Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 2944–2948, May 1988
Biochemistry

Do exons code for structural or functional units in proteins?

(exon size/modules/intron conservation/protein evolution)

THOMAS W. TRAUT

Department of Biochemistry, University of North Carolina School of Medicine, Chapel Hill, NC 27599-7260

Communicated by Mary Ellen Jones, January 4, 1988 (received for review August 27, 1987)

ABSTRACT **In considering the origin and evolution of proteins, the possibility that proteins evolved from exons coding for specific structure–function modules is attractive for its economy and simplicity but is not systematically supported by the available data. However, the number of correspondences between exons and units of protein structure–function that have so far been identified appears to be greater than expected by chance alone. The available data also show (*i*) that exons are fairly limited in size but are large enough to specify structure–function modules in proteins; (*ii*) that the position of introns for homologous domains in the same gene is reasonably stable, but there is also evidence for mechanisms that alter the position or existence of introns; and (*iii*) that it is possible that the observed relationship of exons to protein structure represents a degenerate state of an ancestral correspondence between exons and structure–function modules in proteins.**

After the discovery of introns and exons, it was suggested that protein evolution could have benefited if exons coded for functional units (1) or structural units (2) in proteins. As gene sequences for structurally defined proteins were determined, there were some positive reports for a correspondence between exons and units of protein structure–function for lysozyme (3), hemoglobin (4, 5), ovomucoid (6), and IgG (7). However, other studies claimed a lack of such correspondence for hemoglobin (8), carboxypeptidase (9), and ovalbumin and antitrypsin (10), thereby casting doubt on the possibility of a general pattern.

If a correspondence existed between exons and units of protein structure–function for ancestral genes, to what extent can such a pattern be discerned in the available data?

RESULTS AND DISCUSSION

Size of Exons. An earlier analysis of exon size by Naora and Deacon (11) measured exon size in nucleotides and found that exons from ≈80 genes have a narrow size range centering on 140 base pairs but suggested that there might be discrete classes coding on average for 50, 140, 200, and 300 base pairs. In the present analysis exons are sized by the number of amino acids for which they code. This has a significant effect on 5′ and 3′ exons (≈25% of the data set), since usually only a fraction of the bases in each of these exons codes for amino acids. An effort was made to include a variety of gene types and to limit entries from the same gene family; in the earlier study (11) of the 80 genes examined, 20 were globins and 10 were histones.

Fig. 1 shows the size distribution for exons from 114 genes (>800 total exons). The 5′ exons and 3′ exons are shown separately. Whereas internal exons are expected to contribute the major portions of protein structure, it is possible that the initial and terminal exons code for structure, regulatory signals, or both. Internal exons predominantly code for 20–55 amino acids, whereas 5′ and 3′ exons do not show the same tendency toward a modal size. The 5′ exons are on average significantly smaller ($\bar{\chi}$ = 34.4 amino acids) compared to 3′ exons ($\bar{\chi}$ = 42.6 amino acids) and internal exons ($\bar{\chi}$ = 44.5 amino acids).

Do Exons Code for Units of Protein Structure–Function? This possibility was addressed in less than half of the literature surveyed. Some authors sought for specific correspondence between exons and units of secondary structure or tertiary structure, others looked for a match to a region of the protein known to have an associated function such as signal peptide, transmembrane domain, nucleotide-binding fold, etc. No systematic criteria were used nor were different groups equally stringent in deciding how the limits of the exon had to correspond to the limits of the unit in the protein to designate a positive correspondence. The results as defined by these authors are summarized in Table 1.

No meaningful correspondence for even a single exon was found for carboxypeptidase (9), lactate dehydrogenase (82), glycogen phosphorylase (91), leghemoglobin (93), calmodulin (95), and myosin heavy chain (111), whereas all exons could be assigned a specific function for lysozyme (3), chicken triosephosphate isomerase (79), β-globin (4, 5), transplantation antigen (103), IgG heavy chain (7), β-crystallin (107), glucagon (97), and parathyroid hormone (99). Correspondences between exons and units of protein structure were found for 10 proteins (6, 7, 79, 80, 90, 100, 107–109), and correspondences between exons and units of function were found for 24 proteins (3–5, 39, 76–78, 81, 83–89, 94, 96–99, 101–107).

For the proteins in Table 1, overall >50% of the exons were identified as coding for some unit of structure or function. This value falls far short of a 100% correspondence expected for the general hypothesis that exons code for units of protein structure–function. In favor of the latter hypothesis, the full range of possible exon functions is not always considered. These could include components of the mature protein such as ligand binding modules as well as linker regions (e.g., hinge region of IgG), components for protein transport or activation (signal peptides, activation peptides in zymogens, or sites for covalent modification), and components for RNA processing (5′- and 3′-untranslated exons) and translation. Small exons occur more frequently at the 5′ or 3′ positions; examples where such exons specify only a single amino acid suggest that their function may be to provide a start codon (83, 84, 95, 101) or stop codon (59, 97, 103, 112).

With a view to the limited size range for internal exons, consider more specifically what it is possible for such exons to specify. Domains, as generally identified (114), have a molecular mass range of 3–32 kDa, with an average of 12 kDa. Therefore, exons are generally too small to specify domains. Zehfus and Rose (115) have defined units of compact folded structure; the smallest "primitive compact units" contain 6–39 amino acids. Exons are too large, in general, to specify such compact units, though the average exon might specify two or more such units. The smallest units of protein that have folded structure and also some type of function contain

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.

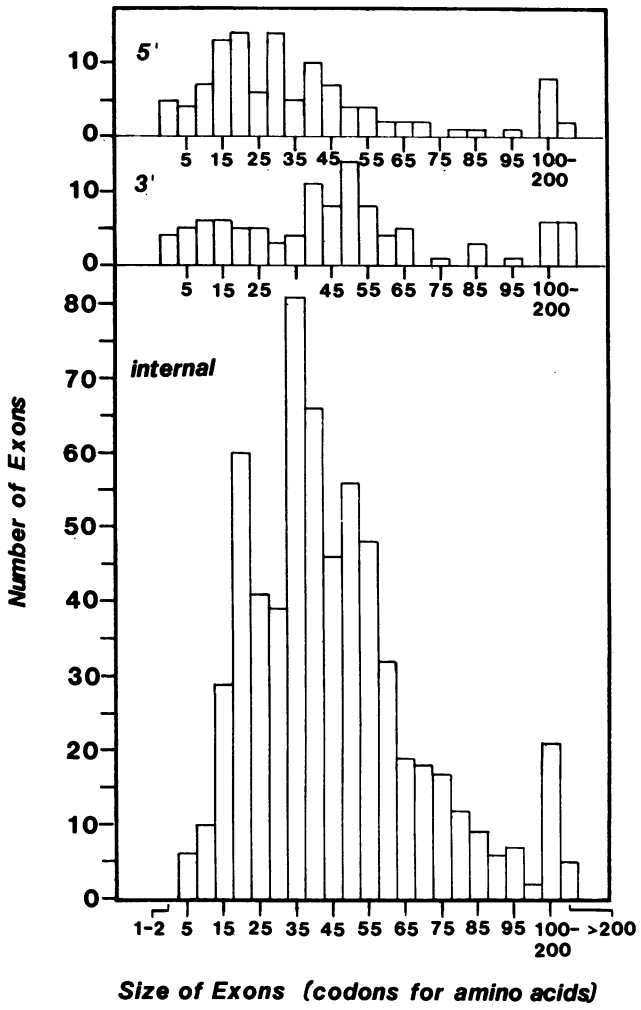

FIG. 1. The number of amino acids coded by exons. Since many genes contain untranslated exons, the first exon of a gene that coded for amino acids was counted as the 5′ exon, the last exon coding for amino acids was counted as the 3′ exon. The exons are from the genes shown in Table 1, as well as from mouse dihydrofolate reductase (12); *Aspergillus nidulans* triosephosphate isomerase (13); human pepsinogen (14); rat amylase (15); rabbit phosphofructokinase (16); rat trypsin (17); human complement protein factor B (18); human catalase (19); hypoxanthine phosphoribosyltransferase from mouse (20) and human (21); chicken aminolevulinate synthase (22); deoxythymidine kinase from chicken (23) and hamster (24); human argininosuccinate synthase (25); human haptoglobin (26); human transferrin (27); chicken conalbumin (28); human apoferritin (29, 30); chicken ovalbumin (31); human metallothionein (32, 33); seal myoglobin (34); human α globin (35); yeast cytochrome *b* (36); rat cytochrome P450 (37); yeast cytochrome oxidase (38); human genes for blue pigment, green pigment, and red pigment (39); human interleukin 2 receptor (40); human antithrombin III (41); rat α_1-acid glycoprotein (42); chicken histone H3 (43); ribosomal protein S14 (44); chicken myosin light chain (45); *Drosophila* tropomyosin (46); chicken α-actin (47); human β-actin (48, 49); bovine αB-crystallin (50); mouse glial fibrillary acidic protein (51); human keratin (52); mouse kallikrein (53); rat α-tubulin (54); human α-tubulin (55); human vimentin (56); human histocompatibility gene SXα (57); rat preprosomatostatin (58); human pronatriodilantin (59); human T-cell growth factor (60); rat insulin (61); human insulin (62); rat growth hormone (63); human growth hormone (64); bovine corticotropin/β-lipotropin precursor (65); human proopiomelanocortin (66); rat prolactin (67); rat interleukin 3 (68); feline c-*sis* (69); human *NRAS* (70); mouse c-*myb* (71); human 27-kDa heat shock protein (72); human pS2 protein (73); human *FGR* (74); and human p53 (75).

20–40 amino acids or more (116), totally consistent with the size range of internal exons. Thus, exons are generally large enough to code for functional units. A general correlation between an increase in protein size and protein function has been found for the enzymes in nucleotide metabolism (117), and it was shown that for selected enzymes subunit size corresponds to the total number of ligand binding functions, suggesting that proteins may be composed of ligand binding modules with an average size of 5 kDa and a range of 2–8 kDa (118). This size is consistent with the values in Fig. 1.

Table 1. Correspondence between exons and units of protein structure–function

Gene products	*n*	Correspondence [no. of gene(s)] Structural	Functional	None
Enzymes	18	4	11	3
Carrier proteins	3	—	2	1
Regulatory proteins	3	2	—	1
Hormones	4	—	4	—
Membrane proteins	8	1	7	—
Structural proteins	4	3	—	1

Included are genes coding for chicken lysozyme (3), rat elastase (76), rat chymotrypsin B (77, 78), triose phosphate isomerase from chicken (79) and maize (80), human calcium-activated neutral protease (81), rat carboxypeptidase (9), mouse lactate dehydrogenase A (82), chicken glyceraldehyde phosphate dehydrogenase (83, 84), mouse glycerol phosphate dehydrogenase (85), alcohol dehydrogenase from human (86) or maize (87), mouse renin (88), human phosphoglycerate kinase (89), chicken pyruvate kinase (90), human glycogen phosphorylase (91), human β-hexosaminidase (92), rabbit phosphofructokinase (16), soybean leghemoglobin (93), human β hemoglobin (4, 5), mouse α-fetoprotein (94), chicken calmodulin (95), chicken ovomucoid (6), mouse immunoglobulin heavy chain (7), rat cholecystokinin (96), rat glucagon (97), human epidermal growth factor (98), rat parathyroid hormone (99), mouse myelin basic protein (100), human myelin proteolipid protein (101), bovine rhodopsin (39, 102), mouse transplantation antigen (103), human acetylcholine receptor α subunit (104), chicken acetylcholine receptor γ and δ subunits (105), human low density lipoprotein receptor (106), mouse β-crystallin (107), chicken collagen α_1-II (108) and α_2 (109, 110), and nematode myosin heavy chain (111).

The Position of Introns. Introns are defined as type 0 (interrupting between codons), 1, or 2 (interrupting after the first or second nucleotide of a codon, respectively). The hypothesis advanced by Gilbert (1) that proteins evolved by shuffling and recombination of exons would appear to require that shuffling of such exons would not alter the reading frame when they are inserted in a new location. One way to accomplish this would be to have only one type of intron; i.e., all recombination events must consistently occur at equivalent positions. For the genes in the present survey, the average composition of introns is type 0 (54%), type 1 (27%), and type 2 (18%).

If, before the advent of splicing enzymes, primitive genes arose by recombination of exons, then self-splicing of the RNA without benefit of splicing enzymes, as exemplified by a number of existing examples (119, 120), might have removed introns. The emergence of splicing enzymes at some later time would then have evolved to recognize the most common sequence demarcating intron–exon junctions. If this is correct, then the current preponderance of type 0 introns reflects the ancestral state.

The distribution of observed intron types is shown in Table 2, and the total frequency of intron occurrence (1 intron per 41 codons) also defines the average size of exons. The possible number and proportion of intron types may be estimated from codon usage and the known consensus exon splice junction MAG–intron–GK, where M is A or C and K is G or U. Use of the consensus splice junction sequence (Table 2) leads to an underestimate of the actual number of introns and, therefore, to an overestimate of exon size; thus this consensus sequence, *per se*, is somewhat too restrictive as a predictor. The second consensus sequence MAG–in-

Table 2. Predicted and observed occurrence of introns

Occurrence of introns	Intron frequency, no. per 100 amino acids		
	Type 0	Type 1	Type 2
Observed	1.34 (54%)	0.67 (27%)	0.45 (18%)
Predicted			
MAG-intron-GK	0.96 (65%)	0.25 (17%)	0.27 (18%)
MAG-intron-GN	2.32 (66%)	0.65 (19%)	0.54 (15%)
NAG-intron-GN	3.70 (64%)	0.97 (17%)	1.11 (19%)

Values in parentheses are percentage of all introns. The predicted frequency for combinations of codons was calculated from a data set (121) for codon usage in humans, mice, rats, and chicken. As examples: the codons (AAG) ↓ (GGN) would define a type-0 splice junction, (NAA)(G ↓ GG) would define a type-1 splice junction, and (NNA)(AG ↓ G)(GNN) would define a type-2 splice junction, where the consensus splice junction is as shown.

tron–GN, where N is any base (Table 2), more accurately reflects the consensus sequence for the present data set, but its use overestimates the number of introns somewhat. Use of a more relaxed consensus sequence, NAG–intron–GN (Table 2), leads to a more generous estimate of the number of possible introns.

Whereas some differences in codon preference between widely diverging organisms have been noted, it is not clear how much effect such changes have on the prevalence of splice junction sequences during evolution. Since the observed size of exons (Table 2) falls between the values predicted for a stringent and a more relaxed splice junction sequence, it is then possible that most genes contain between 50% and 100% of the introns that they are capable of holding. Although there is clear evidence for intron deletion, this process appears to be neither rapid nor extensive for mammalian and avian genes. This suggests that the size of exons at present is not greatly different from exons in ancestral genes and is consistent with the suggestion that ancestral exons specified structure–function modules.

Table 3. Coincidence of intron position for domains duplicated within the same protein

Comparison	*n*	Size of domain,* no. of exons	Introns†			
			Similar positions		Similar positions/ same type	
			n	%	*n*	%
In same protein						
Plasma proteins/ receptors	5	1–5	52	90	51	88
Other enzymes/proteins	4	1–3	28	100	18	64
Between related proteins						
Calmodulins	3	1	8	67	8	67
Myosin alkali light chains plus parvalbumin	4	1	14	93	13	87

Plasma proteins/receptors includes human epidermal growth factor (98), human low density lipoprotein receptor (106), mouse α-fetoprotein (94), human haptoglobin Hp^2 (26), and human transferrin (27); other enzymes/proteins includes mouse renin (88), human pepsinogen (14), calcium-activated protease (81), and chicken ovomucoid (111); calmodulins includes calmodulins from chicken (95), sea urchin (126), and *D. melanogaster* (127); myosine alkali light chains plus parvalbumin includes myosin light chain genes from *D. melanogaster* (127), mouse (127), and chicken (127) and rat parvalbumin (126).

*Homologous domains coded by one or more exons were present in 2–7 copies within any specific protein.

†Intron positions flanking the homologous domains were compared either within the same protein or at corresponding positions between different, related proteins. Introns were scored as being at a similar position if they occurred within 5 codons of the site being compared and as being of the same type if they interrupted codons in the same fashion (for definition of intron types, see Table 2).

Comparison of intron positions for the same gene across many organisms has been used to show that some introns are maintained at the same positions and some are not (122, 124, 125). To limit sources of variability, the position of introns was examined in genes containing two or more homologous domains that were presumably generated by gene duplication plus fusion (Table 3). This made it possible to compare homologous domains and the introns associated with them within the same protein (constant environment) and also between different related proteins (potentially different environments). For comparisons within proteins, 90% or more of introns are maintained at corresponding positions in the duplicated regions (Table 3); this represents a measure of intron loss or acquisition that is very low. Furthermore, of introns found at similar positions, almost all are of the same intron type. This difference in frequency (Table 3) between introns of "same type" and introns "at similar position" is an indicator of the extent of intron sliding.

Comparisons between proteins were made with two groups of proteins containing three or four homologous domains that bind calcium. Again, a fairly high proportion of introns is found to be retained at corresponding positions (Table 3). This intron pattern must be >600 million years old, since it is found in genes from *Drosophila melanogaster*, sea urchin, and chicken.

Modifications of Exons. Illustrated in Fig. 2 is an ancestral gene (gene A) containing four exons and three introns. Each

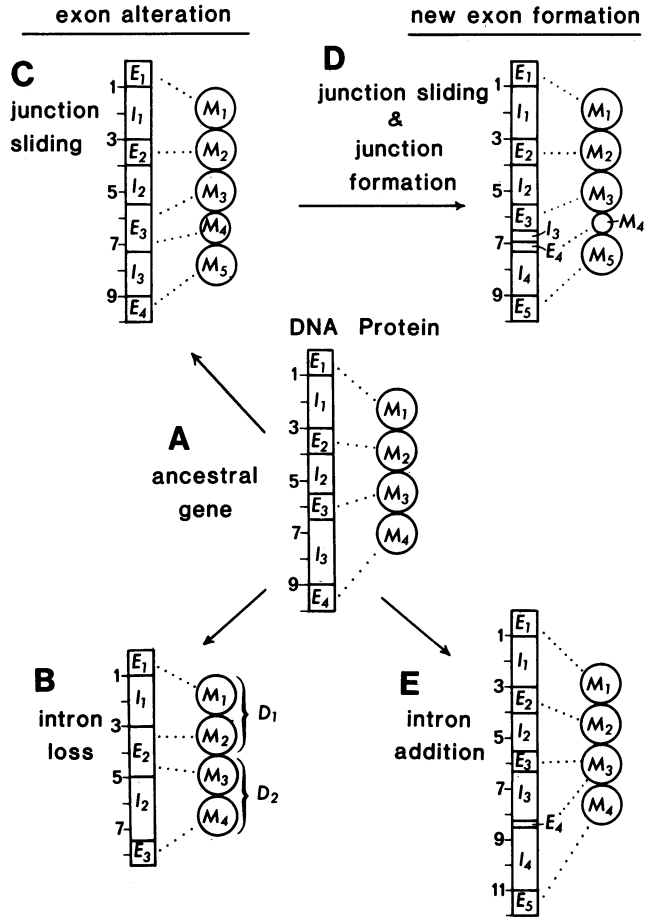

FIG. 2. Mechanisms for altering exons. Purely to facilitate comparison, the exons in the ancestral gene (gene A) are uniform in size, and an arbitrary size scale is shown for the DNA. Genes B–E are the result of various mechanisms. Abbreviations: E, exon; I, intron; D, domain; M, module.

exon is presumed to specify one structure–function module in the protein. The correspondence between exons and modules may become obscured by the mechanisms suggested by Table 3. Intron deletion would cause two exons to become fused into one larger exon. It is now quite possible for the junction between two domains to map to the middle of an exon (Fig. 2, gene B), and this would be interpreted as a lack of correspondence between exons and modules or domains. Junction sliding into an intron could add extra amino acids (perhaps even an additional module) to the protein structure (Fig. 2, gene C). Also, there are two processes by which an additional exon could be formed (Fig. 2, genes D and E). It is likely that such additional exons will code for a much smaller than average number of amino acids, and very small internal exons may be diagnostic for mechanisms that formed genes D and E.

The foregoing analysis shows that in existing genes exons clearly do not consistently relate to units of structure or function in proteins. However, what we observe could be a degenerate pattern that may still give evidence about the origin and evolution of proteins.

This work was supported by Grant DMB-8310902 from the National Science Foundation. I am grateful to Eric Baldwin for suggesting the analysis of codon usage.

1. Gilbert, W. (1978) *Nature (London)* **271,** 501.
2. Blake, C. C. F. (1978) *Nature (London)* **273,** 267.
3. Jung, A., Sippel, A. E., Grez, M. & Schütz, G. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 5759–5763.
4. Craik, C. S., Buchman, S. R. & Beychok, S. (1981) *Nature (London)* **291,** 87–90.
5. Eaton, W. A. (1980) *Nature (London)* **284,** 183–185.
6. Stein, J. P., Catterall, J. F., Kristo, P., Means, A. R. & O'Malley, B. W. (1980) *Cell* **21,** 681–687.
7. Sakano, H., Rogers, J. H., Hüppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R. & Tonegawa, S. (1979) *Nature (London)* **277,** 627–633.
8. Wodak, S. J. & Janin, J. (1981) *Biochemistry* **20,** 6544–6552.
9. Quinto, C., Quiroga, M., Swain, W. F., Nikovits, W. C., Jr., Standring, R. N., Pictet, R. L., Valenzuela, P. & Rutter, W. J. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 31–35.
10. Leicht, M., Long, G. L., Chandra, T., Kurachi, K., Kidd, V. J., Mau, M., Jr., Davie, E. W. & Woo, S. L. C. (1982) *Nature (London)* **297,** 655–660.
11. Naora, A. & Deacon, N. J. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 6196–6200.
12. Nunberg, J. H., Kaufman, R. J., Chang, A. C. Y., Cohen, S. N. & Schimke, R. T. (1980) *Cell* **19,** 355–364.
13. McKnight, G. L., O'Hara, P. J. & Parker, M. L. (1986) *Cell* **46,** 143–147.
14. Sogawa, K., Fujii-Kuriyama, Y., Mizukami, Y., Ichihara, Y. & Takahashi, K. (1983) *J. Biol. Chem.* **258,** 5306–5311.
15. MacDonald, R. J., Crerar, M. C., Swain, W. F., Pictet, R. L., Thomas, G. & Rutter, W. J. (1980) *Nature (London)* **287,** 117–122.
16. Lee, C.-P., Kao, M.-C., French, B. A., Putney, S. D. & Chang, S. H. (1987) *J. Biol. Chem.* **262,** 4195–4199.
17. Craik, C. S., Choo, Q.-L., Swift, G. H., Quinto, C., MacDonald, R. J. & Rutter, W. J. (1984) *J. Biol. Chem.* **259,** 14255–14264.
18. Campbell, R. O. & Porter, R. R. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 4464–4468.
19. Quan, F., Korneluk, R. G., Tropak, M. B. & Gravel, R. A. (1986) *Nucleic Acids Res.* **14,** 5321–5335.
20. Kim, S. H., Moores, J. C., David, D., Respess, J. G., Jolly, D. J. & Friedmann, T. (1986) *Nucleic Acids Res.* **14,** 3103–3118.
21. Patel, P. I., Framson, P. E., Caskey, C. T. & Chinault, A. C. (1986) *Mol. Cell. Biol.* **6,** 393–403.
22. Maguire, D. J., Day, A. R., Borthwick, I. A., Srivastava, G., Wigley, P. L., May, B. K. & Elliott, W. H. (1986) *Nucleic Acids Res.* **14,** 1379–1391.
23. Merrill, G. F., Harland, R. M., Groudine, M. & McKnight, S. L. (1984) *Mol. Cell. Biol.* **4,** 1769–1776.
24. Lewis, J. A. (1986) *Mol. Cell. Biol.* **6,** 1998–2010.
25. Freytag, S. D., Beaudet, A. C., Bock, H. G. O. & O'Brien, W. E. (1984) *Mol. Cell. Biol.* **4,** 1978–1984.
26. Maeda, N., Yang, F., Barnett, D. R., Bowman, B. B. & Smithies, O. (1984) *Nature (London)* **309,** 131–135.
27. Park, I., Schaeffer, E., Sidoli, A., Baralle, F. E., Cohen, G. N. & Zakin, M. M. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 3149–3153.
28. Cochet, M., Gannon, F., Heu, R., Maroteaux, L., Perrin, F. & Chambon, P. (1979) *Nature (London)* **282,** 567–574.
29. Costanzo, F., Colombo, M., Staempfli, S., Santoro, C., Marone, M., Frank, R., Delius, H. & Cortese, R. (1986) *Nucleic Acids Res.* **14,** 721–736.
30. Santoro, C., Marone, M., Ferrone, M., Costanzo, F., Colombo, M., Minganti, C., Cortese, R. & Silengo, L. (1986) *Nucleic Acids Res.* **14,** 2863–2876.
31. Heilig, R., Perrin, F., Gannon, F., Mandel, J. L. & Chambon, P. (1980) *Cell* **20,** 625–637.
32. Karin, M. & Richards, R. I. (1982) *Nature (London)* **299,** 797–802.
33. Heguy, A., West, A., Richards, R. I. & Karin, M. (1986) *Mol. Cell. Biol.* **6,** 2149–2157.
34. Blanchetot, A., Wilson, V., Wood, D. & Jeffreys, A. J. (1983) *Nature (London)* **301,** 732–734.
35. Proudfoot, N. J. & Maniatis, T. (1980) *Cell* **21,** 537–544.
36. Nobrega, F. G. & Tzagoloff, A. (1980) *J. Biol. Chem.* **255,** 9828–9837.
37. Mizukami, Y., Sogawa, K., Suwa, Y., Muramatsu, M. & Fujii-Kuriyama, Y. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3958–3962.
38. Bonitz, S. G., Coruzzi, G., Thalenfeld, B. E., Tzagoloff, A. & Macino, G. (1980) *J. Biol. Chem.* **255,** 11927–11941.
39. Nathans, J., Thomas, D. & Hogness, D. S. (1986) *Science* **232,** 193–202.
40. Ishida, N., Kanamori, H., Noma, T., Nikaido, T., Sabe, H., Suzuki, N., Shimizu, A. & Honjo, T. (1985) *Nucleic Acids Res.* **13,** 7579–7589.
41. Prochownik, E. V., Bock, S. C. & Orkin, S. H. (1985) *J. Biol. Chem.* **260,** 9608–9612.
42. Liao, Y. C., Taylor, J. M., Vannice, J. L., Clawson, G. A. & Smuckler, E. A. (1985) *Mol. Cell. Biol.* **5,** 3634–3639.
43. Brush, D., Dodgson, J. B., Choi, D. R., Stevens, P. W. & Engel, J. D. (1985) *Mol. Cell. Biol.* **5,** 1307–1317.
44. Rhoads, D. D., Dixit, A. & Ronfa, D. J. (1986) *Mol. Cell. Biol.* **6,** 2774–2783.
45. Nabeshima, Y., Fujii-Kuriyama, Y., Muramatsu, M. & Ogata, K. (1984) *Nature (London)* **308,** 333–338.
46. Karlik, C. C. & Fyrberg, E. A. (1986) *Mol. Cell. Biol.* **6,** 1965–1973.
47. Fornwald, J. A., Kuncio, G., Peng, I. & Ordahl, C. P. (1982) *Nucleic Acids Res.* **10,** 3861–3876.
48. Nakajima-Iijima, S., Homada, H., Reddy, P. & Kakunaga, T. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 6133–6137.
49. Ueyama, H., Hamada, H., Battula, N. & Kakunaga, T. (1984) *Mol. Cell. Biol.* **4,** 1073–1078.
50. Quax-Jeuken, Y., Quax, W., van Rens, G., Khan, P. M. & Bloemendal, H. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 5819–5823.
51. Balcarek, J. M. & Cowan, N. J. (1985) *Nucleic Acids Res.* **13,** 5527–5543.
52. Johnson, L. D., Idler, W. W., Zhou, X.-M., Roop, D. R. & Steinert, P. M. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1896–1900.
53. Mason, D. J., Evans, B. A., Cox, D. R., Shino, J. & Richards, R. I. (1983) *Nature (London)* **303,** 300–307.
54. Lemishka, I. & Sharp, P. A. (1982) *Nature (London)* **300,** 330–335.
55. Hall, J. L. & Cowan, N. J. (1985) *Nucleic Acids Res.* **13,** 207–223.
56. Ferrari, S., Battini, R., Kaczmarek, L., Rittling, S., Calabretta, B., de Riel, J. K., Philliponis, V., Wei, J.-F. & Baserga, R. (1986) *Mol. Cell. Biol.* **6,** 3614–3620.
57. Boss, J. M., Mengler, R., Okada, K., Auffray, C. & Strominger, J. L. (1985) *Mol. Cell. Biol.* **5,** 2677–2683.
58. Montminy, M. R., Goodman, R. H., Horovitch, S. J. & Habener, J. F. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 3337–3340.

59. Nemer, M., Chamberland, M., Sirois, D., Argentin, S., Drouin, J., Dixon, R. A. F., Zivin, R. A. & Condra, J. H. (1984) *Nature (London)* **312,** 654–656.
60. Holbrook, N. J., Smith, K. A., Fornace, A. J., Comeau, C. M., Wiskocil, R. L. & Crabtree, G. R. C. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 1634–1638.
61. Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R. & Tizard, R. (1979) *Cell* **18,** 545–558.
62. Ullrich, A., Dull, T. J., Gray, A., Brosius, J. & Sures, I. (1980) *Science* **209,** 613–614.
63. Barta, H., Richards, R. I., Baxter, J. D. & Shine, J. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 4867–4871.
64. De Noto, F. M. & Goodman, H. M. (1981) *Nucleic Acids Res.* **9,** 3719–3730.
65. Nakamishi, S., Teranishi, Y., Watanabe, Y., Notake, M., Noda, M., Kakidami, H., Jingami, H. & Numa, S. (1981) *Eur. J. Biochem.* **115,** 429–438.
66. Cochet, M., Young, A. C. Y. & Cohen, S. N. (1982) *Nature (London)* **297,** 335–339.
67. Chien, Y.-H. & Thompson, E. B. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 4583–4587.
68. Cohen, D. R., Hapel, A. J. & Young, I. G. (1986) *Nucleic Acids Res.* **14,** 3641–3658.
69. Van den Ouweland, A. M. W., van Groningen, J. J. M., Schalken, J. A., van Neck, H. W., Bloemers, H. P. J. & van de Ven, J. M. (1987) *Nucleic Acids Res.* **15,** 959–970.
70. Hall, H. & Brown, R. (1985) *Nucleic Acids Res.* **13,** 5255–5268.
71. Lavu, S. & Reddy, E. P. (1986) *Nucleic Acids Res.* **14,** 5309–5320.
72. Hickey, E., Brandon, S. E., Potter, R., Stein, G., Stein, J. & Weber, L. A. (1986) *Nucleic Acids Res.* **14,** 4127–4145.
73. Jeltsch, J. M., Roberts, M., Schatz, C., Garnier, J. M., Brown, A. M. C. & Chambon, P. (1987) *Nucleic Acids Res.* **15,** 1401–1414.
74. Nishizawa, M., Semba, K., Yoshida, M. C., Yamamoto, T., Sasaki, M. & Toyoshima, K. (1986) *Mol. Cell. Biol.* **6,** 511–517.
75. Lamb, P. & Crawford, L. (1986) *Mol. Cell. Biol.* **6,** 1379–1385.
76. Swift, G. H., Craik, C. S., Stary, S. J., Quinto, C., Lahuie, R. G., Rutter, W. J. & MacDonald, R. J. (1984) *J. Biol. Chem.* **259,** 14271–14278.
77. Craik, C. S., Sprang, S., Fletterick, R. & Rutter, W. J. (1982) *Nature (London)* **299,** 180–182.
78. Bell, G. I., Quinto, C., Quiroga, M., Valenzuela, P., Craik, C. S. & Rutter, W. J. (1984) *J. Biol. Chem.* **259,** 14265–14270.
79. Straus, D. & Gilbert, W. (1985) *Mol. Cell. Biol.* **5,** 3497–3506.
80. Marchionni, M. & Gilbert, W. (1986) *Cell* **46,** 133–141.
81. Miyake, S., Emori, Y. & Suzuki, K. (1986) *Nucleic Acids Res.* **14,** 8805–8817.
82. Li, S. L., Tiano, H. F., Fukasawa, K. M., Yagi, K., Shimizu, M., Sharief, F. S., Nakashima, Y. & Pan, Y. E. (1985) *Eur. J. Biochem.* **149,** 215–225.
83. Stone, E. M., Rothblum, K. N. & Schwartz, R. J. (1985) *Nature (London)* **313,** 498–500.
84. Stone, E. M., Rothblum, K. N., Alevy, M. C., Kuo, T. M. & Schwartz, R. J. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1628–1632.
85. Ireland, R. C., Kotarski, M. A., Johnston, L. A., Stadler, U., Birkenmeier, E. & Kozak, L. P. (1986) *J. Biol. Chem.* **261,** 11779–11785.
86. Duester, G., Jörnvall, H. & Hatfield, G. W. (1986) *Nucleic Acids Res.* **14,** 1931–1940.
87. Bränden, C.-I., Eklund, H., Cambillau, C. & Pryor, A. J. (1984) *EMBO J.* **3,** 1307–1310.
88. Holm, I., Ollo, R., Panthier, J.-J. & Rougeon, F. (1984) *EMBO J.* **3,** 557–562.
89. Michelson, A. M., Blake, C. C. F., Evans, S. T. & Orkin, S. H. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 6965–6969.
90. Lonberg, N. & Gilbert, W. (1985) *Cell* **40,** 81–90.
91. Burke, J., Hwang, P., Anderson, L., Lebo, R., Gorin, F. & Fletterick, R. (1987) *Proteins* **2,** 177–187.
92. Proia, R. L. & Soravia, E. (1987) *J. Biol. Chem.* **262,** 5677–5681.
93. Jensen, E. O., Paludan, K., Hyldig-Nielsen, J. J., Jorgensen, P. & Marcker, K. A. (1981) *Nature (London)* **291,** 677–679.
94. Eiferman, F. A., Young, P. R., Scott, R. W. & Tilghman, S. M. (1981) *Nature (London)* **294,** 713–718.
95. Simmen, R. C. M., Tanaka, T., Ts'ui, K. F., Putkey, J. A., Scott, M. J., Lai, E. C. & Means, A. R. (1985) *J. Biol. Chem.* **260,** 907–912.
96. Deschenes, R. J., Haun, R. S., Funckes, C. L. & Dixon, J. E. (1985) *J. Biol. Chem.* **260,** 1280–1286.
97. Heinrich, G., Gros, P. & Habener, J. F. (1984) *J. Biol. Chem.* **259,** 14082–14087.
98. Bell, G. I., Fong, N. M., Stempien, M. M., Wormsted, M. A., Caput, D., Ku, L., Urdea, M. S., Rall, L. B. & Sanchez-Pescador, R. (1986) *Nucleic Acids Res.* **14,** 8427–8446.
99. Heinrich, G., Kronenberg, H. M., Potts, J. T., Jr., & Habener, J. F. (1984) *J. Biol. Chem.* **259,** 3320–3329.
100. De Ferra, F., Engh, H., Hudson, L., Kamholz, J., Puckett, C., Molineaux, S. & Lazzarini, R. A. (1985) *Cell* **43,** 721–727.
101. Diehl, H.-J., Schaich, M., Budzinski, R.-M. & Stoffel, W. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 9807–9811.
102. Nathans, J. & Hogness, D. S. (1983) *Cell* **34,** 807–814.
103. Steinmetz, M., Moore, K. W., Frelinger, J. G., Sher, B. T., Shen, F.-W., Boyse, E. A. & Hood, L. (1981) *Cell* **25,** 683–692.
104. Noda, M., Furutani, Y., Takahashi, H., Toyosato, M., Tanabe, T., Shimizu, S., Kikyotani, S., Kayano, T., Hirose, T., Inayama, S. & Numa, S. (1983) *Nature (London)* **305,** 818–823.
105. Nef, P., Mauron, A., Stalder, R., Alliod, C. & Ballivet, M. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 7975–7979.
106. Südhof, T. C., Goldstein, J. L., Brown, M. S. & Russel, D. W. (1985) *Science* **228,** 815–822.
107. Inana, G., Piatigorsky, J., Norman, B., Slingsby, C. & Blundell, T. (1983) *Nature (London)* **302,** 310–315.
108. Upholt, W. B. & Sandell, L. J. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 2325–2329.
109. Ohkubo, H., Vogeli, G., Mudryj, M., Avvedimento, V. E., Sullivan, M., Pastan, I. & de Crombrugghe, B. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 7059–7063.
110. Tate, V., Finer, M., Boedtker, H. & Doty, P. (1982) *Cold Spring Harbor Symp. Quant. Biol.* **37,** 1039–1049.
111. Karn, J., Gibb, N. J. & Miller, D. M. (1985) *Cell Muscle Motil.* **6,** 185–237.
112. Greenberg, B. D., Bencen, G. H., Seilhamer, J. J., Lewicki, J. A. & Fiddes, J. C. (1984) *Nature (London)* **312,** 656–658.
113. Hu, M. C., Sharp, S. B. & Davidson, N. (1986) *Mol. Cell. Biol.* **6,** 15–25.
114. Janin, J. & Wodak, S. J. (1983) *Prog. Biophys. Mol. Biol.* **42,** 21–78.
115. Zehfus, M. H. & Rose, G. D. (1986) *Biochemistry* **25,** 5759–5765.
116. Wetlaufer, D. B. (1981) *Adv. Protein Chem.* **34,** 61–92.
117. Traut, T. W. (1988) *CRC Crit. Rev. Biochem.*, in press.
118. Traut, T. W. (1986) *Mol. Cell. Biochem.* **70,** 3–10.
119. Zaug, A. J. & Cech, T. R. (1986) *Science* **231,** 470–475.
120. Kaine, B. P., Gupta, R. & Woese, C. R. (1985) *Proc. Natl. Acad. Sci. USA* **80,** 3309–3312.
121. Maruyama, T., Gojobori, T., Aota, S. & Ikemura, T. (1986) *Nucleic Acids Res.* **14,** r151–r196.
122. Craik, C. S., Rutter, W. J. & Fletterick, R. (1985) *Science* **220,** 1125–1129.
123. Argos, P. & Mohana Rao, J. K. (1985) *Biochim. Biophys. Acta* **827,** 283–297.
124. Gilbert, W., Marchionni, M. & McKnight, G. (1986) *Cell* **46,** 151–154.
125. Lewin, R. (1984) *Science* **226,** 328.
126. Berchtold, M. W., Epstein, P., Beaudet, A. L., Payne, M. E., Heizmann, C. W. & Means, A. R. (1987) *J. Biol. Chem.* **262,** 8696–8701.
127. Smith, V. L., Doyle, K. E., Maune, J. F., Munjaal, K. P. & Beckingham, K. (1987) *J. Mol. Biol.* **196,** 471–485.