q  T  E  N  P  K  E  L  D  E  R  V  T  A  L  K  L  H  Y  N  E  L  G  A  K  V  T  E  R  K  Q  Q  L  E  K  C  L  K  L  S  R  K
                                                                                                                  *E35
GATGAGAAAG GAAATGAATG TCTTAACAGA ATGGCTGGCA GCAACAGATA CAGAATTGAC GAAGAGATCA GCAGTTGAAG GAATGCCAAG TAATTTGGAT TCTGAAGTTG CCTGGGGAAA GGCTACTCAA AAAGAGATTG AGAAACAGAA 5100
 M  R  K  E  M  N  V  L  T  E  W  L  A  A  T  D  T  E  L  T  K  R  S  A  V  E  G  M  P  S  N  L  D  S  E  V  A  W  G  K  A  T  Q  K  E  I  E  K  Q  K

GGCTCACTTG AAGAGTGTTA CAGAATTAGG AGAGTCTTTG AAAATGGTGT TGGGCAAGAA AGAAACCTTG GTAGAAGATA AACTGAGTCT TCTGAACAGT AACTGGATAG CTGTCACCTC CAGAGTAGAA GAATGGCTAA ATCTTTTGTT 5250
 A  H  L  K  S  V  T  E  L  G  E  S  L  K  M  V  L  G  K  K  E  T  L  V  E  D  K  L  S  L  L  N  S  N  W  I  A  V  T  S  R  V  E  E  W  L  N  L  L
*E36                                                                                                                                  *E37
GGAATACCAG AAACACATGG AAACCTTTGA TCAGAACATA GAACAAATCA CAAAGTGGAT CATTCATGCA GATGAACTTT TAGATGAGTC TGAAAAGAAG AAACCACAAC AAAAGGAAGA CATTCTTAAG CGTTTAAAGG CTGAAATGAA 5400
 E  Y  Q  K  H  M  E  T  F  D  Q  N  I  E  Q  I  T  K  W  I  I  H  A  D  E  L  L  D  E  S  E  K  K  K  P  Q  Q  K  E  D  I  L  K  R  L  K  A  E  M  N

TGACATGCGC CCAAAGGTGG ACTCCACACG TGACCAAGCA GCAAAATTGA TGGCAAACCG CGGTGACCAC TGCAGGAAAG TAGTAGAGCC CCAAATCTCT GAGCTCAACC GTCGATTTGC AGCTATTTCT CACAGAATTA AGACTGGAAA 5550
 D  M  R  P  K  V  D  S  T  R  D  Q  A  A  K  L  M  A  N  R  G  D  H  C  R  K  V  V  E  P  Q  I  S  E  L  N  R  R  F  A  A  I  S  H  R  I  K  T  G  K
*E38                                                                                                                        *E39
GGCCTCCATT CCTTTGAAGG AATTGGAGCA GTTTAACTCA GATATACAAA AATTGCTTGA ACCACTGGAG GCTGAAATTC AGCAGGGGGT GAATCTGAAA GAGGAAGACT TCAATAAAGA TATGAGTGAA GACAATGAGG GTACTGTAAA 5700
 A  S  I  P  L  K  E  L  E  Q  F  N  S  D  I  Q  K  L  L  E  P  L  E  A  E  I  Q  Q  G  V  N  L  K  E  E  D  F  N  K  D  M  S  E  D  N  E  G  T  V  N
                                                                                                                *E40
TGAATTGTTG CAAAGAGGAG ACAACTTACA ACAAAGAATC ACAGATGAGA GAAAGCGAGA GGAAATAAAG ATAAAACAGC AGCTGTTACA GACAAAACAT AATGCTCTCA AGGATTTGAG GTCTCAAAGA AGAAAAAAGG CCCTAGAAAT 5850
 E  L  L  Q  R  G  D  N  L  Q  Q  R  I  T  D  E  R  K  R  E  E  I  K  I  K  Q  Q  L  L  Q  T  K  H  N  A  L  K  D  L  R  S  Q  R  R  K  K  A  L  E  I
                                                                                                                   *E41
TTCTCACCAG TGGTATCAGT ACAAGAGGCA GGCTGATGAT CTCCTGAAAT GCTTGGATGA AATTGAAAAA AAATTAGCCA GCCTACCTGA ACCCAGAGAT GAAAGAAAAT TAAAGGAAAT TGATCGTGAA TTGCAGAAGA AGAAAGAGGA 6000
 S  H  Q  W  Y  Q  Y  K  R  Q  A  D  D  L  L  K  C  L  D  E  I  E  K  K  L  A  S  L  P  E  P  R  D  E  R  K  L  K  E  I  D  R  E  L  Q  K  K  K  E  E
                                                                                                                                                *E42
GCTGAATGCA GTGCGCAGGC AAGCTGAGGG CTTGTCTGAG AATGGGGCCG CAATGGCAGT GGAGCCAACT CAGATCCAGC TCAGCAAGCG CTGGCGGCAA ATTGAGAGCA ATTTTGCTCA GTTTCGAAGA CTCAACTTTG CACAAATTCA 6150
 L  N  A  V  R  R  Q  A  E  G  L  S  E  N  G  A  A  M  A  V  E  P  T  Q  I  Q  L  S  K  R  W  R  Q  I  E  S  N  F  A  Q  F  R  R  L  N  F  A  Q  I  H

CACTCTCCAT GAAGAAACTA TGGTAGTGAC GACTGAAGAT ATGCCTTTGG ATGTTTCTTA TGTGCCTTCT ACTTATTTGA CCGAGATCAG TCATATCTTA CAAGCTCTTT CAGAAGTTGA TCATCTTCTA AATACTCCTG AACTCTGTGC 6300
 T  L  H  E  E  T  M  V  V  T  T  E  D  M  P  L  D  V  S  Y  V  P  S  T  Y  L  T  E  I  S  H  I  L  Q  A  L  S  E  V  D  H  L  L  N  T  P  E  L  C  A
                                          *E43
TAAAGATTTT GAAGATCTTT TTAAGCAAGA GGAGTCTCTT AAGAATATAA AAGACAATTT GCAACAAATC TCAGGTCGGA TTGATATTAT TCACAAGAAG AAGACAGCAG CCTTGCAAAG TGCCACCTCC ATGGAAAAGG TGAAAGTACA 6450
 K  D  F  E  D  L  F  K  Q  E  E  S  L  K  N  I  K  D  N  L  Q  Q  I  S  G  R  I  D  I  I  H  K  K  K  T  A  A  L  Q  S  A  T  S  M  E  K  V  K  V  Q
                                                               *E44
GGAAGCCGTG GCACAGATGG ATTTCCAGGG GGAAAAACTT CATAGAATGT ACAAGGAACG ACAAGGGCGA TTCGACAGAT CAGTTGAAAA ATGGCGACAC TTTCATTATG ATATGAAGGT ATTTAATCAA TGGCTGAATG AAGTTGAACA 6600
 E  A  V  A  Q  M  D  F  Q  G  E  K  L  H  R  M  Y  K  E  R  Q  G  R  F  D  R  S  V  E  K  W  R  H  F  H  Y  D  M  K  V  F  N  Q  W  L  N  E  V  E  Q
                                                              *E45
GTTTTTCAAA AAGACACAAA ATCCTGAAAA CTGGGAACAT GCTAAATACA AATGGTATCT TAAGGAACTC CAGGATGGCA TTGGGCAGCG TCAAGCTGTT GTCAGAACAC TGAATGCAAC TGGGGAAGAA ATAATTCAAC AGTCTTCAAA 6750
 F  F  K  K  T  Q  N  P  E  N  W  E  H  A  K  Y  K  W  Y  L  K  E  L  Q  D  G  I  G  Q  R  Q  A  V  V  R  T  L  N  A  T  G  E  E  I  I  Q  Q  S  S  K
                                                                                           *E46
AACAGATGTC AATATTCTAC AAGAAAAATT AGGAAGCTTG AGTCTGCGGT GGCACGACAT CTGCAAAGAG CTGGCAGAAA GGAGAAAGAG GATTGAAGAA CAAAAGAATG TCTTGTCAGA ATTTCAAAGA GATTTAAATG AATTTGTTTT 6900
 T  D  V  N  I  L  Q  E  K  L  G  S  L  S  L  R  W  H  D  I  C  K  E  L  A  E  R  R  K  R  I  E  E  Q  K  N  V  L  S  E  F  Q  R  D  L  N  E  F  V  L
                                                                                      *E47
GTGGCTGGAA GAAGCAGATA ACATTGCTAT TACTCCACTT GGAGATGAGC AGCAGCTAAA AGAACAACTT GAACAAGTCA AGTTACTGGC AGAAGAGTTG CCCCTGCGCC AGGGAATTCT AAAACAATTA AATGAAACAG GAGGAGCAGT 7050
 W  L  E  E  A  D  N  I  A  I  T  P  L  G  D  E  Q  Q  L  K  E  Q  L  E  Q  V  K  L  L  A  E  E  L  P  L  R  Q  G  I  L  K  Q  L  N  E  T  G  G  A  V
                                                                                      *E48
ACTTGTAAGT GCTCCCATAA GGCCAGAAGA GCAAGATAAA CTTGAAAAGA AGCTCAAACA GACAAATCTC CAGTGGATAA AGGTCTCCAG AGCTTTACCT GAGAAACAAG GAGAGCTTGA GGTTCACTTA AAAGATTTTA GGCAGCTTGA 7200
 L  V  S  A  P  I  R  P  E  E  Q  D  K  L  E  K  K  L  K  Q  T  N  L  Q  W  I  K  V  S  R  A  L  P  E  K  Q  G  E  L  E  V  H  L  K  D  F  R  Q  L  E
                                                                                                    *E49
AGAGCAGCTG GATCACCTGC TTCTGTGGCT CTCTCCTATT AGAAACCAGT TGGAAATTTA TAACCAACCA AGTCAGGCAG GACCGTTTGA CATAAAGGAG ATTGAAGTAA CAGTTCACGG TAAACAAGCG GATGTGGAAA GGCTTTTGTC 7350
 E  Q  L  D  H  L  L  L  W  L  S  P  I  R  N  Q  L  E  I  Y  N  Q  P  S  Q  A  G  P  F  D  I  K  E  I  E  V  T  V  H  G  K  Q  A  D  V  E  R  L  L  S
                                                   *E50                                                                            *H3-->
GAAAGGGCAG CATTTGTATA AGGAAAAACC AAGCACTCAG CCAGTGAAGA GGAAGTTAGA AGATCTGAGG TCTGAGTGGG AGGCTGTAAA CCATTTACTT CGGGAGCTGA GGACAAAGCA GCCTGACCGT GCCCCTGGAC TGAGCACTAC 7500
 K  G  Q  H  L  Y  K  E  K  P `S  T  Q  P  V  K  R  K  L  E  D  L  R  S  E  W  E  A  V  N  H  L  L  R  E  L  R  T  K  Q  P  D  R  A  P  G  L  S  T  T
     *E51                                                                                                                <--H3*
TGGAGCCTCT GCCAGTCAGA CTGTTACTCT AGTGACACAA TCTGTGGTTA CTAAGGAAAC TGTCATCTCC AAACTAGAAA TGCCATCTTC TTTGCTGTTG GAGGTACCTG CACTGGCAGA CTTCAACCGA GCTTGGACAG AACTTACAGA 7650
 G  A  S  A  S  Q  T  V  T  L  V  T  Q  S  V  V  T  K  E  T  V  I  S  K  L  E  M  P  S  S  L  L  L  E  V  P  A  L  A  D  F  N  R  A  W  T  E  L  T  D
```

Figure 1.2

```
                                                                                                 *E52
CTGGCTGTCT CTGCTTGATC GAGTTATAAA ATCACAGAGA GTGATGGTGG GTGATCTGGA AGACATCAAT GAAATGATCA TCAAACAGAA GGCAACACTG CAAGATTTGG AACAGAGACG CCCCCAATTG GAAGAACTCA TTACTGCTGC 7800
 W  L  S  L  L  D  R  V  I  K  S  Q  R  V  M  V  G  D  L  E  D  I  N  E  M  I  I  K  Q  K  A  T  L  Q  D  L  E  Q  R  R  P  Q  L  E  E  L  I  T  A  A
                                                              *E53
CCAGAATTTG AAAAACAAAA CCAGCAATCA AGAAGCTAGA ACAATCATTA CTGATCGAAT TGAAAGAATT CAGATTCAGT GGGATGAGGT TCAAGAACAG CTGCAGAACA GGAGACAACA GTTGAATGAA ATGTTAAAGG ATTCAACACA 7950
 Q  N  L  K  N  K  T  S  N  Q  E  A  R  T  I  I  T  D  R  I  E  R  I  Q  I  Q  W  D  E  V  Q  E  Q  L  Q  N  R  R  Q  Q  L  N  E  M  L  K  D  S  T  Q
                                                                                                                                 *E54
ATGGCTGGAA GCTAAGGAAG AAGCCGAACA GGTCATAGGA CAGGTCAGAG GCAAGCTTGA CTCATGGAAA GAAGGTCCTC ACACAGTAGA TGCAATCCAA AAGAAGATCA CAGAAACCAA GCAGTTGGCC AAAGACCTCC GTCAACGGCA 8100
 W  L  E  A  K  E  E  A  E  Q  V  I  G  Q  V  R  G  K  L  D  S  W  K  E  G  P  H  T  V  D  A  I  Q  K  K  I  T  E  T  K  Q  L  A  K  D  L  R  Q  R  Q
                                                                                                                                     *E55
GATAAGTGTA GACGTGGCAA ATGATTTGGC ACTGAAACTT CTTCGGGACT ATTCTGCTGA TGATACCAGA AAAGTACACA TGATAACAGA GAATATCAAT ACTTCTTGGG GAAACATTCA TAAAAGAGTA AGTGAGCAAG AGGCTGCTTT 8250
 I  S  V  D  V  A  N  D  L  A  L  K  L  L  R  D  Y  S  A  D  D  T  R  K  V  H  M  I  T  E  N  I  N  T  S  W  G  N  I  H  K  R  V  S  E  Q  E  A  A  L

GGAAGAAACT CATAGATTAC TGCAGCAGTT CCCTCTGGAC CTGGAGAAGT TTCTTTCCTG GATTACGGAA GCAGAAACAA CTGCCAATGT CCTACAGGAC GCTTCCCGTA AGGAGAAGCT CCTAGAAGAC TCCAGGGGAG TCAGAGAGCT 8400
 E  E  T  H  R  L  L  Q  Q  F  P  L  D  L  E  K  F  L  S  W  I  T  E  A  E  T  T  A  N  V  L  Q  D  A  S  R  K  E  K  L  L  E  D  S  R  G  V  R  E  L
               *E56
GATGAAACCA TGGCAAGATC TCCAAGGAGA AATTGAAACT CACACAGATA TCTATCACAA TCTTGATGAA AATGGCCAAA AAATCCTGAG ATCCCTGGAA GGTTCGGATG AAGCACCCCT GTTACAAAGA CGTTTGGATA ACATGAATTT 8550
 M  K  P  W  Q  D  L  Q  G  E  I  E  T  H  T  D  I  Y  H  N  L  D  E  N  G  Q  K  I  L  R  S  L  E  G  S  D  E  A  P  L  L  Q  R  R  L  D  N  M  N  F
                                        *E57
CAAGTGGAGT GAACTTCAGA AAAAGTCTCT CAACATTAGG TCCCATTTGG AAGCAAGTTC TGACCAGTGG AAGCGTTTGC ATCTTTCTCT TCAGGAACTT CTTGTTTGGC TACAGCTGAA AGATGATGAA CTGAGCCGTC AGGCACCCAT 8700
 K  W  S  E  L  Q  K  K  S  L  N  I  R  S  H  L  E  A  S  S  D  Q  W  K  R  L  H  L  S  L  Q  E  L  L  V  W  L  Q  L  K  D  D  E  L  S  R  Q  A  P  I
                                                *E58
CGGTGGTGAT TTCCCAGCAG TTCAGAAGCA GAATGATATA CATAGGGCCT TCAAGAGGGA ATTGAAAACT AAAGAACCTG TAATCATGAG TACTCTGGAG ACTGTGAGAA TATTTCTGAC AGAGCAGCCT TTGGAAGGAC TAGAGAAACT 8850
 G  G  D  F  P  A  V  Q  K  Q  N  D  I  H  R  A  F  K  R  E  L  K  T  K  E  P  V  I  M  S  T  L  E  T  V  R  I  F  L  T  E  Q  P  L  E  G  L  E  K  L
               *E59
CTACCAGGAG CCCAGAGAAC TGCCTCCTGA AGAAAGAGCT CAGAATGTCA CTCGGCTCCT ACGAAAGCAG GCTGAAGAGG TCAACGCTGA ATGGGACAAA TTGAACCTGC GCTCAGCTGA TTGGCAGAGA AAAATAGATG AAGCTCTTGA 9000
 Y  Q  E  P  R  E  L  P  P  E  E  R  A  Q  N  V  T  R  L  L  R  K  Q  A  E  E  V  N  A  E  W  D  K  L  N  L  R  S  A  D  W  Q  R  K  I  D  E  A  L  E
                                                                                                                                            *E60
AAGACTCCAG GAACTTCAGG AAGCTGCCGA TGAACTGGAC CTCAAGTTGC GCCAAGCTGA GGTGATCAAG GGATCCTGGC AGCCAGTGGG GGATCTCCTC ATTGACTCTC TGCAAGATCA CCTTGAAAAA GTCAAGGCAC TTCGGGGAGA 9150
 R  L  Q  E  L  Q  E  A  A  D  E  L  D  L  K  L  R  Q  A  E  V  I  K  G  S  W  Q  P  V  G  D  L  L  I  D  S  L  Q  D  H  L  E  K  V  K  A  L  R  G  E
                                                                                                                                         *E61
AATTGCACCT CTTAAAGAGA ATGTCAATCG TGTCAATGAC CTTGCACATC AGCTGACCAC ACTGGGCATT CAGCTCTCAC CTTATAACCT CAGCACTTTG GAAGATCTGA ATACCAGATG GAGGCTTCTA CAGGTGGCTG TGGAGGACCG 9300
 I  A  P  L  K  E  N  V  N  R  V  N  D  L  A  H  Q  L  T  T  L  G  I  Q  L  S  P  Y  N  L  S  T  L  E  D  L  N  T  R  W  R  L  L  Q  V  A  V  E  D  R
                  *H4-->                                            *E62 (WW->)                                                         *E63
TGTCAGACAG CTGCATGAAG CCCACAGGGA CTTTGGTCCT GCATCCCAGC ACTTCCTTTC CACTTCAGTT CAGGGTCCCT GGGAGAGAGC CATCTCACCA AACAAAGTGC CCTACTATAT CAACCACGAG ACCCAAACCA CTTGTTGGGA 9450
 V  R  Q  L  H  E  A  H  R  D  F  G  P  A  S  Q  H  F  L  S  T  S  V  Q  G  P  W  E  R  A  I  S  P  N  K  V  P  Y  Y  I  N  H  E  T  Q  T  T  C  W  D
                      <WW*        *E64                                       <--H4*                    *E65
CCACCCCAAA ATGACAGAGC TCTACCAGTC TTTAGCTGAC CTGAATAATG TCAGGTTCTC CGCGTATAGG ACTGCCATGA AGCTCAGAAG GCTCCAGAAG GCCCTTTGCT TGGATCTCTT GAGCCTGTCA GCTGCATGTG ATGCCCTGGA 9600
 H  P  K  M  T  E  L  Y  Q  S  L  A  D  L  N  N  V  R  F  S  A  Y  R  T  A  M  K  L  R  R  L  Q  K  A  L  C  L  D  L  L  S  L  S  A  A  C  D  A  L  D

CCAGCACAAC CTCAAGCAAA ATGACCAGCC CATGGATATC CTGCAGATAA TTAACTGTTT GACTACAATT TATGATCGTC TGGAGCAAGA GCACAACAAT CTGGTCAATG TCCCTCTCTG TGTGGATATG TGTCTCAACT GGCTTCTCAA 9750
 Q  H  N  L  K  Q  N  D  Q  P  M  D  I  L  Q  I  I  N  C  L  T  T  I  Y  D  R  L  E  Q  E  H  N  N  L  V  N  V  P  L  C  V  D  M  C  L  N  W  L  L  N
            *E66                                                                                     *E67
TGTTTATGAT ACGGGACGAA CAGGGAGGAT CCGTGTCCTG TCTTTTAAAA CTGGCATCAT TTCTCTGTGT AAAGCACACT TGGAAGACAA GTACAGATAC CTTTTCAAGC AAGTGGCAAG TTCAACTGGC TTTTGTGACC AGCGTAGGCT 9900
 V  Y  D  T  G  R  T  G  R  I  R  V  L  S  F  K  T  G  I  I  S  L  C  K  A  H  L  E  D  K  Y  R  Y  L  F  K  Q  V  A  S  S  T  G  F  C  D  Q  R  R  L
                                                                                                             *E68
GGGTCTTCTT CTGCATGATT CTATTCAAAT CCCAAGACAG TTGGGTGAAG TTGCTTCCTT TGGGGGCAGT AACATTGAGC CGAGTGTCAG GAGCTGCTTC CAATTTGCCA ATAATAAACC TGAGATTGAA GCTGCTCTCT TCCTTGACTG 10050
 G  L  L  L  H  D  S  I  Q  I  P  R  Q  L  G  E  V  A  S  F  G  G  S  N  I  E  P  S  V  R  S  C  F  Q  F  A  N  N  K  P  E  I  E  A  A  L  F  L  D  W
                                                                                                                                    *E69
GATGCGCCTG GAACCCCAGT CTATGGTGTG GCTGCCCGTC TTGCACAGAG TGGCTGCTGC TGAAACTGCC AAGCATCAAG CCAAGTGTAA CATCTGTAAG GAGTGTCCAA TCATTGGATT CAGGTACAGA AGCCTAAAGC ATTTTAATTA 10200
 M  R  L  E  P  Q  S  M  V  W  L  P  V  L  H  R  V  A  A  A  E  T  A  K  H  Q  A  K  C  N  I  C  K  E  C  P  I  I  G  F  R  Y  R  S  L  K  H  F  N  Y
                                                                                           *E70
TGACATCTGC CAAAGTTGCT TTTTTTCTGG CCGAGTTGCA AAGGGCCATA AAATGCACTA CCCCATGGTA GAGTATTGCA CTCCGACTAC ATCCGGAGAA GATGTTCGCG ACTTCGCCAA GGTACTAAAA AACAAATTTC GAACCAAAAG 10350
 D  I  C  Q  S  C  F  F  S  G  R  V  A  K  G  H  K  M  H  Y  P  M  V  E  Y  C  T  P  T  T  S  G  E  D  V  R  D  F  A  K  V  L  K  N  K  F  R  T  K  R
                                                                                *E71                                 *E72
GTATTTTGCG AAGCATCCCC GAATGGGCTA CCTGCCAGTG CAGACTGTGT TAGAGGGGGA CAACATGGAA ACTCCCGTTA CTCTGATCAA CTTCTGGCCA GTAGATTCTG CGCCTGCCTC GTCCCCCCAG CTTTCACACG ATGATACTCA 10500
 Y  F  A  K  H  P  R  M  G  Y  L  P  V  Q  T  V  L  E  G  D  N  M  E  T  P  V  T  L  I  N  F  W  P  V  D  S  A  P  A  S  S  P  Q  L  S  H  D  D  T  H
                           *E73                         α-Syn B.D. -->                           *E74                                      <--α-Syn B.D.
TTCACGCATT GAACATTATG CTAGCAGGCT AGCAGAAATG GAAAACAGCA ATGGATCTTA TCTAAATGAT AGCATCTCTC CTAATGAGAG CATAGATGAT GAACATTTGT TAATCCAGCA TTACTGCCAA AGTTTGAACC AGGACTCCCC 10650
 S  R  I  E  H  Y  A  S  R  L  A  E  M  E  N  S  N  G  S  Y  L  N  D  S  I  S  P  N  E  S  I  D  D  E  H  L  L  I  Q  H  Y  C  Q  S  L  N  Q  D  S  P
                                                                                                                *E75         L-ZIP 1
CCTGAGCCAG CCTCGTAGTC CTGCCCAGAT CTTGATTTCC TTAGAGAGTG AGGAAAGAGG GGAGCTAGAG AGAATCCTAG CAGATCTTGA GGAAGAAAAC AGGAATCTGC AAGCAGAATA TGATCGCCTG AAGCAGCAGC ATGAGCATAA 10800
 L  S  Q  P  R  S  P  A  Q  I  L  I  S  L  E  S  E  E  R  G  E  L  E  R  I  L  A  D  L  E  E  N  R  N  L  Q  A  E  Y  D  R  L  K  Q  Q  H  E  H  K
                                                                                                                                  L-ZIP 2
AGGCCTGTCT CCACTGCCAT CTCCTCCTGA GATGATGCCC ACCTCTCCTC AGAGTCCCAG GGATGCTGAG CTCATTGCTG AGGCTAAGCT ACTGCGCCAA CACAAAGGAC GCCTGGAAGC CAGGATGCAA ATCCTGGAAG ACCACAATAA 10950
 G  L  S  P  L  P  S  P  P  E  M  M  P  T  S  P  Q  S  P  R  D  A  E  L  I  A  E  A  K  L  L  R  Q  H  K  G  R  L  E  A  R  M  Q  I  L  E  D  H  N  K
                                             *E76
ACAGCTGGAG TCTCAGTTAC ATAGACTGAG ACAGCTCCTG GAGCAGCCCC AGGCTGAAGC TAAGGTGAAT GGCACCACGG TGTCCTCTCC TTCCACCTCT CTGCAGAGGT CAGATAGCAG TCAGCCTATG CTGCTCCGAG TGGTTGGCAG 11100
 Q  L  E  S  Q  L  H  R  L  R  Q  L  L  E  Q  P  Q  A  E  A  K  V  N  G  T  T  V  S  S  P  S  T  S  L  Q  R  S  D  S  S  Q  P  M  L  L  R  V  V  G  S
                     *E77                                                                                                *E78                         *E79
TCAAACTTCA GAATCTATGG GTGAGGAAGA TCTTCTGAGT CCTCCCCAGG ACACAAGCAC AGGGTTAGAA GAAGTGATGG AGCAACTCAA CAACTCCTTC CCTAGTTCAA GAGGAAGAAA TGCCCCCGGA AAGCCAATGA GAGAGGACAC 11250
 Q  T  S  E  S  M  G  E  E  D  L  L  S  P  P  Q  D  T  S  T  G  L  E  E  V  M  E  Q  L  N  N  S  F  P  S  S  R  G  R  N  A  P  G  K  P  M  R  E  D  T
                                                                                                  alt C
AATGTAGGAA GCCTTTTCCA CATGGCAGAT GATTTGGGCA GAGCGATGGA GTCCTTAGTT TCAGTCATGA CAGATGAAGA AGGAGCAGAA TAAATGTTTT ACAACTCCTG ATTCCCGCAT GGTTTTTATA ATATTCGTAC AACAAAGAGG 11400
 M
```

Figure 1.2 (*continued*)

involving exons 71–74 are detected in cardiac muscle (Feener *et al.*, 1989; Bies *et al.*, 1992*b*), and alternative splicing of exon 68 leads to the generation of full-length and truncated dystrophin isoforms in smooth muscle (Feener *et al.*, 1989). At present, no specific role has been ascribed to the alternatively spliced isoforms of dystrophin. Ten per cent of DMD patients die from cardiac involvement and dystrophin is expressed not only in cardiac muscle but also in cardiac Purkinje fibres, suggesting that cardiac conduction abnormalities could significantly contribute to mortality in some patients (Bies *et al.*, 1992*b*). Several cases of X-linked cardiomyopathy (which has little skeletal muscle involvement) are caused by dystrophin gene mutations, and in some of these cases the mutation inactivates the muscle promoter (Towbin *et al.*, 1993; Muntoni *et al.*, 1995). Apparently the C and/or P promoters can compensate for this defect in skeletal, but not cardiac muscle (X-linked cardiomyopathy patients have not been tested for dystrophin expression in cardiac Purkinje fibres).

Although a large number of dystrophin isoforms are expressed in many cell types, few tissues besides striated muscle manifest clear phenotypic consequences of dystrophin gene mutations (Bushby, 1992; Emery, 1993). For example, cognitive deficits are found in about 30% of DMD/BMD patients, and this presumably arises from defective expression of dystrophin in the CNS (Emery, 1993). Mental retardation in DMD patients remains one of the least well understood features of the disease, primarily because there is no clear correlation between specific types or locations of mutations and whether or not a patient displays mental retardation. Cognitive defects generally co-segregate in families, and there have been reports of an association between mutations in the 3′ half of the gene and mental retardation (Bushby, 1992; Rapaport *et al.*, 1991). However, no pathological abnormalities have been observed in the CNS of DMD patients, and it is not clear which of the many dystrophin isoforms (full-length, Dp140, Dp116, or Dp71) are responsible for the cognitive problems. It seems likely that the interaction of multiple isoforms plays a role in this phenomenon.

Electroretinographs of mice and humans have shown that defective expression of dystrophin transcripts arising from internal promoters is associated with abnormal retinal electrophysiology (Pillers *et al.*, 1995). D'Souza *et al.* (1995) identified the Dp260 isoform of dystrophin in the outer plexiform layer (OPL) of the retina, a region that also expresses Dp71. Differential effects on b-wave timing, amplitude, and oscillatory potentials result from mutations in various parts of the gene in mice, suggesting that both isoforms might contribute to the abnormal retinograms observed in dark-adapted retinas (Pillers *et al.*, 1995).

The Dp140 isoforms is expressed from a promoter located in intron 44 and has been detected in various glial components of the CNS, where it may play a role in membrane specialization or attachment of astrocytes (Lidov *et al.*, 1995). It is not known whether mutation of this transcript contributes to any known pathology associated with DMD, although the 5′ end of Dp140 has been noted to overlap a region frequently deleted in patients with mental retardation (Rapaport *et al.*, 1991; Lidov *et al.*, 1995). Dp116 and/or Dp140 are also detected in a glial component surrounding olfactory nerves (Lidov *et al.*, 1995). The Dp116 isoform is expressed from a promoter located in intron 55, and is primarily found in the submembrane region of the external layer of Schwann cells in peripheral nerve (Byers *et al.*, 1993). No evidence has been presented for direct involvement of this isoform in Schwann cell dysfunction. Dp71 is expressed from a promoter in intron 62 and is found in almost every tissue except muscle (Rapaport *et al.*, 1992), but little is known of its potential function. Although most patients with C-terminal mutations (and the mdx^{3cv} mouse mutant) are defective for expression of all dystrophin isoforms, the retinal defects referred to above are the only non-muscle, clinical phenotypes *consistently* correlated with mutation location. The utilization of 5 different strains of *mdx* mice (Chapman *et al.*, 1989) that contain unique point mutations distributed throughout the dystrophin gene may help clarify the role of the various dystrophin transcripts in non-muscle tissues (Fig. 1.1) (Cox *et al.*, 1993; Im *et al.*, 1996).

Dystrophin domains

The dystrophin protein is predicted to fold into several distinct structural domains (see Chapter 2 and Koenig *et al.*, 1987, 1988). The amino-terminus contains an actin-binding domain encoded by exons 1–8. The majority of the molecule (encoded by exons 9–63) forms a large rod-like domain composed of 24 'spectrin-like' repeats interrupted by a few spacer regions. Further downstream is a cysteine-rich domain (exons 64–67) followed by the C-terminal region (exons 68–79), which together bind a large complex of membrane-associated proteins and glycoproteins known as the dystrophin associated protein (DAP) complex (see Chapter 3 and Ervasti *et al.*, 1990). The C-terminal domain can be further divided into several subdomains (see below).

The DAPs are present at greatly reduced levels in DMD and *mdx* muscle, and are stabilized by their interaction with dystrophin (Ervasti *et al.*, 1990). Although the complete function of dystrophin in skeletal muscle has not been fully elucidated, it is thought to provide a critical link between the subsarcolemmal cytoskeleton and the extracellular matrix (Ervasti & Camp-bell, 1993).