Introns and the Origin of Protein-Coding Genes

In their article, Arlin Stoltzfus *et al.* (*1*) use new techniques to assess the validity of the "exon theory of genes." This theory asserts that split genes arose early in evolution by the recombination of mini-genes, each corresponding to a protein domain. Stoltzfus *et al.* rightly conclude that the theory is untenable, but their analysis contains errors and does not accommodate some important evidence.

First, if one assumes that exons originated from mini genes, then why would one expect a mini gene to correspond to a protein domain? Walter Gilbert was the first to make this assumption (*2*), which was later extended to suit W. F. Doolittle's theory (*3*) that introns were present in the earliest genes. However, Stoltzfus *et al.* now appear to contradict Doolittle's original theory, saying that introns may have been introduced later into contiguous genes. But perhaps more significantly, the new data (*1*) can be better interpreted to yield different and more accurate conclusions. I proposed (*4*) that genes originated fully formed in long random primordial genetic sequences, where the random distribution of stop codons permitted only short reading frames (RFs), which precluded the occurrence of any genes longer than 200 amino acids. In fact, a negative exponential distribution of RFs in the random sequences constrained most RFs to zero length, and only rarely did they reach their upper limit of only 200 codons (600 nucleotides). The only way that longer, fully formed genes could have occurred, then, is if long random sequences dominated by clusters of stop codons were skipped during the reading of the consecutive short RFs between them. The consecutively spliced RFs (in the RNA) could then code for a long protein chain, and a protein with a biochemical function could then have emerged from many long proteins with random amino acid sequences. The short coding pieces (RFs) of these genes are what we now call exons, while the usually long intervening sequences are what we call introns.

Both Doolittle's exon theory of genes and the "split-gene" hypothesis I have proposed agree on the original existence of introns in genes, but for different reasons. Doolittle hypothesized that introns arose as spacers between ancient mini-genes, whereas my research shows that introns simply occurred naturally in fully formed genes as a result of the random occurrence of stop codons.

According to the split-gene hypothesis, the primary RNA copy of a gene would contain long introns with clusters of stop codons. This model suggests strongly that the nuclear boundary appeared in the very first cells, to prevent the translation of these primary RNAs by the ribosomes, which would have produced truncated, wasteful, and chaotic polypeptides, and thereby introduce a profound energy drain to the cell. The segregation of the cleanly spliced messenger RNAs (mRNAs) from the primary RNAs by the nuclear boundary is what makes possible the presentation of only the cleanly spliced mRNAs to the ribosomes in the cytoplasm. Thus the first cells that originated with split genes in the primordial pond were typical eukaryotic cells with a nucleus. If this hypothesis or "model" is correct, the structural features of split genes predicted from computer-simulated random sequences can be expected to occur in actual eukaryotic split genes. This is what we find in most known split genes in eukaryotes living today. The eukaryotic sequences exhibit a nearly perfect negative exponential distribution of RFs, with an upper limit of 600 nucleotides (with rare exceptions). Also, with rare exceptions, the exons in all known eukaryotic genes fall within this 600 nucleotide upper limit.

Moreover, if this hypothesis is correct, exons should be delimited by stop codons, especially at the 3′ end of exons (that is, the 5′ ends of introns). Actually they are precisely delimited more strongly at the 3′ ends of exons and less strongly at the 5′ ends in most known genes, as predicted (*5*). These stop codons are the most important functional parts of both splice junctions. The hypothesis thus provides an explanation for the "conserved" splice junctions at the ends of exons and for the loss of these stop codons along with introns when they are spliced out (from the primary RNA copy of the gene). If this hypothesis is correct, splice junctions should be randomly distributed in eurkaryotic DNA sequences, and they are (*6*). The splice junctions present in transfer RNA genes and ribosomal RNA genes, which do not code for proteins and wherein stop codons have no functional meaning, should not contain stop codons, and again, this is observed. Finally, the "lariat" sequence, another short sequence of about five characters that occurs upstream of most exons and aids in the splicing process, also contains stop codons (*7*). Colin Blake, one of the proponents of the Gilbert-Blake theory, has stated (*8*) that this split-gene hypothesis (*4*) comprehensively explains the ultimate origin of introns and the splicing process in primordial eukaryotic genes. Neither the exon theory of genes nor the intron-insertion theory provides an explanation for the structural features present in eukaryotic genes.

According to the split-gene hypothesis, we can predict what should be the correlation between the exons of a gene and the domains of the protein it encodes. The exons of the primordial genes should specify only short RFs, and the contiguously spliced exons should specify only a long protein with no correlation whatsoever between the exons and the amino acid sequence of the protein. Introns originated to circumvent the problem of the random distribution of stop codons in random primordial genetic sequences. Biologically meaningful proteins with functional domains and other structural paraphernalia were chosen only secondarily from the fairly long, random protein sequences coded by the spliced exons of the genes. Under these circumstances, the exons of a gene should correspond only randomly to the domain structure of its protein. This is observed in the present data (*1*).

The split-gene hypothesis does not preclude a rare role for introns once they had occurred in genes in the primordial pond as a result of the random sequence problem—for example, in the recombination of some protein modules that fortuitously and rarely corresponded with individual exons in various genes. Furthermore, the split-gene theory does not preclude the occasional loss of introns or the occasional insertion of introns, subsequent to their original appearance in genes. As noted in an article by Holland and Blake (*8*, p. 26),

> It is important to distinguish between the role and origin of introns, noting that the gene-shuffling hypothesis relates only to possibly an incidental intron function, in response to evolutionary pressures, and not to the origin of the split gene; otherwise the evolutionary potential inherent in the theory would imply non-Darwinian, anticipatory evolution.

Although Stoltzfus *et al.* agree that there is no correlation between the exon structure of genes and the domain structure of proteins, the evidence does not appear to support the hypothesis that introns were later inserted into genes. Moreover, the notion of intron insertion would not explain the ultimate origin of introns in split genes or of any of the structural features of genes. I have demonstrated how highly unlikely it would have been for long contiguous genes (like those in prokaryotes) to have occurred in primordial genetic sequences (*4*), although such genes would have been a prerequisite for intron insertion. The later loss of introns from original split genes—as proposed by W. F. Doolittle—is a much more likely scenario (*3*). Recent computer simulations show that split genes, fully formed with all their structural features and corresponding to complete proteins, likely could

have occurred in random primordial genetic sequences (9).

The prevailing interpretation of existing fossil evidence contradicts this conclusion that eukaryotic genes preceded those of prokaryotes, as prokaryotes occur 3.6 billion years ago, while the first unicellular eukaryotes do not appear until the Cambrian explosion, 3 billion years later. But a different interpretation of the same fossil record, compatible with the lost-introns hypothesis, seems more plausible than the astronomical improbability of a single contiguous gene coding for a specific protein (typical of prokaryotes) occurring purely by chance on Earth—or even in a random DNA molecule with the mass of the whole universe.

Moreover, the high probability that fully formed split genes did occur by chance, in only a small amount of random primordial DNA (9), suggests a more likely scenario: While split genes did occur first, their expression into the first eukaryotic organisms could not have occurred, or the organisms themselves could not have survived in the oxygen-free atmosphere of primordial Earth. But the loss of introns from these genes by exon recombination and gene processing produced a new variety of genes that lacked introns (4, 9). These new, contiguously coded genes found immediate expression in prokaryotic cells, which were able to thrive without oxygen. Only when sufficient oxygen became available, by the beginning of the Cambrian period, did split genes finally find expression in the first viable eukaryotic organisms. And as the fossil record shows, the eukaryotes—unicellular and multicellular alike—veritably bloomed over a short time once the requisite oxygen became available (9).

Periannan Senapathy
Genome International Corporation,
579 D'Onofrio Drive,
Suite 206,
Madison, WI 53719–2054, USA

REFERENCES

1. A. Stoltzfus, D. F. Spencer, M. Zucker, J. M. Logdon, W. F. Doolittle, *Science* **265**, 202 (1994).
2. W. Gilbert, *Nature* **271**, 501 (1978).
3. W. F. Doolittle, *ibid.* **272**, 581 (1978).
4. P. Senapathy, *Proc. Nat. Acad. Sci. U.S.A.* **83**, 2133 (1986).
5. ———, *ibid.* **85**, 1129 (1988).
6. ———, M. B. Shapiro, N. Harris, *Methods Enzymol.* **183**, 252 (1990).
7. N. Harris and P. Senapathy, *Nucleic Acids Res.* **18**, 3015 (1990).
8. S. K. Holland and C. C. F. Blake, in *Intervening Sequences in Evolution and Development*, E. M. Stone and R. J. Schwartz, Eds. (Oxford Univ. Press, New York, 1990), p. 32.
9. P. Senapathy, *Independent Birth of Organisms* (Genome, Madison, WI, 1994), pp. 10–45.

5 August 1994; accepted 19 January 1995

From their analysis of the gene structures of the ancient proteins alcohol dehydrogenase, globins, pyruvate kinase and triosephosphate isomerase, Stoltzfus *et al.* (*1*) were unable to find support for the view that "exons should encode discrete units of folded protein structure." However, this view is not an axiomatic requirement of the theory that introns are ancient and have been more recently lost, and it is improper to dismiss the "introns-early" idea on this basis. We have earlier proposed the "exon microgene" hypothesis (*2*), which assumes the ancient origin of exons and introns and is fully consistent with the lack of any correlation between exons and encoded peptides of credible three dimensional stability.

Rather than ascribing the nonrandom locations of introns to preferential intronic insertions at particular target sites, we suggested that exon-intron and intron-exon boundaries were originally determined early in evolution (before the archaebacteria-prokaryotic-eukaryotic division) by terminating amber codons (TAG) of polynucleotide segments. Each of these segments encoded separately translated peptides that spontaneously assembled to form catalytically-active complexes. By this hypothesis, the terminal AG of the consensus sequence at exon-intron and intron-exon boundaries for protein-encoding genes could derive from these original termination codons (*3*).

Although the fact that exon boundaries generally map to the surface of proteins (*4*) is not required by the idea that exons encode discrete structural units (*5*), it *is* an implicit corollary of the exon microgene theory. According to this theory, the surface locations of exon junctions would partly derive from the need to solvate and stabilize the charged termini of the originally independently-translated exon products.

"Introns-early" theories do not, therefore, require that exons must encode elements of defined protein structure. Because the exon microgene theory is not constrained by such a correlation, the discovery of "new" intron positions in homologs of previously-analyzed genes does not automatically "exceed the expectations" of this theory. Accordingly, the restricted phylogenetic distribution of introns seems more likely to result from the loss of introns from an ancestral intron-rich gene than from the independent gain of introns at identical positions in different species (even allowing for the possible preferential insertion of mobile introns at specific sites).

Another important question asked by Cerff *et al.* (*6*) of proponents of the "introns-late" view, is "how were genes and long contiguous open reading frames assembled in early evolution? (6, p. 527)" This question is not congruous with the introns-late hypothesis, given the statistical limit of the length of open reading frames before they are interrupted by stop codons (*7*). The idea of selective pressure accounting for the accumulation of ever longer open reading frames through evolution is not a satisfying answer, as many enzymes such as triosephosphate isomerase have critical active site residues encoded by different exons across the entire length of the gene.

Stoltzfus *et al.* (*1*) suggest "that the exon theory of genes" (and the ancestral origin of introns) "is untenable" because of the lack of correspondence between exons and discrete units of protein structure. Yet this correspondence is unnecessary for the validity of the "introns-early" view. The exon microgene theory (*2*) accommodates the available phylogenetic, consensus sequence, and protein structural data, *and* provides an accounting of how primordial genes were assembled. To our knowledge, this challenge has yet to be met by an "introns-late" theory.

Bonnie L. Bertolaet
Cancer Center,
University of California, San Diego,
La Jolla, CA 92093–0684, USA
H. Martin Seidel
Ligand Pharmaceuticals, Inc.,
9393 Towne Centre Drive, Suite 100,
San Diego, CA 92121, USA
Jeremy R. Knowles
Department of Chemistry and Biochemistry,
Harvard University,
Cambridge, MA 02138, USA

REFERENCES

1. A. Stoltzfus, D. F. Spencer, M. Zuker, J. M. Logsdon, W. F. Doolittle, *Science* **265**, 202 (1994).
2. H. M. Seidel, D. I. Pompliano, J. R. Knowles, *ibid.* **257**, 1489 (1992).
3. M. B. Shapiro and P. Senapathy, *Nucleic Acids Res.* **15**, 7155 (1987).
4. C. S. Craik, S. Sprang, R. Fletterick, W. J. Rutter, *Nature* **299**, 180 (1982).
5. C. C. F. Blake, *ibid.* **273**, 267 (1978).
6. R. Cerff, W. Martin, H. Brinkmann, *ibid.* **369**, 527 (1994).
7. P. Senapathy, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 2133 (1986).

7 September 1994; accepted 19 January 1995

Response: We did not dismiss all possible schemes for an ancient origin of introns (*1*). Rather, we discussed one scheme that anticipates an ancient exon-protein correspondence because it relies critically on exon shuffling and the modularity of split-gene structure to explain the observed modularity of protein structure. This particular scheme has been called the "introns-early view" (*2*) or "exon theory of genes" (*3*). By comparison with this scheme, the suggestions of Senapathy (*4*, *5*) and Seidel *et al.* (*6*) are somewhat limited, thus they were not discussed previously. Bertolaet *et al.* [see also (*7*)] blur the distinctions between these

different theories by using the words "introns-early" to apply to any ancient-introns scheme, though this conflicts with prior usage (*2*). Senapathy and Bertolaet *et al.* raise a number of interesting questions, most of which must be reformulated on the basis of existing knowledge.

1. The challenge of explaining long protein genes—what is it? Consider the gene for PK (pyruvate kinase), a 530-residue metabolic enzyme included in our analysis (*1*). How did this long gene arise? At least part of the answer lies in the structure of the PK protein (*8*). Roughly half of PK is a 260-residue α/β barrel domain of the type seen in triosephosphate-isomerase and many other enzymes (*8*). The arrangement of secondary structures in these α/β barrel domains is of the form $[\beta\alpha]_8$, where a β/α segment is 25 to 30 residues. The remainder of PK comprises three globular domains, the largest of which is the 150-residue COOH-terminal domain, which itself contains a 50-residue βαβαβ nucleotide-binding fold similar to that seen in many dehydrogenases. The largest structural segment of PK that is not seen elsewhere in PK or in another protein is the 100-residue "B" domain [see figure 3 of (*1*)].

Thus, the question posed by the existence of long proteins such as PK is not "How did a 530-codon PK gene arise spontaneously from a random sequence?" but rather, "How did the constituent parts of a PK gene arise, and how were they brought together?" More generally, to explain the many ancient proteins that have iterated structures or shared domains (*9*) without relying heavily on convergence, a scenario for the evolution of proteins must incorporate genetic processes of duplication and fusion.

Blake suggested in 1978 (*10*) that the genetic rearrangements implied by protein structural comparisons might be a result of exon shuffling, if exons corresponded to the appropriate units of protein structure. Thus, in principle, the exon theory of genes relies on exon duplication and exon shuffling to explain repetition and domain-sharing. The "exon-microgene" scheme cannot rely similarly on exon shuffling: Bertolaet *et al.* state that exon-encoded peptides would not exhibit coherent structures, therefore the fusion or duplication of exon-microgenes at the genetic level would not represent assembly of discrete structural parts at the protein level. Senapathy's scheme would appear to rely solely on convergence to explain structural repetition and domain-sharing.

If introns arose recently, then the processes of duplication and fusion necessary for the early evolution of protein genes must have occurred without the participation of introns. We do not understand why this implication is judged to be problematic by Bertolaet *et al.* and others [for example, see (*11*)], given that (i) intronless gene fusions have been a standard tool of bacterial molecular genetics for two decades; and (ii) there is ample comparative evidence that tandem repetition of motifs and domains, as well as fusions of heterologous domains, have played important roles in (intronless) bacterial protein evolution (*12*).

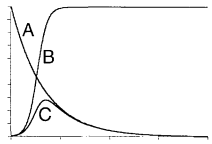

Fig. 1. Some curves useful in describing lengths of exons or ORFs. Only the general shapes of the curves are salient. The horizontal axis represents arbitrary units of length of exons or ORFs; the vertical axis represents arbitrary units of probability density. (Curve A) Exponential function. (Curve B) Sigmoidal function. (Curve C) Combination of the previous two functions. Höglund *et al.* (*15*) have suggested that the actual distribution of exon sizes is best described by the curve C.

2. Does the length distribution of exons suggest an ancient origin? An exponential (or "negative exponential," per Senapathy) frequency distribution, curve A (Fig. 1), is expected for the intervals between random events, whether they are the intervals (in units of time) between radioactive decay events, the exon lengths (in nucleotides) between locations of randomly inserted introns (whether or not there is a target sequence), or the ORF (open reading frame) lengths (in codons) between locations of randomly distributed start and stop codons.

Senapathy (*4, 5*) suggests that the observed distribution of exon lengths is a simple exponential, but it has been known for many years that this there is a substantial deficit of short exon lengths relative to exponential expectations (*13–15*). Höglund and colleagues (*15*) propose that the observed distribution is described by curve C (see Fig. 1), which combines an exponential function (curve A) with a sigmoidal function (curve B) entailing a sharp drop in the probability of short exons. If exons arose originally from random ORFs or microgenes, the curtailing function might represent selection favoring longer peptide modules; in an insertional view of intron origins, the same function might represent selection against introns that insert too close to a pre-existing intron and interfere with splicing. The "longer protein modules" explanation does not account for the fact that the deficit of short exons also applies to 5′ noncoding exons (*14*), which have no peptide products. This latter pattern is consistent with the "intron interference" explanation, and some sort of steric interference phenomenon is further supported by results of manipulative genetic experiments in which short exons are artificially created (*15*).

3. Is there a 600-nucleotide limit on the lengths of exons or ORFs? The "statistical limit" of 600 nucleotides [Bertolaet *et al.*, following Senapathy (*4*), and above] has no apparent empirical or mathematical basis. The reader may refer to Hawkins (*14*) for multiple examples of real exons that are longer than 600 nucleotides. The probability that a randomly generated ORF will have a length of at least **L** codons is $\mathbf{s}^{\mathbf{L}}$, where **s** is the frequency of sense codons. A graph of $\mathbf{s}^{\mathbf{L}}$ as a function of **L** will be like curve A (Fig. 1), a continuous curve that decreases smoothly and monotonically toward zero, but never reaches it (that is, there is no threshold or maximum length limit).

With the use of the above formula, one may verify that there is a nonzero probability of obtaining an ORF of 200 codons in a random sequence of 600 nucleotides or more. For instance, in a 20-kb (kilobase pair) random DNA sequence, one expects 1875 ORFs, of which 0.12 ORFs are expected to have 200 or more codons [counting all six reading frames and defining an ORF according to Senapathy (*2*) as the interval between two nonsense codons]. Thus, on average, one in every eight random 20-kb sequences will contain a 200-codon ORF. Furthermore, it is not necessary to insist upon 200-codon ORFs as a starting condition for the origin of protein genes, since these genes may have originated from shorter sequences, then grown larger through duplications, fusions, and flanking accretions. In a 20-kb random DNA sequence, one expects 170 ORFs of at least 50 codons and 15 ORFs of at least 100 codons.

4. Do split genes carry vestiges of a primordial process of nonsense codon removal? In primordial genomes, long proteins might have been synthesized by transcribing adjacent mini-genes together, then splicing out any nonsense codons between them so as to make one long ORF (*5*). Seidel *et al.* (*6*) and Bertolaet *et al.* suggest that there is evidence for this view because the upstream part of the exonic "shadow sequence" flanking introns (roughly, AG|intron|GU) resembles the nonsense codon UAG. However, it is difficult to view this matching AG dinucleotide as a vestige of an ancient process of nonsense codon removal, given that it would represent a nonsense codon that is not removed by splicing. That is, the sequence AG|intron|GU in the pre-mRNA is spliced to yield AGGU in the mRNA, thus the sequence UAG|intron|GU (representing the junction between two micro-genes in a primor-

dial pre-mRNA) would have been spliced to yield UAGGU, leaving the unwanted nonsense codon in place.

Senapathy (*5*) postulates vestiges of nonsense codons within the AT-rich conserved sequences at the upstream and downstream ends of introns (instead of within the exons). Although this idea is attractive, the problem of phasing has not been addressed. For instance, the putative upstream and downstream nonsense codons identified by Senapathy (*5*) are not in the same phase. On one hand, if UAG, UAA and UGA triplets were spliced out of transcripts regardless of phase, the resulting exons would usually have different reading frames, and their lengths would be (on average) threefold shorter than the lengths of ORFs in the same sequence—this would contradict Senapathy's other major hypothesis (*4*), namely that the size distribution of exons reflects the size distribution of random ORFs. On the other hand, if UAG, UAA, and UGA triplets were spliced out of transcripts only when they occurred in the correct phase at the end of a mini-gene, one wonders what sort of splicing mechanism would have been capable of distinguishing in-phase occurrences of (for example) the triplet UAG from out-of-phase occurrences of the same triplet, such as NUA GNN and NNU AGN.

5. Can the number of intron positions per gene exceed the expectations of the exon-microgene theory? In general, if exons evolved from microgenes (or random ORFs) of an average size of 30 codons (or some other size, such as 20 or 45 codons, which the reader may substitute below), the expected number of intron positions in a gene would be $(L/30) - 1$, where L is the length of the gene in codons. The observed versus the expected numbers of distinct intron positions found in homologous copies of some genes of interest (*16*) are as follows: Globins: 12 vs. 4; superoxide dismutase: 18 vs. 4; small G proteins: 55 vs. 5; glyceraldehyde-3-phosphate dehydrogenase: 47 vs. 10; actin: 34 vs. 11; tubulin: 40 vs. 14. More intron positions will undoubtedly be discovered as additional homologs of these genes are sequenced.

For cases such as these, an ancestral gene containing all known intron positions (the "ancestral intron-rich gene" of Bertolaet *et al.*) would have had exons with an average size less than eight codons, far less than the size of modern exons in animal genes [about 45 codons (*13–15*)], and considerably less than the expected size of random ORFs (about 21 codons). Furthermore, the majority of these putative ancestral exons are difficult to view as microgenes, even as tiny microgenes, since they would not have begun and ended with phase-0 introns (that is, they would not have begun and ended with complete codons).

Thus, there is a substantial excess of intron positions (especially non–phase-0 intron positions) with regard to the expectations of any theory in which exons arise from separate microgenes, minigenes, or ORFs. These expectations do not arise from assuming an exon-protein correspondence (as Bertolaet *et al.* suggest). Advocates of the exon theory of genes have responded to the problem posed by the recent origin of most intron positions by invoking a hypothetical recent process of "sliding" (by which an intron shifts its position a small distance upstream or downstream), but there is, as yet, no evidence for this process.

Though the exon theory of genes would be supported by the discovery of an ancient exon-protein correspondence, the results of our recent analysis did not refute the theory (as Senapathy and Bertolaet *et al.* imply) but only served to emphasize the long-standing lack of reliable evidence of this type [see (*17*) and others (*18*)]. The exon theory of genes is not in great danger of being refuted: some versions are becoming unfalsifiable due to an ever-increasing reliance on events on intron "sliding" and episodes of "streamlining" (genome-wide loss of introns) to accommodate data that would otherwise be considered contradictory; other versions of the theory concede that recent insertions explain observed patterns in the data, but still maintain superfluous propositions about the antiquity of introns [that is, ancient introns existed but were lost, then new introns were added later by insertion (*19*)].

The restricted distribution of spliceosomal introns (which elicits ad hoc proposals of "streamlining") and the large numbers of different intron positions found in extant genes (which elicits ad hoc proposals of "sliding") are problematic for any scheme in which exons arise from primordial minigenes, including the scheme proposed by Senapathy (*4*, *5*) and the "exon microgene" scenario (*6*). Perhaps these ideas can be developed further by abandoning claims based on the "200-codon statistical limit," by addressing the problems of phasing inherent in claims about vestiges of nonsense codons, and by incorporating mechanisms that allow for (i) the patterns of repetition and domain-sharing observed in ancient proteins; (ii) the nonexponential distribution of exon lengths; (iii) the recent origin of most intron positions; and (iv) the paucity or absence of spliceosomal introns in multiple outgroups to the intron-rich eukaryotes.

Arlin Stoltzfus
David F. Spencer
Canadian Institute for Advanced Research (CIAR), Program in Evolutionary Biology, Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia, B3H 4H7 Canada
Michael Zuker
CIAR Program in Evolutionary Biology, Institute for Biomedical Computing, Box 8036, Washington University, St. Louis, MO 63110, USA
John M. Logsdon Jr.
Department of Biology, Indiana University, Bloomington, IN 47405, USA
W. Ford Doolittle
Canadian Institute for Advanced Research (CIAR), Program in Evolutionary Biology, Department of Biochemistry, Dalhousie University

REFERENCES AND NOTES

1. A. Stoltzfus, D. F. Spencer, M. Zuker, J. M. Logsdon, W. F. Doolittle, *Science* **265**, 202 (1994).
2. W. F. Doolittle, *Am. Nat.* **130**, 915 (1987).
3. W. Gilbert, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 901 (1987).
4. P. Senapathy, *Proc. Natl. Acad. Sci.* **83**, 2133 (1986).
5. ———, *ibid.* **85**, 1129 (1988).
6. H. M. Seidel, D. L. Pompliano, J. R. Knowles, *Science* **257**, 1489 (1991).
7. E. N. Trifonov, *J. Mol. Evol.* **40**, 337 (1995).
8. H. Muirhead *et al. EMBO J.* **5**, 475 (1986).
9. A. D. MacLachlan, *Nature* **285**, 267 (1980); J. S. Richardson, *Adv. Protein Chem.* **34**, 167 (1981).
10. C. C. F. Blake, *Nature* **273**, 267 (1978).
11. R. Cerff, W. Martin, H. Brinkmann, *ibid.* **369**, 527 (1994).
12. P. Bork and R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8990 (1992); G. M. Pao and M. H. Saier, *J. Mol. Evol.* **40**, 136 (1995).
13. M. W. Smith, *J. Mol. Evol.* **27**, 45 (1988).
14. J. D. Hawkins, *Nucleic Acids Res.* **16(21)**, 9893 (1988).
15. M. Höglund, T. Säll, D. Röhme, *J. Mol. Evol.* **30**, 104 (1990).
16. References for numbers of intron positions are as follows: globins: Stoltzfus *et al.* (*1*); superoxide dismutase: J.M.L., unpublished compilation from database sequences; glyceraldehyde-3-phosphate dehydrogenase [R. Kersanach *et al.*, *Nature* **367**, 387 (1994)]; small G proteins [W. Dietmaier and S. Fabry, *Curr. Genet.* **26**, 497 (1994)]; actin: K. Weber and W. Kabsch, [*EMBO J.* **13**(6), 1280 (1994)]; tubulin: M-F. Liaud, H. Brinkmann, R. Cerff, *Plant Molec. Biol.* **18**, 639 (1992).
17. J. Rogers, *FEBS Lett.* **268**(2), 339 (1990).
18. N. J. Dibb, *ibid.* **325**(1), 135 (1993); W. F. Doolittle and A. Stoltzfus, *Nature* **361**, 403 (1993); S. A. Benner *et al.*, in *The RNA World*, R. F. Gesteland and J. F. Atkins, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1993), pp. 27–70; K. Weber and W. Kabsch, *EMBO J.* **13**(6), 1280 (1994).
19. L. Hurst, *Nature* **371**, 381 (1994); J. S. Mattick, *Curr. Opin. Genet. Dev.* **4**, 823 (1994).

27 September 1994; accepted 19 January 1995