Hamster preproglucagon contains the sequence of glucagon and two related peptides

Graeme I. Bell, Robert F. Santerre* & Guy T. Mullenbach

Chiron Corporation, 4560 Horton Street, Emeryville, California 94608, USA and *Division of Molecular and Cell Biology, Lilly Research Laboratories, Indianapolis, Indiana 46285, USA

Glucagon is a 29-amino acid polypeptide hormone synthesized by the A cells of the endocrine pancreas[1–4]. Its primary site of action is the liver where it stimulates glycogenolysis, gluconeogenesis and ketogenesis. In mammals, biosynthetic studies have shown that glucagon is derived from a precursor of molecular weight (M_r) approximately 18,000 which is five to six times larger than glucagon[5]. Glucagon-containing polypeptides and immunoreactants of various sizes have also been described from stomach, intestine, brain and salivary gland[3]. Here, we have determined the structure of hamster pancreatic preproglucagon from the sequence of its cDNA. This 180-amino acid precursor contains the sequence of glucagon and two glucagon-like polypeptides arranged in tandem. The precursor also contains the sequences of several non-pancreatic glucagon-containing polypeptides which suggests that, in mammals, both pancreatic and non-pancreatic glucagon and glucagon-containing polypeptides may be derived from a common precursor by tissue-specific processing. We have tentatively identified each of the glucagon-like immunoreactants which have been described with respect to the sequence of proglucagon and have proposed a scheme for the processing of pancreatic proglucagon.

The structure of preproglucagon mRNA was deduced from the 1,118 base pair (bp) sequence of the insert in colony pshglu1 (Fig. 1). The sequence contains a single open reading frame beginning at the methionine codon at nucleotides 104–106 (the only one in any of the three frames) which predicts the sequence of the 180-amino acid preproglucagon. Thus, the coding region of hamster preproglucagon mRNA is 540 nucleotides. The 3′-untranslated region of the mRNA is 475 bases and contains two polyadenylation signals, AAUAAA[6,7], nucleotides 811–816 and 1,098–1,103, beginning 408 and 21 bases, respectively, before the poly(A) tract. The 5′-untranslated region is at least 103 nucleotides. Electrophoresis of glyoxylated islet RNA in an agarose gel[8] and subsequent hybridization[9] indicated that hamster preproglucagon mRNA is 1,250 bases (data not shown). As the cloned cDNA insert is 1,118 nucleotides exclus-

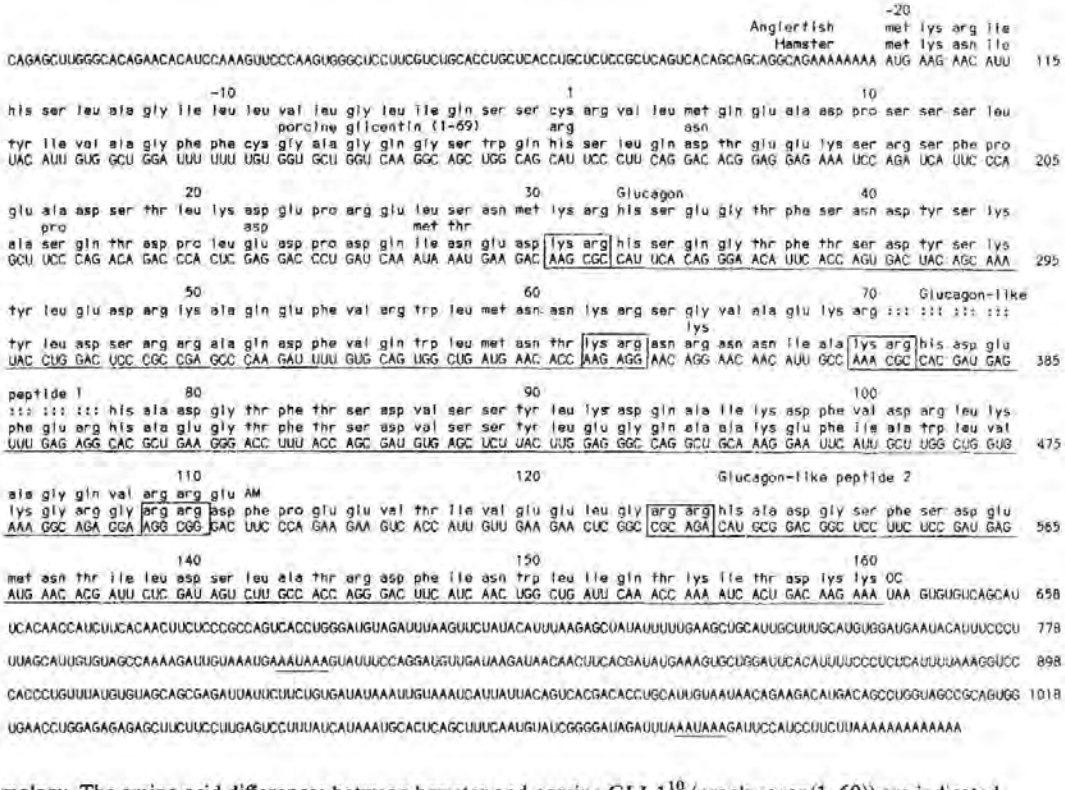

Fig. 1 Primary structure of hamster pancreatic preproglucagon mRNA and protein and comparison with anglerfish preproglucagon and porcine GLI-1. The predicted amino acid sequence of hamster preproglucagon is numbered by designating the first amino acid of hamster GLI-1 as 1. The amino acids constituting the signal peptide are given negative numbers. The basic dipeptides which may be involved in processing are boxed. The regions corresponding to glucagon and GLP-1 and -2 are indicated. The number of the nucleotide at the end of each line is indicated. The two AAUAAA sequences in the 3′-untranslated region are underlined. Hamster and anglerfish[14] preproglucagon are aligned and colons indicate gaps introduced to maximize homology. The amino acid differences between hamster and porcine GLI-1[10] (proglucagon(1–69)) are indicated.

Methods: RNA was prepared[23] from islets of Langerhans obtained from the pancreases of 400 female Syrian hamsters (*Mesocricetus auratus*)[24]. The tissue preparation was 50% islets and contained ~2×10^7 cells. The yield of poly(A)-containing RNA was 35 μg. Double-stranded cDNA was prepared as described by Land *et al.*[25] and inserted into the *Pst*I site of pBR322 using the GC-tailing technique[26]. Transformation of *Escherichia coli* strain HB101 generated a library of 25,000 tetracycline-resistant transformants; 1,000 of these were grown in arrays on Whatman 541 filter paper[27]. Following amplification of the plasmid DNA *in situ*, the colonies were screened for those containing glucagon sequences by hybridization[28] with a ^{32}P-labelled heptadecadeoxynucleotide[29] pool whose sequences 3′ GT(T/C) ACC (G/A)A(A/G/C/T) TAC TA(A/G) TG 5′ were complementary to all 32 possible sequences of mRNA encoding glucagon(24–29). Two colonies containing plasmids with inserts of 1,200 and 1,000 bp hybridized: pshglu1 and 2, respectively. The colonies were rescreened with the insert in pshglu1 and no additional hybridizing colonies were observed. The frequency of colonies containing plasmids coding for hamster insulin and somatostatin was also determined by hybridization with human insulin[30] and somatostatin[15] probes: 20/500 and 2/1,000 colonies hybridized with the insulin and somatostatin probes, respectively. The sequence of the insert in pshglu1 was determined by the procedure of Maxam and Gilbert[31] on both strands except for 165 bp and 210 bp at the 5′ and 3′ ends, respectively, and across all restriction sites used to initiate sequence determinations. A partial sequence of the insert in pshglu2 indicated no differences in the region from nucleotides 240 to 329.

```
                              his ser gln gly thr phe thr ser asp  Glucagon
his asp glu phe glu arg his ala glu gly thr phe thr ser asp        GLP-1
                         his ala asp gly ser phe ser asp glu       GLP-2
                         his ser asp gly thr phe thr ser glu       Secretin
                         his ser asp ala val phe thr asp asn       VIP
                         tyr ala glu gly thr phe ile ser asp       GIP
                         his ala asp gly val phe thr ser asp       PHI-27
                         tyr ala asp ala ile phe thr asn ser       pGRF

tyr ser lys tyr leu asp ser arg arg ala gln asp phe val gln        Glucagon
val ser ser tyr leu glu gly gln ala ala lys glu phe ile ala        GLP-1
met asn thr ile leu asp ser leu ala thr arg asp phe ile asn        GLP-2
leu ser arg leu arg asp ser ala arg leu gln arg leu leu gln        Secretin
tyr thr arg leu arg lys gln met ala val lys lys tyr leu asn        VIP
tyr ser ile ala met asp lys ile arg gln gln asp phe val asn        GIP
phe ser arg leu leu gly gln leu ser ala lys lys tyr leu glu        PHI-27
tyr arg lys val leu gly gln leu ser ala arg lys leu leu gln        pGRF

trp leu met asn thr                                               Glucagon
trp leu val lys gly arg gly                                       GLP-1
trp leu ile gln thr lys ile thr asp lys lys                       GLP-2
gly leu val                                                       Secretin
ser ile leu asn                                                   VIP
trp leu leu ala gln lys gly lys lys ser asp trp lys his asn        GIP
ser leu ile                                                       PHI-27
asp ile met ser arg gln gln gly glu ser asn gln glu arg gly        pGRF

ile thr gln                                                       GIP
ala arg ala arg leu                                               pGRF
```

Fig. 2 Comparison of the sequences of hamster glucagon-like polypeptides with other members of the glucagon–secretin family. The peptides are: human glucagon (all mammalian glucagons seem to have identical sequences)[32]; hamster GLP-1 and -2 (this paper); porcine secretin, vasoactive intestinal polypeptide (VIP) gastric inhibitory polypeptide (GIP) and PHI-[32–34]; and human pGRF[35].

ive of the poly(A) tract, it may represent a nearly full-length copy of the mRNA. Also, as an approximately 900-base form of hamster pancreatic preproglucagon mRNA was not observed, the proximal AAUAAA, nucleotides 811–816, is not a normal signal for polyadenylation.

The organization of hamster preproglucagon was determined by comparing its sequence with that of other glucagon-containing polypeptides. This analysis suggested that the first 20, mainly hydrophobic, amino acids constitute the signal peptide. Thus, hamster preproglucagon is 160 amino acids and its predicted molecular weight of 18,675 is in good agreement with the value of 18,000 determined by Patzelt *et al.*[5] for the rat precursor. Glucagon is proglucagon(33–61) and is flanked by 52 and 99 amino acids at its amino and carboxy terminus, respectively. Proglucagon(1–69) possesses 90% homology with porcine intestinal glucagon-like immunoreactant 1 (GLI-1) or glicentin[10] and is probably the corresponding hamster protein (Fig. 1). Proglucagon(1–30) is 80% homologous to a porcine polypeptide, called glicentin-related pancreatic polypeptide (GRPP), which is secreted from the pancreas concomitantly with glucagon[11]. Proglucagon(33–69) is the pancreatic glucagon precursor initially described by Tager and Steiner[12] and which Bataille *et al.*[13] recently characterized from porcine intestine. Because pancreatic proglucagon contains the sequences of both pancreatic and intestinal glucagon-containing polypeptides, a common precursor may be synthesized in both tissues. The carboxy-terminal segment of proglucagon, residues 70–160, contains two glucagon-like peptides (GLP) of 37 and 35 amino acids, GLP-1 and -2 (Fig. 1). Each polypeptide is flanked by a pair of basic amino acids which can be sites of proteolytic processing. However, there is no evidence to suggest that the Arg–Arg at residues 109, 110 and 124, 125 are cleaved. In fact, the Arg–Arg at residues 49, 50 in the glucagon moiety is not cleaved. In addition, spacer oligopeptides of 6 and 13 residues separate glucagon and GLP-1, and GLP-1 and GLP-2, respectively. GLP-1 and -2 are related but not identical to other members of the glucagon–secretin family of gastrointestinal hormones which have been described (Fig. 2).

Lund *et al.*[14] have characterized an anglerfish pancreatic preproglucagon. This 124-amino acid precursor (M_r 14,500) is 56 amino acids smaller than hamster preproglucagon and this difference is due to the absence of the 13-amino acid spacer peptide and second glucagon-like peptide (GLP-2) in the anglerfish precursor (Fig. 1). Interestingly, this is the first example in which the organization of a prohormone has not been conserved during vertebrate evolution (compare mammalian and fish preproinsulin and preprosomatostatin[15,16]). Also, in contrast to mammals, anglerfish has another preproglucagon of ~12,500 M_r (refs 17, 18); however, its sequence has not been reported. Thus, there may be at least three different types of pancreatic proglucagon in vertebrates. Comparing hamster and anglerfish preproglucagon, the signal peptide, the amino-terminal peptide (corresponding to GRPP), glucagon and GLP-1 possess 25, 10, 69 and 48% amino acid homology and 47, 33, 76 and 66% nucleotide homology, respectively. The low level of sequence conservation in the signal peptide region is not unexpected because the absolute sequence of this region is not as important as the maintenance of its hydrophobic character[19]. Although the sequence of the amino-terminal peptide, that is, proglucagon(1–30), is not conserved, the size is and this segment may be required for proper processing of the precursor. Interestingly, in mammals, the

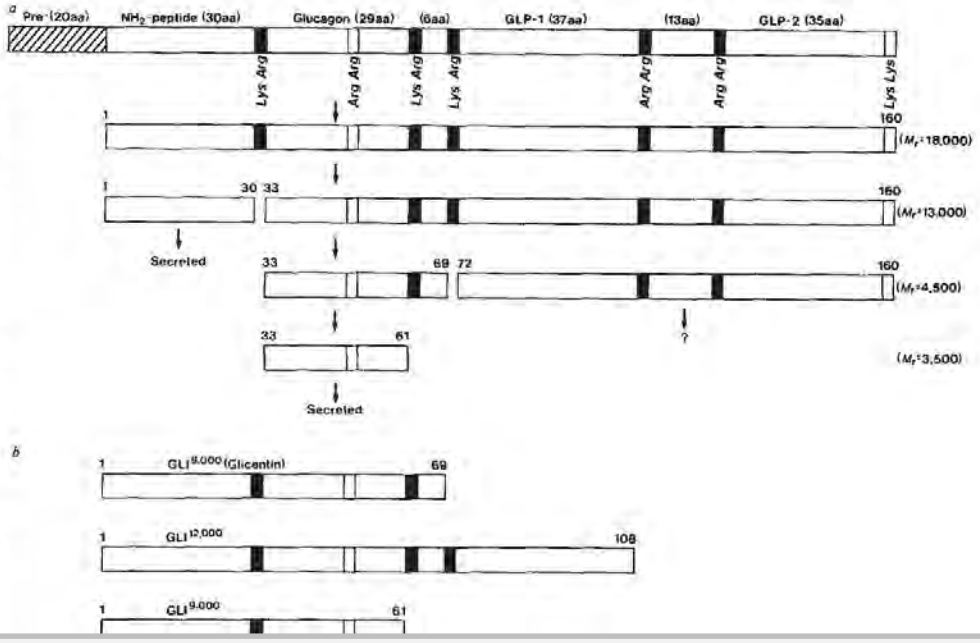

Fig. 3 Schematic representation of the processing of pancreatic preproglucagon and the structure of glucagon-containing polypeptides. *a*, Possible pathway for the proteolytic processing of pancreatic proglucagon. The basic dipeptides are indicated and those which are potential sites for cleavage are dark boxes. The numbers in parentheses at the end of each line are the sizes of the glucagon-containing intermediates determined by Patzelt *et al.*[5]. The numbers above the lines are the amino acids at the ends of the polypeptide in relation to the sequence of preproglucagon (Fig. 1). *b*, Structure of non-pancreatic glucagon-containing polypeptides. GLI 8,000 and GLI 12,000 are major polypeptides in the intestine and GLI 9,000 accumulates in the serum of animals with renal

sequence of proglucagon(1–30) is conserved to a greater extent than the C-peptide of proinsulin. For example, there is 80% homology between hamster and porcine proglucagon(1–30) and only 48% homology between their insulin C-peptides. The corresponding values in a comparison of hamster and human are 83% and 65%, respectively (G.I.B., unpublished). The spacer peptide which separates glucagon and GLP-1 is 6 amino acids in hamster and 5 in anglerfish and the sequence is different. The GLP-1 region possesses extensive homology between hamster and anglerfish, especially the segment corresponding to hamster proglucagon(78–100). However, hamster GLP-1 has a 6-amino acid amino-terminal extension which is absent in anglerfish. The sequence conservation in the GLP-1 region suggests that this peptide has a biological function. We have also compared hamster GLP-2 with the anglerfish GLP and they possess only 29% amino acid sequence homology. The 5′- and 3′-untranslated portions of hamster and anglerfish mRNA possess no significant regions of homology.

The pancreas and intestine are the major sites of synthesis of glucagon and glucagon-containing polypeptides. Figure 3 is a possible scheme for the processing of proglucagon in the pancreas. This model is based on the structure of preproglucagon presented here, the sizes and order of appearance of intermediates in the processing of rat proglucagon and the assumption that processing occurs at basic dipeptides. As indicated, both glucagon and proglucagon(1–30) are secreted[11]. However, the fate of proglucagon(72–160) and the two glucagon-like polypeptides is unknown. Intestinal glucagon-containing polypeptides are present at less than 1% the levels of pancreatic glucagon and the major polypeptides have M_rs of 8,000 and 12,000 (ref. 20). The 8,000-M_r protein is GLI-1 and corresponds to proglucagon(1–69) (Figs 1, 3). The sequence of the 12,000-M_r polypeptide has not been determined but its size, biochemical and immunological properties[3,20] are consistent with it being proglucagon(1–108), and thus it would include both glucagon and GLP-1 (Figs 1, 3). In addition, a 9,000-M_r peptide with glucagon-like immunoreactivity has been described which accumulates in the plasma of animals with renal failure[1–3,20]. Its size, biochemical and immunological properties suggest that it may correspond to proglucagon(1–61) (Fig. 3).

Glucagon and insulin have a major role in the regulation of plasma glucose levels. The biological role(s) of the intestinal glucagon-containing polypeptides is unclear although GLI-1 seems to inhibit gastric acid secretion[21]. Both the pancreatic and intestinal glucagon-containing polypeptides are probably derived from a common precursor which is processed differently in these two tissues. The processing of proglucagon can potentially generate at least 11 unique polypeptides, 8 of which have been identified biochemically or immunochemically. As the sequence of proglucagon is now known, it will be possible to synthesize polypeptides and to produce antisera to specific segments or polypeptides contained within the precursor[22]. The processing of the precursor in different tissues and the function of this family of glucagon-containing polypeptides in normal and disease states can then be critically examined. The difference in structure between mammalian and anglerfish proglucagon is unusual. The presence of an additional glucagon-like peptide in the mammalian precursor suggests that duplication or loss of a segment of the proglucagon gene has occurred. Examination of the structure of the gene may elucidate the evolutionary history of this hormone.

We thank our colleagues at Lilly Research Laboratories who helped in the isolation of the islets, Ms M. Quiroga and N. Fong for their technical assistance, Dr W. J. Rutter for his encouragement and suppport, Drs Leslie B. Rall and Pablo Valenzuela for their assistance in writing and Dr R. Najarian and Ms Maureen Appling who prepared the figures and typed the manuscript.

Received 2 December 1982; accepted 16 February 1983.

1. Foa, P., Bajaj, J. & Foa, N. (eds) *Glucagon: Its Role in Physiology and Clinical Medicine* (Springer, Berlin, 1977).
2. Unger, R. H. & Orci, L. (eds) *Glucagon. Physiology, Pathophysiology and Morphology of the Pancreatic A-Cell* (Elsevier, New York, 1981).
3. Conlon, J. M. *Diabetologia* **18**, 85–88 (1980).
4. Unger, R. H. & Orci, L. *New Engl. J. Med.* **304**, 1518–1524 (1981).
5. Patzelt, C., Tager, H. S., Carroll, R. S. & Steiner, D. F. *Nature* **282**, 260–266 (1979).
6. Proudfoot, N. J. & Brownlee, G. G. *Nature* **263**, 211–214 (1976).
7. Fitzgerald, M. & Shenk, T. *Cell* **24**, 251–260 (1981).
8. Carmichael, G. C. & McMaster, G. K. *Meth. Enzym.* **65**, 380–391 (1980).
9. Thomas, P. S. *Proc. natn. Acad. Sci. U.S.A.* **77**, 5201–5205 (1980).
10. Thim, L. & Moody, A. J. *Regulatory Peptides* **2**, 139–150 (1981).
11. Moody, A. J., Holst, J. J., Thim, L. & Lindkaer, Jenson, S. *Nature* **289**, 514–516 (1981).
12. Tager, H. S. & Steiner, D. F. *Proc. natn. Acad. Sci. U.S.A.* **70**, 2321–2325 (1973).
13. Bataille, D. *et al. FEBS Lett.* **146**, 79–86 (1982).
14. Lund, P. K., Goodman, R. H., Dee, P. C. & Habener, J. F. *Proc. natn. Acad. Sci. U.S.A.* **79**, 345–349 (1982).
15. Shen, L. P., Pictet, R. L. & Rutter, W. J. *Proc. natn. Acad. Sci. U.S.A.* **79**, 4575–4579 (1982).
16. Chan, S. J. *et al. J. biol. Chem.* **256**, 7595–7602 (1981).
17. Lund, P. K., Goodman, R. H. & Habener, J. F. *J. biol. Chem.* **256**, 6515–6518 (1981).
18. Shields, D., Warren, T. G., Roth, S. E. & Brenner, M. J. *Nature* **289**, 511–514 (1981).
19. Steiner, D. F., Quinn, P. S., Chan, S. J., Marsh, J. & Tager, H. S. *Ann. N.Y. Acad. Sci.* **343**, 1–16 (1980).
20. Tager, H. S. & Markese, J. *J. biol. Chem.* **254**, 2229–2233 (1979).
21. Kirkegaard, P. *et al. Nature* **297**, 156–157 (1982).
22. Lerner, R. A. *Nature* **299**, 592–596 (1982).
23. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. *Biochemistry* **18**, 5294–5299 (1979).
24. Santerre, R. F. *et al. Proc. natn. Acad. Sci. U.S.A.* **78**, 4339–4343 (1981).
25. Land, H., Grez, M., Hauser, H., Lindenmaier, W. & Schutz, G. *Nucleic Acids Res.* **9**, 2251–2266 (1981).
26. Villa-Komaroff, L. *et al. Proc. natn. Acad. Sci. U.S.A.* **75**, 3727–3731 (1978).
27. Gergen, J. P., Stern, R. H. & Wensink, P. C. *Nucleic Acids Res.* **7**, 2115–2136 (1979).
28. Comb, M., Seeburg, P. H., Adelman, J., Eiden, L. & Herbert, E. *Nature* **295**, 663–666 (1982).
29. Beaucage, S. L. & Caruthers, M. H. *Tetrahedron Lett.* **22**, 1859–1862 (1981).
30. Bell, G. I. *et al. Nature* **282**, 525–527 (1979).
31. Maxam, A. M. & Gilbert, W. *Meth. Enzym.* **65**, 499–560 (1980).
32. Dayhoff, M. O. *Atlas of Protein Sequence and Structure* Vol. 5, Suppl. 2, 125–126 (National Biomedical Research Foundation, Washington, DC, 1976).
33. Tatemoto, K. & Mutt, V. *Proc. natn. Acad. Sci. U.S.A.* **78**, 6603–6607 (1981).
34. Jornval, H. *et al. FEBS Lett.* **123**, 205–216 (1981).
35. Guillemin, R. *et al. Science* **218**, 585–587 (1982).

A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing

Scott D. Putney*, Walter C. Herlihy† & Paul Schimmel*

* Department of Biology and † Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Complete amino acid sequences have been established for 19 muscle-related proteins and these proteins are each sufficiently abundant to suggest that their mRNA levels are about 0.4% or higher. Based on these considerations, a simple theoretical analysis shows that clones for most of these proteins can be identified within a complementary DNA library by sequencing cDNA inserts from 150–200 randomly selected clones. This procedure should not only rigorously identify specific clones, but it could also uncover amino acid sequence variants of major muscle proteins such as the troponins[1–6]. We have determined sequences for about 20,000 nucleotides within 178 randomly selected clones of a rabbit muscle cDNA library, and report here that in addition to finding sequences encoding the two known skeletal muscle isotypes of troponin C[7–9], we have discovered sequences encoding two forms of troponin T. Over the region of nucleotide sequence overlap in the troponin T clones, the new isotype diverges significantly from its counterpart[10]. Altogether, clones for 13 of the 19 known muscle-specific proteins were identified, in addition to the clone for the new troponin T isotype.

To identify a clone for a particular protein by DNA sequencing, the nucleotide sequence of a cDNA clone encoding a portion of the protein sequence must be determined. We determined sequences of cDNA fragments isolated from a library of rabbit muscle cDNA cloned into M13 phage[11]. Before cloning, the cDNA was restricted with *Msp*I, *Taq*I or *Sau*3AI so that cDNA fragments of ~250 base pairs (bp) were actually cloned. Sequences of ~110 nucleotides from 178 different phage inserts were determined. The sequences were translated