Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 6571–6575, October 1985
Cell Biology

Repeating modular structure of the fibronectin gene: Relationship to protein structure and subunit variation

(exons/introns/RNA splicing/protein domains)

ERICH ODERMATT, JOHN W. TAMKUN, AND RICHARD O. HYNES

Center for Cancer Research and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

Communicated by Mary Lou Pardue, June 3, 1985

ABSTRACT **Analysis of the exon–intron structure of the rat fibronectin gene shows that exons correspond precisely with repeating structural units in the protein and that alternative use of some exons produces fibronectin subunits that differ by the presence or absence of certain structural modules. Secondary structure predictions suggest that the repeating structure of the protein is further subdivided into smaller structural units and that these also correspond with exons in the gene.**

The idea that structural or functional units within proteins might be encoded by separate exons within genes and that such units might reassort during evolution (1) has received a good deal of attention over the past several years. Though some proteins show a good correspondence between exons and structural units of proteins (2–15), others do not show an obvious relationship (16–19).

We have been interested in the structure of fibronectins and in their origin. Fibronectins comprise a closely related group of glycoproteins involved in cell adhesion to extracellular materials and are important in various biological phenomena, including embryonic cell migration, oncogenic transformation, wound healing, hemostasis, and thrombosis (20, 21). These proteins consist of well-defined domains specialized for a variety of binding functions. Furthermore, the protein sequences of fibronectins contain multiple homologous but nonidentical repeats of three types (22). Two of these (types I and II) consist of disulfide-bonded loops around 50 amino acids long. The third type of repeat (type III) is around 90 amino acids long and is not disulfide-bonded; there are at least 15 type III repeats per monomer of ≈250,000 daltons (22–25). Different fibronectin monomers differ in the number of type III repeats and in the presence or absence of segments of sequence that do not fit any of the three types of homology (23–25). These different subunits are all encoded by a single gene and must arise by alternative splicing of the primary transcript (23–29).

An interesting problem posed by these results is the way in which the various functional domains and repeating structural units that make up a fibronectin subunit are encoded in the exons of the gene. We report here that the type III repeats are indeed represented by a repeating structure in the gene, comprising two exons, which together encode a complete repeat. Furthermore, analyses of the gene structure presented here and elsewhere (27, 29) show that variation in the number of type III repeats arises by a form of alternative splicing in which the sequences encoding an entire repeat can be included or omitted.

MATERIALS AND METHODS

Enzymes were purchased from New England Biolabs, except for the calf alkaline phosphatase (Boehringer Mannheim). The rat genomic clone λrFN-2 containing 18 kilobases (kb) of the ≈50-kb fibronectin gene is described elsewhere (27). Fragments derived from the phage DNA were sequenced by the method of Maxam and Gilbert (30).

RESULTS AND DISCUSSION

Repeating Structure of the Fibronectin Gene. We have analyzed 10 kb of the rat fibronectin genomic clone, λrFN-2 (ref. 27, see Fig. 1). This stretch includes 12 exons encoding the cell- and heparin-binding domains from the central region of fibronectin (23). The sequences of the intron–exon boundaries are given in Fig. 2. This region of fibronectin contains seven type III repeats ($n - 6$ to n) and one stretch of a different type of sequence (23, 27). The amino acid sequence is shown in Fig. 3. An intron is present at the end of each type III repeat (arrowheads); the repeating structure of the protein reflects a repetitive unit in the gene. However, with one exception, each type III repeat is encoded by *2* exons, not 1 (Fig. 3).

Variations and Alternative Splicing. This basic repeating pattern is interrupted at two points. The first of these is in repeat $n - 4$. This type III repeat was not present in any of the cDNA clones isolated from rat liver (23) nor is it present in the amino acid sequence of bovine plasma fibronectin (22). However, it *is* present in the sequence of one of two cDNA clones isolated from a human cell line (24, 25) and is incorporated into mRNA in several rat cell lines, as shown by RNA transfer blotting (unpublished data). The sequence of this repeat was determined (Fig. 4). It is almost identical with the sequence determined from human cDNA (24, 25) and is a typical type III homology. However, unlike the other six type III repeats shown in Fig. 3, this repeat is encoded by a single exon.

This exon is not included in fibronectin mRNA by liver cells (24, 25), which synthesize and secrete plasma fibronectin (38), but is included in mRNA by other cell types synthesizing cellular fibronectin, which contains subunits absent from plasma fibronectin (38, 39). Therefore, this exon is utilized in a tissue-specific fashion to generate different fibronectin subunits. Given the structure of the fibronectin gene described here, the facultative inclusion of this repeat must be by means of alternative splicing involving skipping of this exon in hepatocytes and its inclusion in other cell types. While this manuscript was in preparation, Vibe-Pedersen *et al.* (29) reported that this type III repeat is also encoded by a single exon in the human fibronectin gene and is alternatively spliced. It is plausible to argue that this exon-skipping mechanism requires this repeat to be encoded in a single exon to avoid the possibility of inclusion of only one portion of a type III repeat. A likely evolutionary origin of this exon is deletion of the intron interrupting the typical type III unit.

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

Abbreviation: kb, kilobase(s).

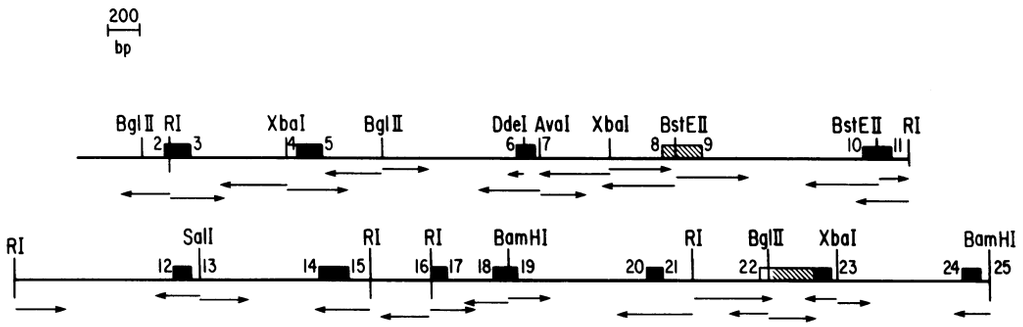

FIG. 1. Structure of the rat fibronectin gene in the region encoding the cell- and heparin-binding domains. Approximately 10 kb of clone λrFN-2 (27) were analyzed by restriction enzyme digestion and DNA sequencing. The restriction enzyme sites marked are those used in sequencing and do not comprise a complete listing of sites. The exons are shown as boxes and correspond precisely with the cDNA sequences (23). All intron–exon boundaries (numbered 2–25) were sequenced (see Fig. 2). The 5′ *Eco*RI site lies in the exon encoding the cell attachment site of fibronectin (23). The 3′ *Bam*HI site lies in the intron separating the last type III repeat of the heparin-binding region from the first type I repeat of the fibrin-binding domain (23, 27). Between these two sites there are 12 exons encoding 6½ type III repeats and the region of difference between subunits of plasma fibronectin (see refs. 23, 27, and 31). The two exons involved in alternative splicing are shown as hatched boxes (see text). The protein sequence encoded by the exons is given in Figs. 3 and 4. bp, Base pairs.

The second variation on the repeating pattern of two exons to one type III repeat falls between the last two repeats (Fig. 3). As described previously, the exon encoding the NH_2-terminal half of the last repeat (*n*) also encodes a 120 amino acid sequence of a different type (27). Use of alternative 3′ splice sites within this exon allows inclusion or omission of this difference segment (23, 27). Some subunits of plasma fibronectin contain this segment, whereas others do not (31).

Fig. 5 shows the genomic structure of a typical type III repeat (Fig. 5*A*), the type III repeat involved in exon skipping (Fig. 5*B*), and the type III repeat involved in exon subdivision (Fig. 5*C*). *Exon skipping* appears to give rise to the difference between cellular and plasma fibronectins, and *exon subdivision* accounts for differences between subunits of plasma fibronectin. These two forms of alternative splicing both occur in rat and the exon skipping also occurs in human fibronectin (24, 25, 29). One human cDNA clone covering the second region of alternative splicing (25) includes only the first 89 amino acids of the difference segment (see Fig. 5*D*), whereas another clone from a different source includes all 120 residues of this segment (40). The structure of the human gene is not known. The rat gene contains a sequence ATGAGGAG immediately following the 89 amino acid sequence (27). Replacement of the 5′ adenylate residue with a guanylate residue in the human cDNA sequence (40) gives a 5′ splice site, which apparently can sometimes be used to splice out the 93 bases encoding 31 amino acids following the 89 amino acid segment. Therefore, the human gene could, in principle, produce up to five variant splicing patterns (Fig. 5*D*). Exactly how this region of the human gene is in fact used will have to await sequencing of this segment and S1 nuclease analysis.

Relationship Between Gene Structure and Protein Structure. If all type III repeats were encoded by single exons, one would conclude that there was good correspondence between the units of structure in the gene and in the protein, with

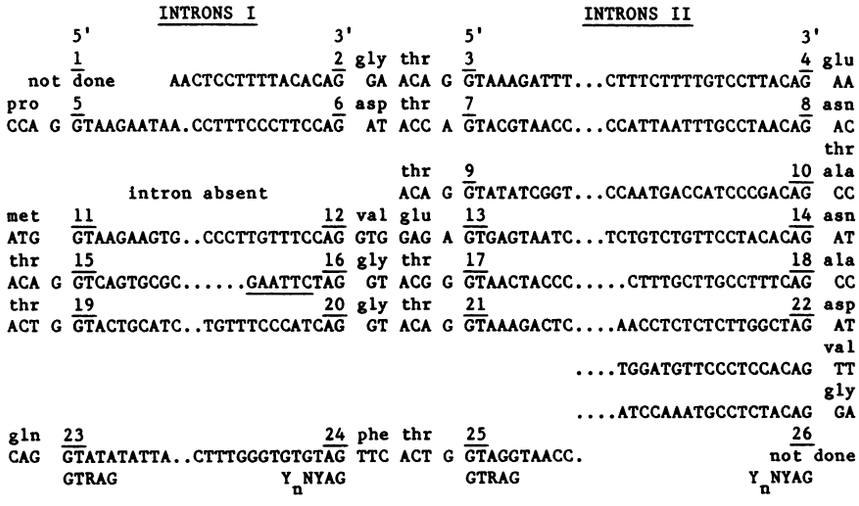

FIG. 2. Intron–exon boundaries of the rat fibronectin gene. The boundaries are arranged according to the repeating structure of the gene and the protein and are numbered as in Figs. 1 and 3. The consensus sequences for intron boundaries (32) are shown at the bottom (Y = pyrimidine; R = purine; N = any base). Introns II separate type III repeats from each other and always fall between the first and second bases of a codon. Introns I, where present, interrupt the type III repeats in various places (Fig. 3) and the positions of the introns within codons also vary. The 3′ splice site of intron I of repeat $n - 2$ (boundary 16) is immediately preceded by an *Eco*RI site (underlined). Repeat $n - 4$ lacks intron I. The amino acid following intron II of repeat $n - 4$ (boundary 10) is alanine or threonine, depending on which 5′ splice site (boundary 7 or 9) is used—i.e., whether or not repeat $n - 4$ is included (see also ref. 24). The penultimate repeat is sometimes followed directly by the last repeat and sometimes followed by different segments of coding sequence (23). Therefore, intron II of repeat $n - 1$ has three different potential 3′ splice sites (boundary 22), all of which are shown (see refs. 23 and 27).

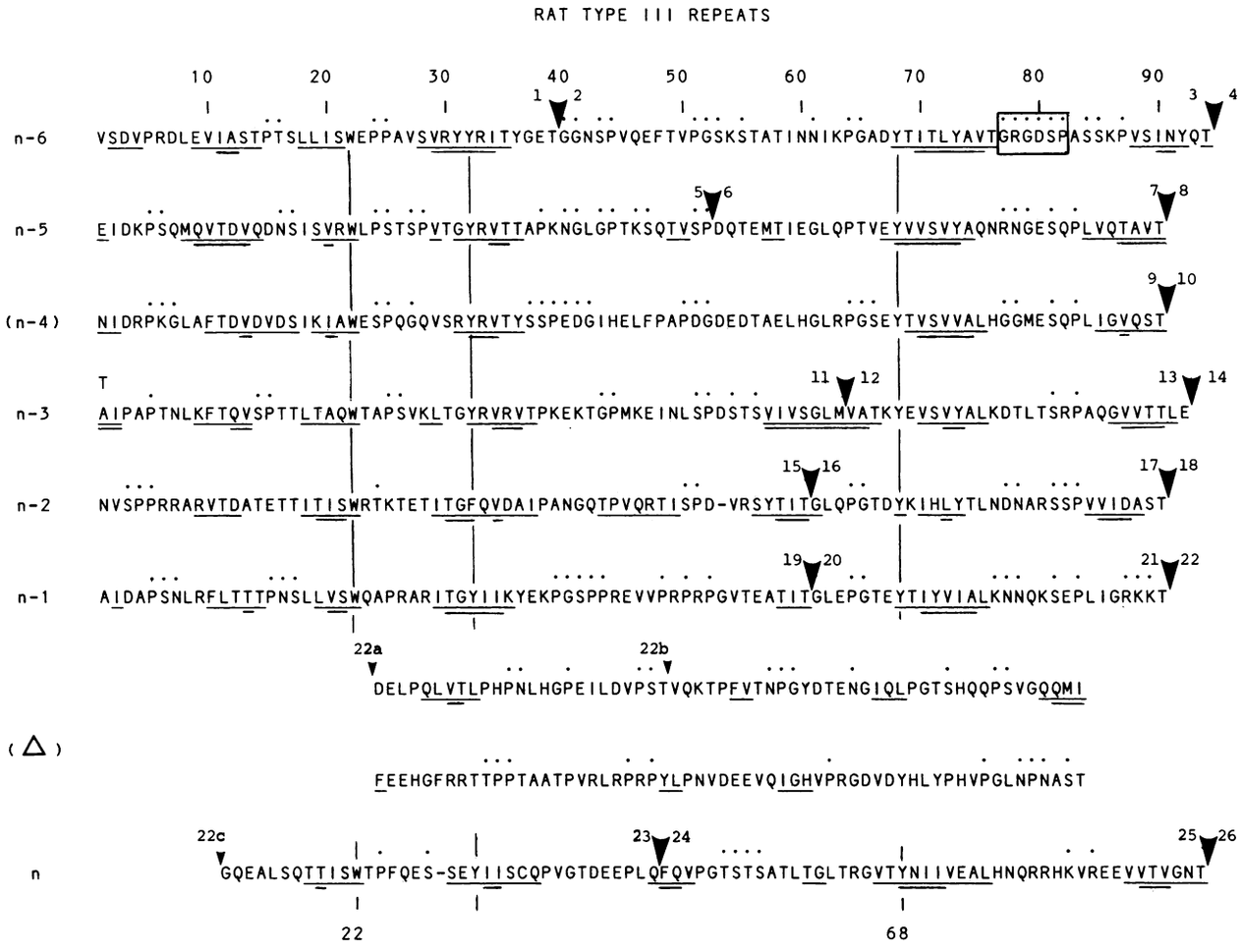

FIG. 3. Repeating structure of the cell- and heparin-binding regions of rat fibronectin. The amino acid sequence deduced from cDNA (23) and genomic DNA sequences (Figs. 1 and 4) is shown. Residues 1–49 conform with unpublished cDNA sequence (J. E. Schwarzbauer, personal communication). Repeat $n - 4$ is based on the genomic DNA sequence (Fig. 4). The sequence is arranged to show the homologous repeating structure. Note particularly the conserved aromatic residues at positions 22 and 68 in each type III repeat (see also refs. 22 and 23 and text). The segment marked (△) is not a type III repeat and is not always present in mRNA. Repeat $n - 4$ can also be omitted. When it *is* included the first amino acid of repeat $n - 3$ is alanine; when repeat $n - 4$ is omitted, the first amino acid of repeat $n - 3$ is threonine (see also Fig. 2 and ref. 24). Positions of the introns (see Fig. 1) are shown by arrowheads. Note that an intron separates each type III repeat from the next and that most repeats are also interrupted by another intron (see Fig. 2). Predicted β-structures (33, 34) are denoted by underlining, with a single line representing a weaker prediction (Chou–Fasman average ≥ 1.1; Robson average ≥ 10) and double underlining representing a stronger prediction (Chou–Fasman average ≥ 1.2; Robson average ≥ 20). β-turns are denoted by dots and are based on predictions by three methods (33–35). The predictions agreed well between the various methods. The hexapeptide shown to be involved in cell attachment (36) is boxed in repeat $n - 6$. The sequence is given in the single-letter code (37).

introns separating protein structural units. But what is the significance of introns' being present in the middle of type III homologies and, furthermore, varying in their exact location? To investigate whether the exons correspond with some smaller structural unit, we have analyzed the predicted secondary structure of the type III repeats using computer programs (33–35).

We observed striking similarities among the structures

```
CCATTAATTTGCCTAACAG AC ATT GAC CGC CCT AAA GGA CTG GCA TTC ACT GAT GTG GAT GTC GAT TCC ATC AAA ATT
                       asn ile asp arg pro lys gly leu ala phe thr asp val asp val asp ser ile lys ile

GCC TGG GAA AGC CCA CAG GGG CAA GTT TCC AGG TAC AGG GTG ACC TAC TCA AGC CCT GAG GAT GGA ATC CAT GAG
ala trp glu ser pro gln gly gln val ser arg tyr arg val thr tyr ser ser pro glu asp gly ile his glu

CTT TTC CCT GCG CCT GAT GGT GAC GAG GAC ACG GCA GAG CTG CAC GGC CTC AGG CCG GGT TCT GAG TAC ACA GTC
leu phe pro ala pro asp gly asp glu asp thr ala glu leu his gly leu arg pro gly ser glu tyr thr val

AGT GTG GTT GCC TTG CAC GGT GGC ATG GAG AGC CAG CCC CTG ATT GGA GTC CAG TCC ACA G GTATATCGGTGACTGCA
ser val val ala leu his gly gly met glu ser gln pro leu ile gly val gln ser thr
```

FIG. 4. Sequence of the exon encoding a type III repeat that is spliced in a tissue-specific fashion. The DNA sequence (see Fig. 1) and corresponding amino acid sequence are shown. The DNA sequence of the exon differs at 15 of 270 positions from the human cDNA sequence (24). Most of these changes (11) are third base changes. The amino acid sequence differs at only 5 of 90 positions from the sequence of human fibronectin (24). The positions that differ are marked by underlining. The corresponding amino acids in the rat and human sequences are, respectively, aspartic acid/glutamic acid, histidine/glutamine, glycine/aspartic acid, glycine/aspartic acid, and valine/threonine. The sequences of the flanking introns are remarkably similar to those recently reported for the equivalent human exon (29).

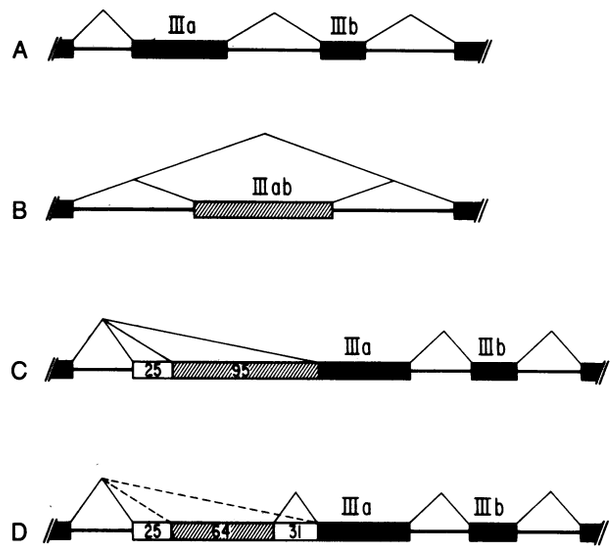

FIG. 5. Two different types of alternative splicing of the fibronectin gene transcript. (*A*) A typical type III repeat of rat fibronectin is encoded by two exons (IIIa and IIIb). (*B*) One repeat that is included by some cell types but not by others is encoded in a single exon that can be omitted by exon skipping during splicing. (*C*) The last type III repeat is encoded by two exons. One of these (IIIb) is similar to the 3′ exon of a typical type III repeat, but the 5′ exon is longer and contains a 5′ extension encoding segments of coding sequence (25 and 95 amino acids long) that can be included or not. The 3′ segment of this exon (marked IIIa) is always present; exon subdivision during splicing gives rise to three different mRNAs (refs. 23 and 27). (*D*) Possible structures of the human fibronectin gene in the region encoding the difference segment and the last type III repeat. One published human cDNA sequence (25) contains only the first 89 amino acids corresponding to the rat difference segment. The last 31 amino acids can also be included as shown by a different human cDNA clone (40). Possible splicing patterns are shown as dashed lines (see text for discussion).

predicted for the seven type III repeats. In each case, several segments of β-structure were predicted (underlined in Fig. 3) and these fell in homologous positions in each repeat, corresponding with blocks of conserved amino acid sequence (23). In contrast with the consistency of the β-predictions, only scattered α-helical segments were predicted and were not in homologous positions in different repeats (not shown). These predictions are consistent with CD results showing that there is no α-helix but quite a lot of β-structure (41–43).

The first 40 amino acids of each repeat are predicted to have the same β-sheet structure. The β-sheets show a pattern of alternating hydrophobic residues, which could allow either a sheet with a hydrophobic face or a cylinder with a hydrophobic core. Consistent with such structures, it is known from CD analyses that this region of fibronectin contains highly ordered aromatic residues (41, 43, 44). Whatever the exact structure of this segment of each repeat, it seems clear that it is a conserved structural unit. This conserved unit is never interrupted by an intron (see Fig. 3).

In contrast, the next 25 residues of each repeat are not well conserved (23) and no consistent secondary structure predictions were obtained for this region (see Fig. 3). This region appears to be heavily interrupted by β-turns and only three repeats ($n - 3$, $n - 2$, and n) show any significant stretches of predicted β-structure. This segment (residues 40–65) is the region that is interrupted by introns (Fig. 3).

Finally, the region between residue 65 and the end of each repeat again appears to have a highly conserved structure (Fig. 3) consisting of two β-segments, separated by a hydrophilic loop containing β-turns. A reasonable prediction for the structure of this segment in each type III repeat is a β-hairpin. The hexapeptide that constitutes the active site for cell adhesion (36) lies in the loop of repeat $n - 6$ (Fig. 3). These β-strands also have alternating hydrophobic residues, including conserved aromatics, allowing for a hydrophobic side to the β-hairpin. Again, this conserved structural unit is never interrupted by introns (Fig. 3). We have also performed secondary structure predictions on the sequences of six more human type III repeats that lie NH_2-terminal to those analyzed here (25). All six fit well with the structural picture outlined here.

These computer analyses suggest that each type III repeat consists of two conserved structural minidomains or modules separated by a long poorly conserved segment. Each of the structural modules lies in a separate exon. The introns separating these two modules always lie in the unconserved and relatively unstructured loop, although their exact position varies. A similar correspondence in position between variable protein loops and introns has been noted in proteolytic enzymes (45, 46) and it has been suggested that introns move along genes generating variations in coding sequence.

The difference segment (23, 27) between the last two type III repeats appears to have very little regular structure, which is not surprising given its high content of proline (16%). The exon encoding this segment and part of type III repeat *n* encodes two completely different structures, the proline-rich segment and a β-structured module of type III. An analogy with this can be found in the collagen gene, in which the repeating 54-base-pair exons encoding collagen triple helix can occur fused with exons encoding nonhelical pro segments (7, 8). Perhaps this large hybrid fibronectin exon arose during evolution either by removal of an intron fusing two exons or by extension of the exon to the 5′ side.

Therefore, the remarkable repeating structure of fibronectin is reflected in the exon structure of the gene. Each type III repeat is encoded in a pair of exons and each of these exons appears to encode a separate structural, and possibly functional, unit within the protein. This repeating structure strongly suggests that the gene arose by endoduplication during evolution. The repeating pattern of exon pairs is modified in two places, apparently by exon fusions or extensions. Each of these variant repeats is involved in alternative splicing to generate different subunits of fibronectin. Thus, this region of the gene consists of separate modules that are spliced together, and sometimes selectively removed, to construct a set of related complex proteins.

We thank Jean Schwarzbauer for helpful advice and discussions, Phil Auron, Gary Quigley, and Will Gilbert for help with computer analyses, and Richard Black and Susan Podshadley for preparing the manuscript. This research was supported by a grant from the Public Health Service, National Cancer Institute (PO1 CA 26712). E.O. was supported by a postdoctoral fellowship from the Swiss National Science Foundation and J.W.T. was supported by a predoctoral fellowship from the Whitaker Health Sciences Foundation.

1. Gilbert, W. (1978) *Nature (London)* **271,** 501.
2. Jung, A., Sippel, A. E., Grez, M. & Schutz, G. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 5759–5763.
3. Go, M. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1964–1968.
4. Stein, J. P., Catterall, J. F., Kristo, P., Means, A. R. & O'Maley, B. W. (1980) *Cell* **21,** 681–687.
5. Jensen, E. O., Paludan, K., Hyldig-Nielsen, J. J., Jorgensen, P. & Marcker, I. C. A. (1981) *Nature (London)* **241,** 677–679.
6. Sargent, T. D., Jagodzinski, L. L., Yang, M. & Bonner, J. (1981) *J. Mol. Cell. Biol.* **10,** 871–883.
7. Yamada, Y., Avvedimento, V. E., Mudry, J. M., Ohkubo, H., Vogeli, G., Irani, M., Pastan, I. & de Crombrugghe, B. (1980) *Cell* **22,** 887–892.
8. Wozney, J., Hanahan, D., Morimoto, R., Boedtker, H. & Doty, P. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 712–716.
9. Inana, G., Piatigorsky, J., Norman, B., Slingsby, C. P. & Blundell, T. (1983) *Nature (London)* **302,** 310–315.

10. Nathans, J. & Hogness, D. S. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 4851–4855.
11. Sakano, H., Rogers, J. H., Huppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R. & Tonegawa, S. (1979) *Nature (London)* **277,** 627–633.
12. Tonegawa, S. (1983) *Nature (London)* **302,** 575–581.
13. Honjo, T. (1983) *Annu. Rev. Immun.* **1,** 499–528.
14. Hood, L., Steinmetz, M. & Goodenow, R. (1982) *Cell* **28,** 685–687.
15. Ny, T., Elgh, F. & Lund, B. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 5355–5359.
16. Beryajati, C., Place, A. R., Powers, D. A. & Sofer, W. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 2717–2721.
17. Firtel, R. A. (1980) *Cell* **24,** 6–7.
18. Karn, J., Brenner, S. & Barnett, L. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 4253–4257.
19. Quax, W., Egberts, W. V., Hendriks, W., Quax-Jenken, Y. & Bloemendahl, H. (1983) *Cell* **35,** 215–223.
20. Hynes, R. O. & Yamada, K. M. (1982) *J. Cell Biol.* **95,** 369–377.
21. Yamada, K. M. (1983) *Annu. Rev. Biochem.* **52,** 761–799.
22. Petersen, T. E., Thogersen, H. C., Skorstengaard, K., Vibe-Pedersen, K., Sahl, P., Sottrup-Jensen, L. & Magnusson, S. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 137–141.
23. Schwarzbauer, J. E., Tamkun, J. W., Lemischka, I. R. & Hynes, R. O. (1983) *Cell* **35,** 421–431.
24. Kornblihtt, A. R., Vibe-Pedersen, K. & Baralle, F. E. (1984) *EMBO J.* **3,** 221–226.
25. Kornblihtt, A. R., Vibe-Pedersen, K. & Baralle, F. E. (1984) *Nucleic Acids Res.* **12,** 5853–5868.
26. Kornblihtt, A. R., Vibe-Pedersen, K. & Baralle, F. E. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3218–3222.
27. Tamkun, J. W., Schwarzbauer, J. E. & Hynes, R. O. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 5140–5144.
28. Hirano, H., Yamada, Y., Sullivan, M., de Crombrugghe, B., Pastan, I. & Yamada, K. M. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 46–50.
29. Vibe-Pedersen, K., Kornblihtt, A. R. & Baralle, F. E. (1984) *EMBO J.* **3,** 2511–2516.
30. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
31. Schwarzbauer, J. E., Paul, J. I. & Hynes, R. O. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1424–1428.
32. Sharp, P. A. (1981) *Cell* **23,** 643–646.
33. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13,** 222–244.
34. Garnier, J., Osguthorpe, D. J. & Robson, B. (1978) *J. Mol. Biol.* **120,** 97–120.
35. Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. & Fletterick, R. J. (1983) *Biochemistry* **22,** 4894–4904.
36. Pierschbacher, M. D. & Ruoslahti, E. (1984) *Nature (London)* **309,** 30–33.
37. IUPAC–IUB Commission on Biochemical Nomenclature (1968) *Eur. J. Biochem.* **5,** 151–153.
38. Tamkun, J. W. & Hynes, R. O. (1983) *J. Biol. Chem.* **258,** 4641–4647.
39. Paul, J. I. & Hynes, R. O. (1984) *J. Biol. Chem.* **259,** 13477–13487.
40. Bernard, M. P., Kolbe, M., Weil, D. & Chu, M. L. (1985) *Biochemistry* **24,** 2698–2704.
41. Alexander, S. S., Colonna, G. & Edelhoch, H. (1979) *J. Biol. Chem.* **254,** 1501–1505.
42. Odermatt, E., Engel, J., Richter, H. & Hormann, H. (1982) *J. Mol. Biol.* **159,** 109–123.62.
43. Tooney, N. M., Amrani, D. L., Homandberg, G. A., McDonald, J. A. & Mosesson, M. W. (1982) *Biochem. Biophys. Res. Commun.* **108,** 1085–1091.
44. Welsh, E. J., Frangou, S. A., Morris, E. R., Rees, D. A. & Chavin, S. I. (1983) *Biopolymers* **22,** 821–831.
45. Craik, C. S., Sprang, S., Fletterick, R. & Rutter, W. J. (1982) *Nature (London)* **299,** 180–182.
46. Craik, C. S., Rutter, W. J. & Fletterick, R. (1983) *Science* **220,** 1125–1129.