Genes-in-Pieces Revisited

Walter Gilbert

Mammalian genes are discontinuous, broken up along the DNA into alternating regions: coding sequences or exons, which are interspaced with other sequences, and introns that will be spliced out of the RNA transcript. What is the meaning of this arrangement?

In 1977, I conjectured that genes in eukaryotic cells arose as collections of exons brought together by recombination within intron sequences, and that the introns were the remnants of a process that speeded up evolution (*1, 2*). This hypothesis predicts that the exons code for useful portions of protein structure: functional regions, folding elements, domains, or subdomains—any segment that can be sorted independently during evolution (*3*). As relics of the recombination process that brought the exons together, the introns would be long, random sequences that would drift rapidly in sequence and size since the last act that assembled the gene.

Ford Doolittle (*4*) realized that there is no reason for this speeding of the evolutionary process to be restricted to eukaryotes, and he argued that the earliest organisms should have had split genes; the present day intron-less genomes of prokaryotes and of lower eukaryotes would then be the result of streamlining, a consequence of the evolutionary pressure for rapid replication. How well have these ideas fared?

Today we can draw some general conclusions about intron patterns. Introns are common in vertebrate genes, essentially absent in prokaryotes, and rare in lower eukaryotes such as yeast, with some dramatic exceptions such as the bithorax complex genes of drosophila. The distribution of exon sizes is rather narrow and peaks at 40 to 50 amino acid residues, but the sizes of introns scatter randomly and range from 50 to 10,000 to 20,000 bases in length. Since the introns are, on the average, so much larger than the exons, vertebrate genes have turned out to be much larger than we expected a decade ago, about 10 to 30 times larger than the coding sequence. The largest gene yet found is 200 kilobases long and 50-kilobase genes are not uncommon.

Where a protein has a structure containing repeated domains, the repeat is reflected in the intron distribution. For example, the basic domain of the immunoglobulins, the immunoglobulin fold, is carried on a single exon and repeated from one to five times as we compare β_2-microglobulin (*5*), the immunoglobulins (*6*), the histocompatibility antigens (*7*), and the T-cell receptors (*8*). The triple repeat of albumin (*9*), α-fetoprotein (*10*), or ovomucoid (*11*) clearly arose from a tripling of the underlying exon-intron structure. The helix of collagen was built up as a 40-fold repeat of an exon that bears a half turn (*12*). Thus the idea that introns serve to assemble the genes for proteins having repeating structures is well borne out.

Exons can often be correlated with functional elements of the encoded proteins. The hydrophobic signal sequences that tag a molecule for export are often carried on separate exons. Transmembrane, hinge, and cytoplasmic portions of immunoglobulins are similarly isolated. However, is there any pattern of the use of the same exon in different genes, when the same element of structure or function is required by different proteins? This would be the acid test of the theory. At last a dramatic example has been found.

In two important papers in this issue Südhof, Goldstein, Brown, and Russell (page 815), and Südhof, Russell, Goldstein, Brown, Sanchez-Pescador, and Bell (page 893), demonstrate the existence of exon shuffling. By analyzing the intron-exon structure of the gene for the low-density lipoprotein (LDL) receptor, the membrane receptor that binds the LDL particle and leads to its internalization within the cell, these workers found that the functional subdivisions of this protein are reflected in detail in the way this gene is broken up into 18 exons. More remarkably, however, Südhof *et al.* have shown that much of the LDL receptor gene is made up of exons recruited from other genes.

First, there is a stretch of 400 amino acids of the LDL receptor that shows 33 percent homology with the precursor for epidermal growth factor (EGF). This region of homology is encoded by eight contiguous exons in each gene. Of the nine introns involved, five are at identical positions, one has migrated a few codons, and three do not have mates. The most reasonable interpretation is that this region came as a whole from some common ancestral element and that the missing introns have subsequently been lost.

Second, a 40-amino-acid cysteine-rich stretch is repeated three times within this 400-amino-acid homologous segment. This sequence is encoded by a single exon, repeated thrice in both the LDL receptor and the EGF precursor genes, and arising once, as a separate exon, in the blood-clotting protein factor IX (*13*). Protein sequence homology suggests that this same exon will also be found in two other blood-clotting proteins, factor X and protein C (*14*).

Third, the LDL binding domain of the LDL receptor contains seven repeats of a 40-amino-acid sequence. Four of these occur on separate exons, while the other three occur on a single exon, as though two introns had been lost. This repeat element also occurs once in the complement factor C9; these authors predict that this will turn out to be a recurrence of that exon. The LDL receptor gene is thus a mosaic of exons derived from many diverse sources.

This work shows that introns have been used to assemble those genes that are the late products of evolution. But where did the introns come from?

There are two extreme alternatives. Either the introns are the vestigial linkers between useful coding sequences, left over from tying together simple reading frames at the beginning of evolution, or they arose by the insertion of DNA sequences into genes that were originally continuous. In the recent evolutionary period, we have only evidence for loss, as in the case of preproinsulin (*15*). We could interpret other anomalies, such as the ovalbumin—α-antitrypsin comparison (*16*) or the differences between the actin genes of plants (*17*), sea urchins (*18*), and vertebrates (*19*) as the loss of multiple introns. Can we distinguish between the possibility that preexisting introns have been lost by the lower eukaryotes and prokaryotes and the alternative, that introns have been created in

The author's address is 107 Upland Road, Cambridge, Massachusetts 02140.

the evolution that led to the vertebrates, either by insertion into a continuous gene or by the tying together of simpler gene elements?

A test is possible by examining old genes—genes whose products were in existence before the separation of the prokaryotes and eukaryotes. Such products are the enzymes for basic biochemical processes, which have the same three-dimensional structure in all cells. These ancient genes have a nonsplit structure in prokaryotes and in yeast. Do they have introns in the vertebrates? If so, are the introns randomly placed or do they correlate with structural elements?

The genes for three glycolytic enzymes whose three-dimensional structures are known have now been analyzed: glyceraldehyde phosphate dehydrogenase (GAPDH) (*20*), pyruvate kinase (PK) (*21*), and triose phosphate isomerase (TIM) (*22*), all from the chicken. All three genes have many introns; GAPDH has 11, PK has 9, and TIM has 6. The exons are quite regular in size, clustered about 30 to 40 amino acid residues in length, so the number of introns simply reflects the size of the protein. Three introns in GAPDH mark major domain borders as do two in PK, a third domain border in PK is not split. As we look more deeply into the tertiary structures of the domains encoded by the exons, we see that in PK and TIM most of the exons are compact pieces of the protein, modules in Go's sense (*23, 24*), each carrying one or two α-helical and β-sheet elements. The introns often mark turns or edges of secondary structure. The structures seem to be assembled out of the exon peptides; the intron positions are not random. But if these proteins were assembled to include the introns, then yeasts and prokaryotes, where the corresponding genes do not have introns, must have lost these dividers.

To attack this problem in another way we (*25*) have examined the TIM gene from a higher plant, maize, to see if its structure resembled that in the vertebrates. Maybe higher plants and animals would have similar gene structures that resemble the original gene in their single-cell common forebear more closely than do those genes of the lower eukaryotes that have evolved through so many more generations. Two-thirds of the maize TIM gene has been sequenced, revealing five introns. Three are at identical positions in corn and chicken, one has moved three codons over, and there is one extra intron in corn, at a bend in the last α-helix, an intron presumably lost from chicken TIM. The ancestral gene must have been already broken up in the eukaryotic progenitor cell before the time that the first algae and animal cells separated, probably more than a billion years ago, and the "lower" eukaryotes, such as yeasts and insects, must have lost introns as they evolved. The same argument applies to the prokaryotes; the organisms that went into symbiosis to form the eukaryotic cell probably all had genes made up of exons tied together by introns.

These ideas imply that the structure of the exon polypeptides must be telling us something profound about the "rules for creating proteins." Not only are proteins put together as mosaics of simpler structures, combinatorial assemblies of a smaller number of minigenes, but the folding principles may become apparent if we can understand the structure of the exon products and find the rules by which they were fitted together.

References

1. W. Gilbert, *Nature (London)* **271**, 501 (1978).
2. ——, *ibid., in Eucaryotic Gene Regulation: ICN-UCLA Symposia on Molecular and Cellular Biology*, R. Axel, T. Maniatis, C. F. Fox Eds. (Academic Press, New York, 1979), vol. 14, pp. 1-12.
3. C. C. F. Blake, *Nature (London)* **277**, 598 (1979); and see also *ibid.* **273**, 267 (1978) and *ibid.* **306**, 535 (1983).
4. W. F. Doolittle, *ibid.* **272**, 581 (1978).
5. J. R. Parnes and J. G. Seidman, *Cell* **29**, 662 (1982).
6. T. Honjo, *Annu. Rev. Immunol.* **1**, 499 (1983).
7. L. Hood, M. Steinmetz, B. Malissen, *ibid.*, p. 529.
8. M. Malissen *et al.*, *Cell* **37**, 1101 (1984).
9. T. D. Sargent, L. L. Jagodzinski, M. Yang, J. Bonner, *Mol. Cell. Biol.* **1**, 871 (1981).
10. F. A. Eiferman, P. R. Young, R. W. Scott, S. M. Tilghman, *Nature (London)* **294**, 713 (1981).
11. J. P. Stein, J. F. Catterall, P. Kristo, A. R. Means, B. W. O'Malley, *Cell* **21**, 681 (1980).
12. H. Boedtker and S. Aho, *Biochem. Soc. Symp.* **49**, 67 (1984).
13. D. S. Anson, K. H. Choo, D. J. G. Rees, F. Giannelli, K. Gould, J. A. Huddleston, G. G. Brownlee, *EMBO J.* **3**, 1053 (1984).
14. R. F. Doolittle, D.-F. Feng, M. S. Johnson, *Nature (London)* **307**, 558 (1984).
15. F. Perler, A. Efstratiadis, P. Lomedico, W. Gilbert, R. Kolodner, J. Dodgson, *Cell* **20**, 555 (1980).
16. M. Leicht, G. L. Long, T. Chandra, K. Kurachi, V. J. Kidd, M. Mace, E. W. Davie, S. L. C. Woo, *Nature (London)* **297**, 655 (1982).
17. D. M. Shah, R. C. Hightower, R. B. Meagher, *J. Mol. Appl. Genet.* **2**, 111 (1983).
18. A. D. Cooper and W. R. Crain, *Nucleic Acids Res.* **10**, 4081 (1982).
19. H. U. Ama *et al.*, *Mol. Cell Biol.* **4**, 1073 (1984).
20. E. M. Stone, K. N. Rothblum, R. J. Schwartz, *Nature (London)* **313**, 498 (1985).
21. N. Lonberg and W. Gilbert, *Cell* **40**, 81 (1985).
22. D. Straus and W. Gilbert, in preparation.
23. M. Go, *Nature (London)* **291**, 90 (1981).
24. ——, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1964 (1983).
25. M. Marchionni and W. Gilbert, unpublished.

15 April 1985