program that will attract bipartisan support for years to come—is not automatic. It is critically dependent on cooperation and assistance from the science community itself.

I am especially worried about the continued inability—or unwillingness—of the members of the science community to agree among themselves about priorities or to abide by their decisions when they can agree. Considering all the complaints I hear from that community—and I find that the level of complaint is much the same no matter what the R & D budget looks like—I would not think it necessary to remind them that these are tough times. I will add that, for anyone depending on federal funding, they are going to remain tough times for quite a while.

My experience in the past 2 years reinforces my conviction that the disciplines which present well-considered, unified agendas for research have the best chance of getting support for their programs. After all, in the absence of agreed-upon recommendations, what can we expect the nonscientists who allocate funds to base their decisions on?

There are three choices, none of them good. It may be that funding increases will simply be deferred until the community can come to some consensus. Or decisions may be based on such nonscientifically relevant factors as preservation of politically popular facilities. Or disaffected minority viewpoints, when they are the dominant messages transmitted to the decision-makers, may well carry the day. The central point is that the community has to be willing to establish its own priorities and then stand by them in the public arena.

From my perspective, I would say that the coming year could prove very important for the future of American basic research. The Administration's proposals have been very well received so far. There is every reason to expect that we will see broad bipartisan support for most of the elements of the plan. This favorable reception, if it is supported by the science community and by industry, may set a course for a healthy and beneficial new degree of integration of science and technology in American life.

Reference

1. G. A. Keyworth, II, *Science* **219**, 801 (1983).

RESEARCH ARTICLE

Splice Junctions: Association with Variation in Protein Structure

Charles S. Craik, William J. Rutter, Robert Fletterick

Abstract. *A comparison between eukaryotic gene sequences and protein sequences of homologous enzymes from bacterial and mammalian organisms shows that intron-exon junctions frequently coincide with variable surface loops of the protein structures. The altered surface structures can account for functional differences among the members of a family. Sliding of the intron-exon junctions may constitute one mechanism for generating length polymorphisms and divergent sequences found in protein families. Since intron-exon junctions map to protein surfaces, the alterations mediated by sliding of these junctions can be effected without disrupting the stability of the protein core.*

Numerous studies have revealed the existence of families of structurally and functionally homologous proteins (*1*). The members of these protein families present fundamentally similar tertiary structures yet can exhibit quite divergent amino acid sequences and can be of quite different size. Small variations in polypeptide length are usually manifest as loops on the protein surface. The structural and functional relationships among the members of protein families imply a kinship among their respective genes. Presumably they are descendents from a single primordial gene by duplication and subsequent divergence. However, the pathway for this diversification is not clear. Point mutations produce amino acid substitutions, but a mechanism for production of deletions or additions of peptides in the internal region of the proteins is not obvious.

Eukaryotic genes are fragmented; the coding regions (exons) are interrupted by untranslated segments (introns) that are removed by a splicing system. The introns are excised from the initial RNA transcript and the exons are joined prior to translation of the messenger RNA (mRNA) into the protein product. It has been postulated that the exons represent genetic building blocks that code for discrete structural or functional domains of the proteins (*2*). This hypothesis appears tenable for some systems but clearly fails for others (*3*). Particularly intriguing is the fact that the positions of introns in the genetic sequence map to the surface of the protein (*4*). This implies a relation between the intron-exon structure of the gene and the tertiary structure of the gene product. An analysis of gene structure and variation in protein sequence within gene families shows that intron-exon junction positions correspond with length variations within members of the protein family. This leads to the hypothesis that translation of intron-exon junctions along the genetic sequence (intron-exon junctional sliding) may be one mechanism to account for peptide sequence length variability within protein families.

For these studies, gene families were selected in which the gene sequences, amino acid sequences, and protein structures of family members are known. This information is available for a family of mammalian trypsin-like proteolytic enzymes that typically contain serine at the catalytically active site (the serine proteases) (*5*) and for a homologous bacterial proteinase. Similar information exists for a metabolic enzyme dihydrofolate reductase.

A comparison of gene structure and protein structure for these families re-

The authors are members of the Department of Biochemistry and Biophysics, University of California, San Francisco 94143.

quires accurate alignment of the protein sequences and, in particular, an account of the additional amino acids in the eukaryotic enzymes. The sequences for the serine proteases have been aligned by Delbaere *et al.* (*6*) and Greer (*7*) on the basis of the three-dimensional equivalence of the protein tertiary structures. Volz *et al.* (*8*) have also suggested an alignment for the dihydrofolate reductases. The positions of the introns for rat trypsin, chymotrypsin, and elastase are indicated on these aligned protein sequences in Table 1 (*9–11*). The intron positions for mouse dihydrofolate reductase are shown with the aligned protein sequences in Table 2 (*12*). Thirteen intron positions are found for the serine proteases and dihydrofolate reductases, 11 (85 percent) correspond to regions of the polypeptide chain that show length variations among the bacterial and eukaryotic enzymes.

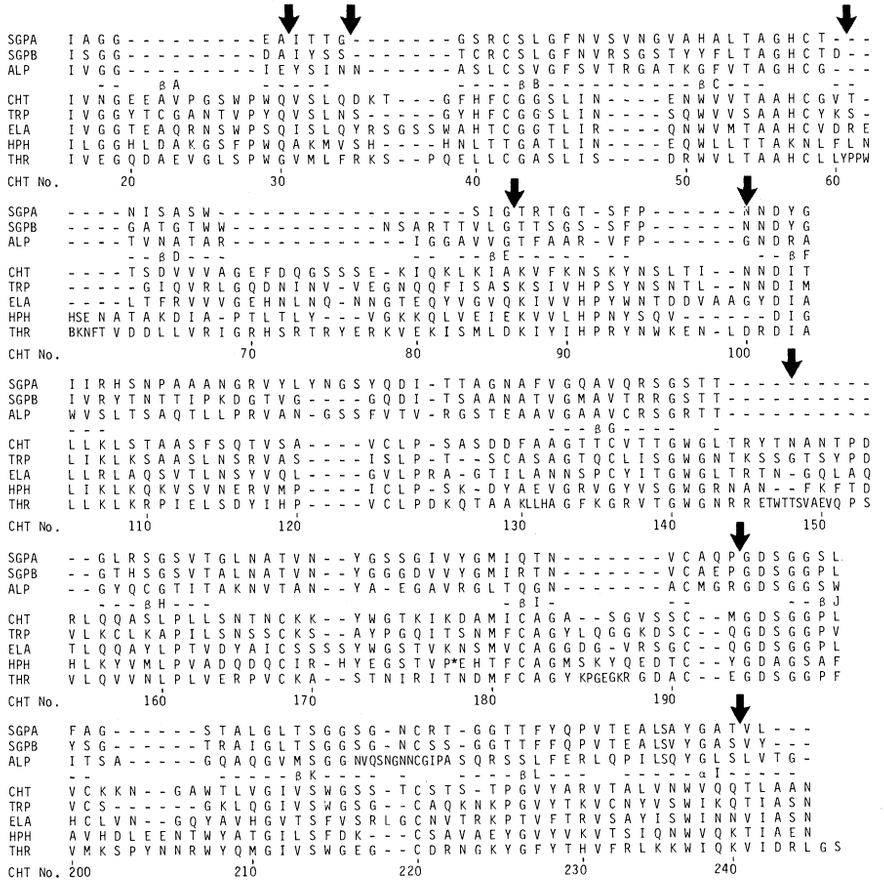
Serine Proteases

The variation in tertiary structure of porcine elastase compared with the bacterial homolog, *Streptomyces griseus* protease A (SGPA) is shown in Fig. 1. The α carbon atoms in the core of these two proteins (and chymotrypsin, trypsin, *S. griseus* protease B, and α-lytic protease) align to within 2 to 3 Å (*13*). Thus, the hydrophobic cores of the two domains are conserved with the structural differences arising primarily from variations in the length of 13 surface loops. The intron positions of the genes for trypsin (61, 150, 192) (*9*), chymotrypsin (34, 61, 87, 148, 192) (*10*), and elastase (30, 61, 100, 146, 192, 240) (*11*) map to 7 (64 percent) of these loops (see Fig. 1). The splice junction at position 192 corresponds to a deletion in the eukaryotic sequences when compared to the bacterial sequences while the six other cases involve insertions. The polypeptide length variation ranges from an insertion in thrombin of 17 amino acids corresponding to the splice junction found in the chymotrypsin gene at position 148, to a deletion of two amino acids from all of the eukaryotic sequences at the splice junction position 192.

Some of the variations in the polypeptide that coincide with splice junctions account for functional differences among the enzymes of the family. The splice junction at position 30 in the elastase genomic structure coincides with an addition of nine amino acids from threonine 20 through serine 28 when compared with SGPA (Table 1). This insert positions isoleucine 16 of elastase 8 Å closer to aspartate 194 (compared with SGPA) allowing an ionic interaction that is required for the active enzyme (*14*). There is an insert of eight amino acids in haptoglobin and an insert of 11 amino acids in thrombin at position 61. This loop may be involved in the complex of prothrombin and factor X (*15*). The splice junction mapping to amino acid 148 occurs within an insertion of ten amino acids which form the so-called autolysis loop (*16*) that is cleaved in chymotrypsin with no known change in its catalytic activity. Haptoglobin and thrombin also have additions at this position which may account for some aspect of their specialized physiological function (*17*). The intron at position 192 maps where two amino acids are deleted in the eukaryotic enzyme when compared with the bacterial proteases; that dipeptide may account for the specificity difference between elastase and α-lytic protease (*18*). A comparison of the NH_2-terminal sequences (not included in Table 1) of factor IX and prothrombin suggest the insertion of 112 amino acids (known as the S fragment) to prothrombin (*19*). This coincides with the intervening sequence at amino acid position 140 in the gene for factor IX (*20*). This additional peptide loop also appears to be involved in the interaction between prothrombin and factor X (*15, 21*).

Functional differences among the serine proteases have not been mapped to the insertions at positions 34, 100, and 240. The amino acids (namely, YPSGSSW) (*22*) inserted at position 34 of elastase form a surface loop which is absent in chymotrypsin and trypsin. One (trypsin), two (chymotrypsin), or three (elastase) amino acids are inserted at position 61 in these proteases compared to SGPA with no obvious functional correlation. This contrasts with the haptoglobin and prothrombin cases discussed above. An insertion of six amino acids at

Table 1. Alignment of serine protease family amino acid sequences. Arrows refer to the position of an intron-exon junction found in the genes for chymotrypsin B (CHT) E.C. 3.4.21.1 (*9*) (positions 34, 61, 87, 148, 192), elastase II (ELA) E.C. 3.4.21.11 (*10*) (positions 30, 61, 100, 146, 192, 240), or trypsin I (TRP) E.C. 3.4.21.4 (*11*) (positions 61, 150, 192). HPH refers to haptoglobin heavy chain and THR signifies thrombin (E.C. 3.4.21.5). SGPA and SGPB denote the bacterial proteases *S. griseus* protease A (E.C. 3.4.21) and *S. griseus* protease B (E.C. 3.4.21), respectively, while ALP refers to the bacterial α-lytic protease (E.C. 3.4.21.12). The alignment is based on tertiary structures of the homologs (*6, 7*). Secondary structural segments are marked. The asterisk (*) in the HPH sequence between positions 177 and 178 is where the sequence of residues EKKTPKSPVGVQPILN (*22*) have been removed to maximize homology.



position 100 of elastase relative to SGPA coincides with a splice junction but has no known associated function (*23*). Finally, an addition of three (or five in thrombin) amino acids at the COOH-terminal α-helix in the eukaryotic enzymes coincides with the splice junction at position 240.

Dihydrofolate Reductase

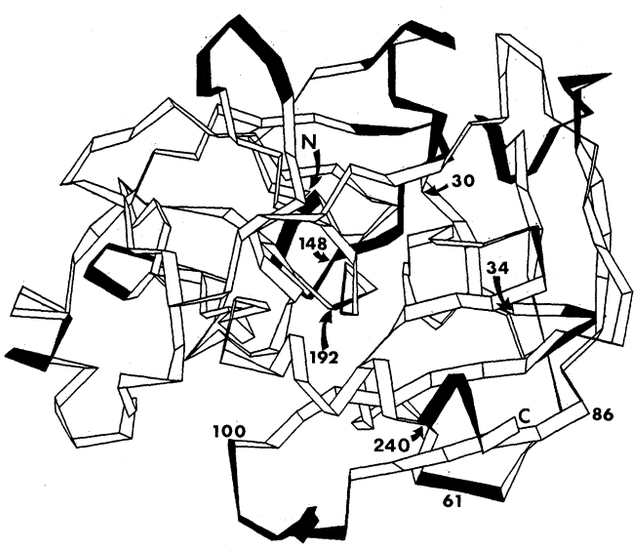
The gene structure of a single member of the dihydrofolate reductase family and the amino acid sequence of seven members of the family are known. The mouse gene has five splice junctions at positions 28, 45, 80, 122, and 161 (*12*), four of these (80 percent) coincide with additions in the polypeptide sequence, when the eukaryotic enzymes are compared to the bacterial homologs (Table 2 and Fig. 2). Since these additions occur at about 15 to 20 Å from the active site, they presumably have not significantly altered the catalytic function of the enzyme; however, the structural basis for the differences in the catalytic characteristics between the eukaryotic and bacterial enzymes is unknown. The addition of a single proline to the eukaryotic sequences at position 25 occurs two amino acids from the splice junction at position 28; it has no apparent effect on structure (*8*). The amino acids inserted in the eukaryotic sequences at positions 40 through 46 form a simple surface loop (TSSVEGK in chicken liver dihydrofolate reductase) that is 20 Å from the active site and coincides with a splice junction at position 45. The splice junction at position 122 coincides with an addition of two amino acids to helix αF in the eukaryotic enzyme. This addition extends the helix with no obvious effect on the structure of the enzyme. Finally, the insertion of six amino acids into an edge β strand, βG, occurs at the splice junction at 161. This insertion is accommodated by the polypeptide chain forming an external antiparallel β loop.

Discussion

Our results primarily concern the development of gene families, in particular, the variation in length of the gene products. The bacterial serine proteases comprise about 180 amino acids while the homologous mammalian enzymes comprise about 240 amino acids. Similarly, the bacterial dihydrofolate reductases are 160 amino acids and the mammalian enzymes are approximately 180 amino acids. Within both groups of proteins, the tertiary structures are similar. The length variation among the proteins discussed above ranges from −2 to 17 amino acids but are usually less than ten amino acids. These changes could in principle be affected by insertion of an exon; however, the added sequences are smaller than the usual exon size (*24*), and furthermore none of the segments are flanked by introns. We therefore conclude that exon insertion is an unlikely cause for this variation. However, these added segments do frequently coincide with intron positions in the genes. The evidence is more consistent with the hypothesis that length variation is caused by extension or contraction of exons at the intron junctions (Fig. 3).

Smith earlier suggested that the gaps and additional amino acid sequences that he and his colleagues observed in a homologous series of glutamate dehydrogenases might occur at splice junctions (*25*). In the two gene families that we studied, many of the splice junctions mark regions where the eukaryotic proteins have been significantly altered by additions or deletions of polypeptide chain segments. Some of these changes seem to account for new functions of the protein, but others provide no obvious advantage to the protein or the organism. Our rudimentary understanding of the structural and functional differences within a family is insufficient to recognize all significant relationships of these proteins. However, all changes introduced by junctional sliding would not be expected to result in a functional alteration. If the newly acquired polypeptide is the product of a mutated splicing event, then the expressed DNA sequence would extend beyond its previous intron-exon boundary. This should occur stochastically and may not always enhance function.

The sliding junction model reasonably explains the observed coincidence of splice junctions with additions, deletions, and variability in surface loops on related proteins. This hypothesis is sup-



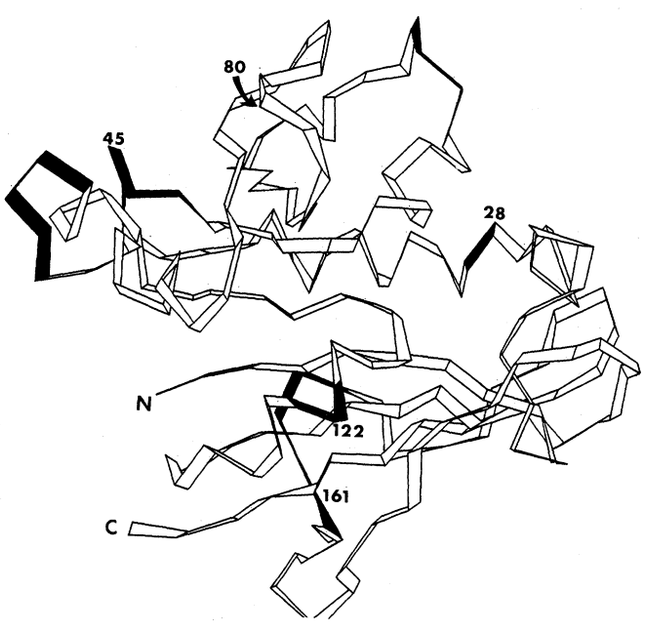


Fig. 1 (left). A drawing of the α-carbon backbone for porcine elastase. The solid regions represent length variations among the bacterial and eukaryotic serine proteases. Position 192 represents a deletion; the others are additions. The numbering scheme is that of bovine chymotrypsinogen. The positions corresponding to splice junctions in the eukaryotic serine protease genes are designated by numbers. The NH_2 and the COOH termini of the polypeptide chain are designated N and C, respectively. Fig. 2 (right). A drawing of the α-carbon backbone for chicken liver dihydrofolate reductase. The solid regions correspond to additions in comparing the bacterial and eukaryotic enzymes. The numbering scheme is that of chicken liver dihydrofolate reductase. The positions corresponding to splice junctions in the eukaryotic gene are designated by numbers.

ported by observations that alternate splicing of primary transcripts can generate different gene products [for example, hemoglobin (*26, 27*), amylase (*28*), calcitonin (*29*), ovomucoid (*30*), insulin (*31, 32*) and fibrinogen (*33*)]. In our study, many splice junctions (85 percent) coincide with regions of variability but there are other variable regions (8 of 19) that have no obvious relation to splice junctions. Where no correspondence is found, introns may have been present in other progenitor genes and have contributed to change in members of a family but may be absent in the gene sample investigated. Comparative studies of trypsin, chymotrypsin, and elastase genes (Table 1), the actin genes (*34*), and the insulin genes (*35*) indicate that the location of introns is not constant in the gene family. Introns can therefore be itinerant, but their residence within a gene is usually much longer than random mutational events, thus junctional sliding can occur and mediate significant changes in the gene product. Nevertheless, we believe it is unlikely that all length variation is associated with junctional sliding. Other mechanisms, such as those involving recombination or direct insertions, can also explain length polymorphism.

Table 2. Alignment of the dihydrofolate reductase family. The intron positions of the murine gene (*12*) for dihydrofolate reductase E.C. 1.5.1.3 are marked with arrows. SFII, LCA, and ECO are the bacterial sequences from *Streptococcus faecium* II, *Lactobacillus casei*, and *Escherichia coli* MB1428, respectively. CLI (chicken liver), BLI (bovine liver), 1210 (mouse lymphoma), and PLI (porcine liver) are the eukaryotic sequences. The alignment is based on the tertiary structure of the homologs (*8*). Secondary structural segments are marked.

SFII	M F I S M W A Q D K N G L I G K D G L L P W - R L P N D M R F F R E H T M - - - - - - - D K I L
LCA	T A F L W A Q N R D G L I G K D G H L P W - H L P D D L H Y F R A Q T V - - - - - - - G K I M
ECO	M I S L I A A L A V D R V I G M E N A M P W - N L P A D L A W F K R N T L - - - - - - - N K P V
	- - β A - - - - - - - - - α B - - - - - - - - - -
CLI	V R S L N S I V A V C Q N M G I G K D G N L P W P P L R N E Y K Y F Q R M T S T S H V E G K Q N A V
BLI	V R P L N S I V A V C Q D M G I G K N G D L P W P P L R N E F Q Y F Q R M T T V S S V E G K Q N L V
1210	V R P L N C I V A V S Q N M G I G K N G D L P W P P L R N E F K Y F Q R M T T T S S V E G K Q N L V
PLI	V R P L N C I V A V S Q N M G I G K N G D L P W P P L R N E Y K Y F Q R M T T T S S V E G K Q N L V
Cl #	1 10 20 40 50
SFII	V M G R K T Y E G M G - - K L S L P Y R H I I V L T T Q K D F K V E K N A E V L H S I D E L L A Y A K
LCA	V V G R R T Y E S F P - - K R P L P E R T N V V L T H Q E D Y Q A Q G - A V V V H D V A A V F A Y A K
ECO	I M G R H T W E S I G - - - R P L P G R K N I I L S S Q P - G T D D R - V T W V K S V D E A I A A C G
	β B - - - - α C - - - β C - - - - β D - - - - - - α E - - - -
CLI	I M G K K T W F S I P E K N R P L K D R I N I V L S R E L K E A P K G A H Y L S K S - L D D A L A L L
BLI	I M G R K T W F S I P E K N R P L K N R I N I V L S R E L K E P P K G A H F L A K S - L N D A L E L I
1210	I M G R K T W F S I P E K N R P L K D R I N I V L S R E L K E P P R G A H F L A K S - L D D A L R L I
PLI	I M G R K T W F S I P E K N R P L K D R I N I V L S R E L K E P P Q G A H F L A K S - L D D A L K L T
	51 60 70 80 90 100
SFII	D I P E - - - - - - - D I Y V S G G S R I F Q A L L - - P E T K I I W R T L I D A E F E G D T F I G E
LCA	Q H L D Q - - - - - - E L V I A G G A Q I F T A F K - - D D V D T L L V T R L A G S F E G D T K M I P
ECO	N V P - - - - - - - - E I M V I G G G R V Y E Q F L - - P K A Q K L Y L T H I D A E V E G D T H F P D
	- - β E - - - - - - - - α F - - - - - - - - - β F - - -
CLI	D S P E L K S K V D M V W I V G G T A V Y K A A M E K P I N H R L F V T R I L H E F E S D T F F P E
BLI	E D P E L T N K V B X V X I V G G S S V Y K Z A M D K P G H V R L F V T R I M Q E F E A D A F F P E
1210	E Q P E L A S K V D M V W I V G G S S V Y E Q A M N E P G H L R L F V T R I M Q E F E S D T F F P E
PLI	E Q P E L K D K V D M V W I V G G S S V Y K E A M N K P G H I R L F V T R I M K E F E S D T F F P E
	101 110 120 130 140 150
SFII	I D F T S F E L V E E H E G I - - V N Q E N Q Y - P H R F Q K W Q K M S K V V
LCA	L N W D D F T K V S S - - - - - - - R T V E D T N - P A L T H T Y E V W Q K K A
ECO	Y E P D D W E S V F S - - - - - - - E F H N A D A Q N S H S Y C F K I L E R R
	- - - - - - - - - β G - - - - - - - - - β H - - -
CLI	I D Y D K F K L L T E Y P G V P A D I Q E E D G - - - I Q Y K F E V Y Q K S V L A Q
BLI	I D F E K Y K L L P E Y P G V P L D V E E D Q G - - - I K Y K F E V Y E K N
1210	I D L E K Y K L L P E Y P G V L S E V Q E E D G - - - I K Y K F E V Y E K K D
PLI	I D L E K Y K L L S E C S G V P S D V Q E E K G - - - I K Y K F E V Y E K N N
	151 160 170 180

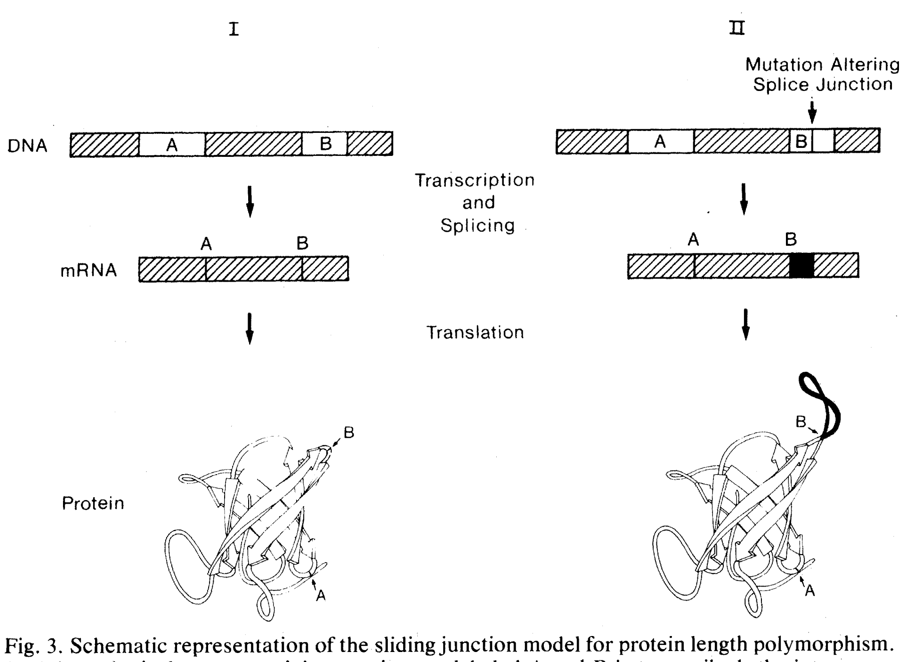


Fig. 3. Schematic representation of the sliding junction model for protein length polymorphism. (I) A hypothetical gene containing two introns labeled A and B is transcribed; the introns are excised precisely, and the RNA is spliced to form the functional mRNA. The mRNA is then translated into protein. The letters A and B on the protein mark the splice junctions. (II) A mutation alters the point of excision of intron B. This results in an extension of the exon and hence an added segment in the functional mRNA. On translation, this produces a protein which has an additional surface loop at the splice junction B.

In contrast to the general observation of variability at intron junctions, there are two instances in the present set of genes in which an intron exists within a region of strict length conservation. One is found in the serine proteases (the splice at position 87) and the other for the dihydrofolate reductases (the splice at position 80) (*36*). The α- and β-actin gene family for seven species also shows length conservation for all of the 11 intron positions (*34*). In the case of actin, the conservation may be required for function since the surface is essential to its function. Such a crucial function is not obvious for the two cases reported here.

There are two basic models that can be envisioned for the role of introns in the evolution of eukaryotic proteins. In one, exon encoded protein domains fuse by intron-mediated recombination; this exon shuffling model may account for the correspondence of domain boundaries with introns for certain proteins including the immunoglobulins (*37*), α-fetoprotein (*38*), and ovomucoid (*30*). In these and other proteins duplication of domains is accompanied by duplication of introns. The other model involves the disruption of a preexisting gene (the intron intrusion model); introns intrude upon, and under certain circumstances may escape from, a functional gene sequence. When present, introns can produce sequence variation with the consequence that coding sequences can be added or deleted. The altered DNA may encode new domains, variable length loops, deletions, or it may terminate.

The distinction between these related models is that in the first model, exon shuffling is an important feature of the origin of the gene (*39*). In the second model, intron intrusion may occur at any time, but presumably in most cases after formation of the gene. Apart from the origin and mobility of introns, junctional

sliding provides a means for diversification of genes. This important function may be one evolutionary reason for the existence of segmented genes (*40*). Eukaryotic organisms contain and selectively express families of related genes (for example, isozymes) to a greater degree than prokaryotic organisms. In distant evolutionary development, the presence of introns in the coding region provides an enhanced possibility for variation of the gene product. The localization of the intron-exon junctions on the protein surface tends to maximize positive or neutral changes, that is, the maintenance of function, instead of negative effects that would be expected if the intron-exon junction sites occurred in the central core region (*41*).

References and Notes

1. M. Dayhoff *et al.*, *Protein Sequence Database* (National Biomedical Research Foundation, Washington, D.C., 1982).
2. Gilbert's original hypothesis [W. Gilbert, *Nature (London)* **271**, 501 (1978)] suggests that the exons of eukaryotic genes code for functional domains and that the partitioned arrangement of coding information may serve to mediate the rapid evolution of new and unique proteins. Blake [C. C. F. Blake, *ibid.* **277**, 598 (1979)] suggested that exons may represent the domains of folded proteins. These ideas were formulated for the exon-encoded domain structure of the immunoglobulin gamma heavy chain (*37*) and have become an established notion for the immunoglobulin genes. Further support of this theory came from studies on the exon-encoded heme-binding domain of the β-globin gene [C. S. Craik, S. R. Buchman, S. Beychok, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 1384 (1980)] and from studies on the lysozyme gene [A. Jung, A. E. Sippel, M. Grez, G. Schütz, *ibid.*, p. 5759].
3. Examples of proteins with exons having no characterized independent structure or function are carboxypeptidase A [C. Quinto *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 31 (1982) and C. S. Craik, unpublished], α_1-antitrypsin [M. Leicht *et al.*, *Nature (London)* **297**, 655 (1982)], and α-amylase [R. J. MacDonald, M. M. Crerar, W. F. Swain, R. L. Pictet, G. Thomas, W. J. Rutter, *ibid.* **287**, 117 (1980)].
4. C. S. Craik, S. Sprang, R. Fletterick, W. J. Rutter, *Nature (London)* **299**, 180 (1982).
5. The serine proteases are members of a large family of homologous proteins which require the amino acid, serine, at their active site and appear to use the same mechanism for catalysis. Members include enzymes involved in digestion (trypsin, chymotrypsin, elastase), blood coagulation (thrombin), clot dissolution (plasmin), complement fixation (C1 protease), pain sensing (kallikrein), and fertilization (acrosomal enzyme). Haptoglobin, a member of the extended family which lacks the active site serine residue, is a plasma glycoprotein that forms a strong and stable complex with hemoglobin to aid the recycling of heme iron. For a discussion of the diversity of these enzymes, see R. M. Stroud [*Sci. Am.* **231**, 74 (July 1974)] and H. Neurath, K. A. Walsh, and W. P. Winter [*Science* **158**, 1638 (1967)].
6. L. T. J. Delbaere, W. L. B. Hutcheon, M. N. G. James, W. E. Thiessen, *Nature (London)* **257**, 758 (1975).
7. J. Greer, *J. Mol. Biol.* **153**, 1027 (1981).
8. K. W. Volz *et al.*, *J. Biol. Chem.* **257**, 2528 (1982).
9. G. Bell, C. Quinto, C. S. Craik, unpublished data.
10. G. Swift, R. MacDonald, C. S. Craik, unpublished data.
11. Q. L. Choo and C. S. Craik, unpublished data.
12. G. F. Crouse, C. C. Simonsen, R. N. McEwan, R. T. Schimke, *J. Biol. Chem.* **257**, 7887 (1982).
13. L. T. J. Delbaere, G. D. Brayer, M. N. G. James, *Can. J. Biol. Chem.* **57**, 135 (1979).
14. D. Shotton and J. Watson, *Nature (London)* **225**, 811 (1970).
15. J. Greer, *J. Mol. Biol.* **153**, 1043 (1981).
16. D. M. Blow, J. J. Birktoft, B. S. Hartley, *Nature (London)* **221**, 337 (1969).
17. J. Greer, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 3393 (1980).
18. L. T. J. Delbaere, G. D. Brayer, M. N. G. James, *Nature (London)* **279**, 165 (1979).
19. K. Katayama *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 4990 (1979).
20. K. H. Choo, K. G. Gould, D. J. G. Rees, G. G. Brownlee, *Nature (London)* **299**, 178 (1982).
21. S. Magnusson, T. E. Peterson, L. S. Jensen, H. Claeys, *Proteases and Biological Control*, E. Reich, D. Rifkin, E. Shaw, Eds. (Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1975), pp. 123–149.
22. Abbreviations for the amino acid residues are given in the standard one letter code: A, Ala, alanine; C, Cys, cysteine; D, Asp, aspartic acid; E, Glu, glutamic acid; F, Phe, phenylalanine; G, Gly, glycine; H, His, histidine; I, Ile, isoleucine; K, Lys, lysine; L, Leu, leucine; M, Met, methionine; N, Asn, asparagine; P, Pro, proline; Q, Gln, glutamine; R, Arg, arginine; S, Ser, serine; T, Thr, threonine; V, Val, valine; W, Trp, tryptophan; Y, Tyr, tyrosine.
23. Kallikrein (E.C. 3.4.21.8) is a serine protease involved in the processing of hormone precursors. Rat pancreatic kallikrein [G. H. Swift, J. C. Dagorn, P. A. Ashley, S. W. Cummings, R. J. MacDonald, *Proc. Natl. Acad. Sci. U.S.A.* **79** 7263 (1982)] contains an insertion of 15 amino acids (relative to SGPA) at position 100 referred to as the kallikrein autolysis loop. The proteolytic cleavage that occurs within this extended surface loop results in no known change in the catalytic activity of the enzyme.
24. H. Naora and N. J. Deacon, *ibid.*, p. 6196.
25. E. L. Smith, in *International Symposium on Frontiers of Bioorganic Chemistry and Molecular Biology, 1978*, S. N. Ananchenko, Ed. (Pergamon, New York, 1980), p. 39.
26. R. A. Spritz *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 2455 (1981).
27. M. Busslinger, N. Moschonas, R. Flavell, *Cell* **27**, 289 (1981).
28. O. Hagenbuchle, R. Bovey, R. A. Young, *ibid.* **21**, 179 (1980).
29. M. G. Rosenfeld *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 1717 (1982).
30. J. P. Stein, J. M. Catterall, P. Kristo, A. R. Means, B. W. O'Malley, *Cell* **21**, 681 (1980).
31. C. S. Craik, O. Laub, G. I. Bell, S. Sprang, R. J. Fletterick, W. J. Rutter, in *Gene Regulation*, B. O'Malley and C. F. Fox, Eds. (Academic Press, New York, 1982), p. 35.
32. O. Laub and W. J. Rutter, *J. Biol. Chem.*, in press.
33. G. R. Crabtree and J. A. Kant, *Cell* **31**, 159 (1982).
34. J. A. Farnwald, G. Kuncio, I. Peng, C. P. Ordahl, *Nucleic Acids Res.* **10**, 3861 (1982).
35. G. I. Bell, R. L. Pictet, W. J. Rutter, B. Cordell, E. Tischer, H. M. Goodman, *Nature (London)* **284**, 26 (1980).
36. Although length polymorphism does not coincide with the splice junction at position 87 in the serine proteases or at position 80 in the dihydrofolate reductases, the bacterial and eukaryotic proteins show sequence hypervariability in these regions (see Tables 1 and 2).
37. H. Sakano *et al.*, *Nature (London)* **277**, 598 (1979).
38. F. A. Eiferman, P. R. Young, R. W. Scott, S. M. Tilghman, *Nature (London)* **294**, 713 (1981).
39. Darnell [J. E. Darnell, Jr., *Science* **202**, 1257 (1978)] and Doolittle [W. F. Doolittle, *Nature (London)* **272**, 581 (1978)] have discussed the possibility that introns were part of a primordial gene. Eukaryotes would have utilized the noncontiguous nature of such a gene while prokaryotes eventually would have eliminated the introns, thereby "streamlining" their genomes.
40. Other reasons for split genes have been discussed. For example, see F. Crick, *Science* **204**, 264 (1979) and J. Lazawska, C. Jacq, P. P. Slonimski, *Cell* **22**, 333 (1980).
41. While this article was in press, a paper appeared in *Cell* **32**, 707 (1983) by C. R. King and J. Piatigorsky showing that one gene for murine α-crystalline generates two gene products which differ by a 23 amino acid insertion at an intron-exon boundary. By analogy to our study on the evolutionary development of gene families, these data substantiate our hypothesis that introns can produce sequence variation that results in added or deleted coding sequences.

42. We thank Joseph Kraut and David Matthews for providing us with the tertiary structure of chicken liver dihydrofolate reductase; Galvin Swift and Raymond MacDonald; Graeme Bell and Carmen Quinto; and Qui-Lim Choo for communicating their sequencing data on the genes for elastase, chymotrypsin, and trypsin, respectively, prior to publication; J. Kraut for discussion on the association of splice points with variability in dihydrofolate reductase; and Leslie Spector for technical assistance. Supported by NIH grant AM26081 (to R.J.F.) and NIH grant AM21344 and GM28520 (to W.J.R.), and an American Cancer Society postdoctoral fellowship (to C.S.C.).

6 December 1982; revised 4 April 1983