Testing the Exon Theory of Genes: The Evidence from Protein Structure

Arlin Stoltzfus,* David F. Spencer, Michael Zuker, John M. Logsdon Jr., W. Ford Doolittle

A tendency for exons to correspond to discrete units of protein structure in protein-coding genes of ancient origin would provide clear evidence in favor of the exon theory of genes, which proposes that split genes arose not by insertion of introns into unsplit genes, but from combinations of primordial mini-genes (exons) separated by spacers (introns). Although putative examples of such correspondence have strongly influenced previous debate on the origin of introns, a general correspondence has not been rigorously proved. Objective methods for detecting correspondences were developed and applied to four examples that have been cited previously as evidence of the exon theory of genes. No significant correspondence between exons and units of protein structure was detected, suggesting that the putative correspondence does not exist and that the exon theory of genes is untenable.

Spliceosomal introns are present in the nuclear protein-coding genes of most eukaryotic organisms, but they have not been detected in several eukaryotic protist phyla or in eubacteria, archaebacteria, and organelles (*1*). Two major theories have emerged in the continuing debate on the origin of these introns. The exon theory of genes—sometimes called the introns-early view—proposes that (i) exons are the descendants of ancient mini-genes and introns are the descendants of the spacers between them; (ii) genes large enough to encode contemporary proteins were first assembled from sets of exons; (iii) the machinery of splicing originated in an ancient RNA world; and (iv) introns were lost completely from both kingdoms of bacteria as well as several protist groups (*2–4*). In contrast, the insertional theory of intron origins—also known as the introns-late view—holds that (i) split genes arise from uninterrupted genes by insertion of introns; (ii) genes large enough to encode contemporary proteins first arose (presumably from smaller genes) without the participation of introns; (iii) the machinery of spliceosomal splicing arose from fragmented self-splicing introns; and (iv) spliceosomal introns were never present in the ancestors of those organisms that now lack them (*1, 3, 5, 6*).

A. Stoltzfus, D. F. Spencer, and W. F. Doolittle are with the Canadian Institute for Advanced Research (CIAR) Program in Evolutionary Biology, Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia, B3H 4H7 Canada. M. Zuker is with the CIAR Program in Evolutionary Biology, Institute for Biomedical Computing, Washington University, St. Louis, MO 63110, USA. J. M. Logsdon Jr. is with the Department of Biology, Indiana University, Bloomington, IN 47405, USA.

*To whom correspondence should be addressed.

Testable implications of the exon theory of genes arise from the possibility that exons might retain some of the properties expected of mini-genes. As first suggested by Blake (*7*), this theory implies that exons should encode discrete units of folded protein structure, suitable for combinatorial assembly. Originally, a correspondence of exons with globular domains was anticipated (*7*). However, it soon became clear that exons are generally too short for such a correspondence, suggesting that globular domains might be encoded by blocks of exons, with individual exons encoding smaller elements (*8*). Subsequent reports have implicated familiar elements such as secondary structures (*9–11*) and motifs (*8, 12*), and have also introduced novel divisions of protein structure, including "least-extended" units (*9, 13, 14*), peptides "circumscribed by a sphere 28 Å in diameter" (*4*), and "compact modules" (*15*).

Not all genes and proteins are relevant to the search for this possible correspondence. Chimeric genes assembled by exon shuffling in recent evolutionary times may exhibit correspondences that are not ancient, where "ancient" refers (here and below) to the period of time prior to the divergence of archaebacteria, eubacteria, and eukaryotes. Instead, the process of exon shuffling may have favored the propagation of chance correspondences between split gene structure and protein structure subsequent to an insertional origin of introns in an early eukaryote (*3, 5, 6*). Therefore, evidence for an ancient origin of introns must be sought in genes and proteins of ancient origin, not recently assembled ones. This important point, suggested in 1983 by Blake (*8*), has been accepted by advocates of both theories of intron origins (*3–6, 9–12*).

Nevertheless, evidence for exon shuffling in the assembly of a few recently evolved genes continues to be confounded with evidence for the assembly of all genes from exons. Dorit *et al.* (*16*) made exhaustive comparisons of exon sequences in a search for chimeric genes suggestive of exon shuffling. The putative chimeras identified in this search were genes for animal-specific proteins (*17*)—a result fully consistent with the insertional theory of intron origins—yet these few examples were used to infer the size of a putative underlying universe of exons (*16*) as though they represented all cases, an assumption that begged the question of whether a primordial universe of exons ever existed (*17*).

Such confusion can be avoided by maintaining a clear methodological distinction between ancient genes and more recently assembled genes, in parallel with a logical distinction between the exon theory of genes and the exon-shuffling hypothesis. The exon-shuffling hypothesis maintains that introns sometimes serve as sites of illegitimate recombination, providing a direct route for the assembly of chimeric genes from sets of exons (*18*). This hypothesis says nothing about the origin of introns, but instead addresses the expected effects of the presence of introns on the formation of new genes (*2, 3*). The exon-shuffling hypothesis has been adequately confirmed by the tendency for intron positions to correspond to the recombinant junctions of some chimeric sequences, none of which is ancient (*17, 19*).

In contrast, the introns-early view, proposed by Darnell (*2*) and Doolittle (*2, 3*) and aptly named the exon theory of genes by Gilbert (*4*), addresses the ultimate origin of the exons and introns of split protein-coding genes. Although this theory is sometimes presented in textbooks as an established fact (*20*), the supporting evidence consists of (i) examples of shared intron positions that are interpreted to be ancient (*21*) and (ii) a putative ancient correspondence between exons and units of protein structure.

The case for this ancient correspondence rests on a small set of proposals regarding globins (*13*), lysozyme (*14*), several nicotinamide adenine dinucleotide (NAD)–dependent dehydrogenases (*12*), pyruvate kinase (PK) (*10, 11*), and triose phosphate isomerase (TPI) (*9–11*). A three-intron proposal regarding globins (*13*) was hailed as a success while three intron positions were known (*3, 4*), but conflicted with subsequently discovered introns (*22*) (see below). Predictions of intron positions on the basis of correspondences proposed for lysozyme genes have

not fared well (*23*). Most of the 11 intron positions known in genes for glyceraldehyde-3-phosphate dehydrogenase (an NAD-dependent dehydrogenase) in 1985 were justified relative to its structure by Stone *et al.* (*12*), but 36 additional intron positions have subsequently been identified (*21*). A comparison of several NAD-dependent dehydrogenases does not reveal the expected pattern of shared introns in regions encoding their shared NAD-binding motifs (*24*). Gilbert and Glynias (*10*) predicted an unknown ancestral intron position for PK on the basis of intron positions detected in animal genes, but none of the seven additional intron positions from plant and fungal genes—already known at the time the prediction was made—is within 48 codons of the predicted site (see below). Taken together, reports regarding TPI suggest the unlikely circumstance that a single split gene structure simultaneously corresponds to (i) the eightfold repeats of α helices and β strands that constitute its β-barrel domain (9), (ii) 13 compact modules (*15*), and (iii) 12 "least extended" or compactly folded peptides (*9, 10, 25*).

Exons do not neatly correspond with any single aspect of protein structure, and proposing a special type of correspondence for each individual protein would be without value. The question that remains is whether there is a general tendency for exons to encode some discrete unit of protein structure, even if the underlying signal is partially obscured by noise. Accordingly, quantitative methods for evaluating the statistical significance of correspondences have been developed and applied uniformly to four representative ancient proteins—alcohol dehydrogenase (ADH) (an NAD-dependent dehydrogenase), globins, PK, and TPI—each of which has been cited as providing evidence for the exon theory of genes.

Analysis of Observed Gene Structures

For each of the four examples studied, the amino acid sequences inferred from intron-containing copies of the gene have been aligned with each other (*26, 27*) and with a homologous reference protein whose crystal structure has been determined (*28*). The resulting complete set of known intron positions was then mapped in relation to structural features of the corresponding protein (Figs. 1 to 4). The data represented in the figures, along with Cα coordinates from the corresponding crystal structures (*28*), serve as the basis for our analyses.

If exons in ancient genes encode defined units of protein structure, such as globular domains (*7*), modules (*15*), and secondary structural elements (*9–11*), then the positions of introns should tend to correspond to boundary regions between structural units, or at least to fall close to their ends. Such a correspondence can be scored by measuring the distance from an intron to the nearest boundary region, with lower scores indicating greater correspondence.

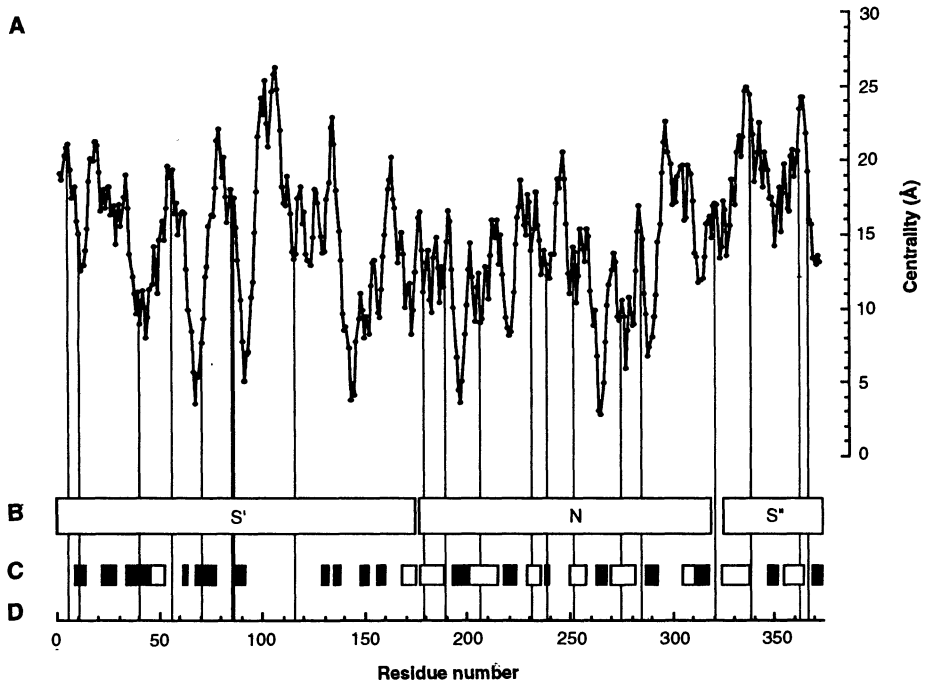

Fig. 1. Intron positions of class I ADH genes in relation to structural features of the reference protein, the 374-residue enzyme from horse liver (*28*). (**A**) Centrality plot, showing the distance of each Cα atom from the center of mass of the domain in which the atom resides (the average distance for all residues is 14.6 Å). (**B**) Like other NAD-dependent dehydrogenases, ADH has two domains: one for substrate binding (composed of two noncontiguous segments, S′ and S″), and one for nucleotide binding (N). Domains are delimited according to the encompassed secondary structures, so that domain N is defined as residues 176 to 318 inclusive (*37*). (**C**) Locations of α helices (open boxes) and β strands (filled boxes). (**D**) The 20 known intron positions, indicated by vertical lines, are completely represented in four genomic sequences (*26*) as follows: human, 6–0, 40–0, 86–1, 115–2, 189–0, 276–0, 321–1, 367–2; rat, 6–0, 40–0, 86–2, 115–2, 189–0, 276–0, 321–1, 367–2; maize, 11–1, 57–0, 71–2, 179–1, 207–0, 232–1, 253–0, 285–0, 339–0; and *Aspergillus*, 239–0, 363–1. Positions are given in the codon-phase notation of Dibb and Newman (*36*), based on an alignment with the reference protein.

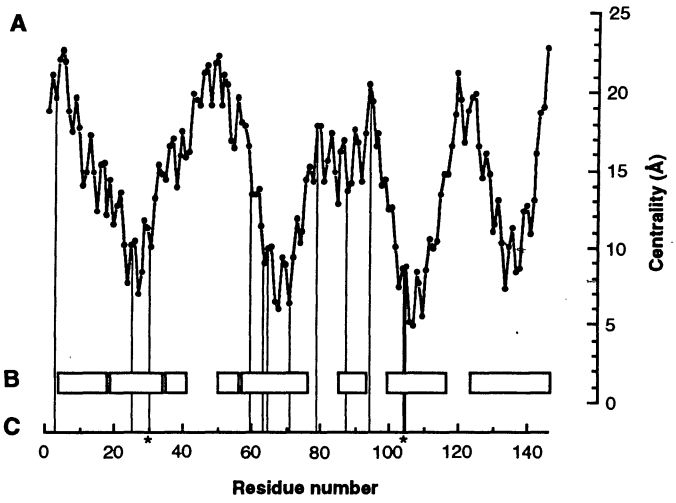

Fig. 2. Intron positions of globin-related genes in relation to structural features of the 146-residue β subunit of horse hemoglobin (*28*). A characteristic globin fold consisting of eight (or, sometimes, seven) α helices is common to the globular domains of hemoglobins, myoglobins, leghemoglobins, various invertebrate globins, and eubacterial hemeproteins. (**A**) Centrality plot (average centrality, 14.4 Å). (**B**) Locations of α helices. (**C**) The 12 known intron positions, which are represented by a minimal set of seven genomic sequences (*27*) as follows: codworm (*Pseudoterranova decipiens*), 3–1, 30–2, 64–1, 105–0; soybean, 30–2, 71–0, 105–0; *Chlamydomonas eugametos*, 25–0, 79–0, 94–0; *Artemia salina*, 30–2, 104–2; *Caenorhabditis elegans*, 59–2; *Chironomus thummi*, 65–1; and *Paramecium caudatum*, 87–0. Asterisks, the two intron positions commonly present in vertebrate globin genes [the ones on which Gō's (*13*) hypothesis was based]. Other conventions are as indicated for Fig. 1.

For example, for the 14 distinct intron positions identified in TPI genes, the average distance to the nearest boundary between secondary structures (α helices or β strands) is 5.9 base pairs (bp), closer than the random expectation of 6.5 bp. Thus, TPI introns exhibit a measurable tendency to fall between or at the edges of secondary structural elements.

Whether or not this tendency is statistically significant can be addressed by generating a reference distribution of scores measured from sets of introns that are positioned randomly with respect to protein structure. Of 1000 reference sets of introns, each consisting of 14 nonidentical positions randomly distributed over the length of a hypothetical TPI gene (29), 379 sets showed an equal or greater tendency to avoid secondary structural elements. That is, if introns are randomly positioned with regard to protein structure, the probability that a set of introns would avoid secondary structures as well as the observed set is 379/1000, or 0.38, which is not significant (Table 1). For ADH, globins, and PK, there is no significant tendency for intron positions to avoid secondary structures (Table 1).

Because ADH (Fig. 1) and PK (Fig. 3) possess multiple globular domains, similar methods can be applied to test whether there is an excess of introns at or near domain boundaries. Although nearly all intron positions interrupt domains, there is a small (insignificant) excess of introns near domain boundaries for both ADH (P = 0.24) and PK (P = 0.25).

The correspondence of introns to boundaries between other structural elements can be addressed in a similar manner. Gō and Nosaka (*15*) have proposed that TPI has 13 modules, with boundaries as shown in Fig. 4. The average distance to an intermodule boundary region for the observed set of TPI introns is 14.9 bp, slightly greater than the average of 14.8 bp for random sets (based on 1000 simulations). Thus, no significant correspondence exists between exons and these divisions of TPI.

Gō has proposed for globins (*13*) and lysozyme (*14*) that residues associated with introns (residues encoded by codons that abut or contain an intron) tend to be located near the center of globular domains; a similar claim has been made for TPI (*9, 11*). The centrality of a residue can be defined as its physical distance from the center of the globular domain in which it resides. The centrality of intron positions for ADH, globins, PK, and TPI can be judged visually by examining Figs. 1 to 4, which include a plot of the centrality of each residue in the protein.

The average centrality for globin intron positions is 12.9 Å, which is lower, but not significantly, than the reference mean of 14.4 Å (P = 0.12) (Table 2); much of the excess centrality is attributable to the two intron positions (Fig. 2) that stimulated Gō's original conjecture (*13*). For ADH and PK, intron positions show no tendency to map to central locations in a domain, whereas intron positions for TPI show a slight but insignificant tendency (P = 0.42) (Table 2).

Gō also proposed that ancestral globin exons encoded least extended polypeptide structures (*13*), and similar claims have been made for lysozyme (*14*), TPI (*9–11*), and PK (*10, 11*). The extensity of a polypeptide fragment corresponding to an exon, here called an exon-encoded peptide, can be quantified by a measure of the dispersion of its Cα atoms in three-dimensional space, with lower extensity scores indicating a greater degree of compactness. We have used various metrics for extensity of exon-encoded peptides, including the maximum distance between any pair of residues (that is, the diameter) and the radius of gyration

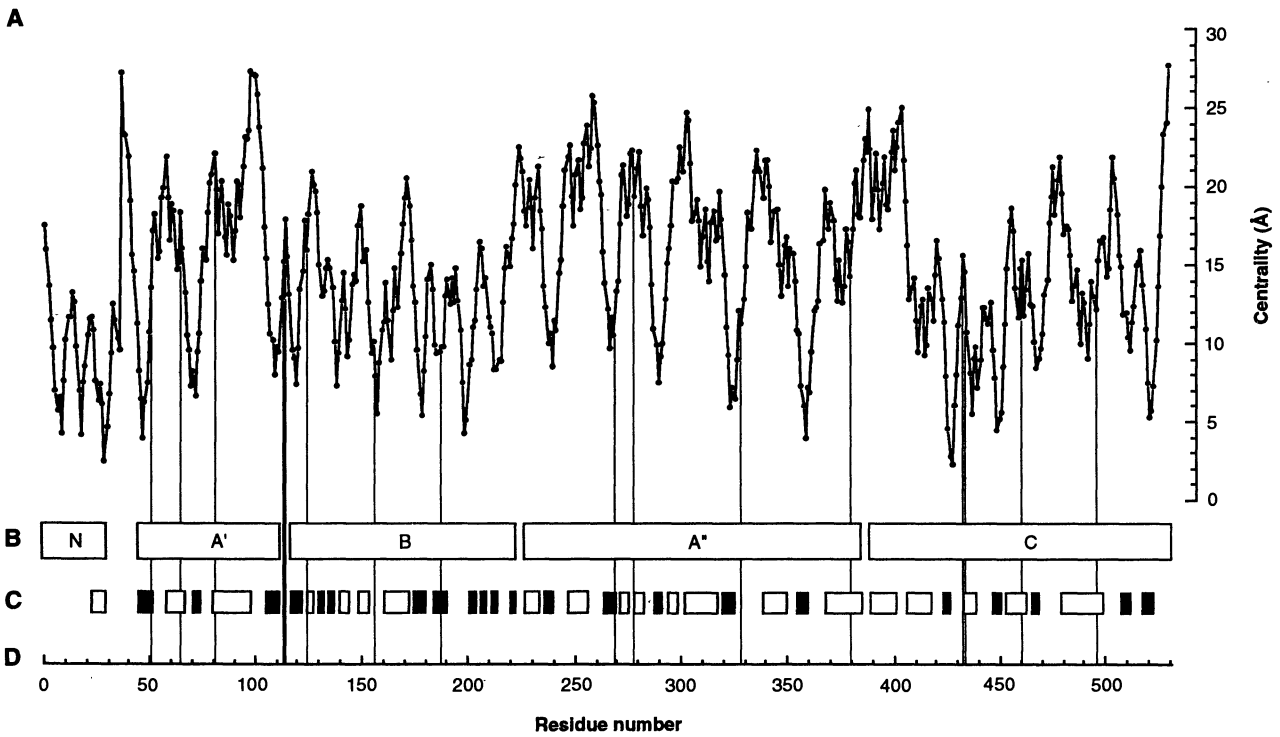

Fig. 3. Intron positions of PK genes in relation to structural features of the 530-residue feline muscle enzyme (*28*). (**A**) Centrality plot (average centrality, 14.6 Å). (**B**) The four domains of PK, with boundaries defined by secondary structural elements assigned to domains by Muirhead *et al.* (*28*). Domain A (A′ and A″), an eight-stranded β barrel (similar in structure and overall dimensions to TPI), is interrupted by domain B and is flanked by amino- and carboxyl-terminal domains (N and C, respectively). (**C**) Elements of secondary structure. (**D**) The 16 known intron positions are represented in a minimal set of three genomic sequences (*26*) as follows: chicken, 51–1, 82–0, 126–0, 188–1, 278–2, 329–0, 380–0, 435–2, 496–1; *Aspergillus*, 51–1, 66–1, 82–0, 114–0, 156–2, 433–2, 461–0; and potato, 115–0, 270–0. Other conventions are as indicated for Fig. 1. Gilbert and Glynias (*10*) recently predicted that an intron would be found in a PK gene near codon 222, on the basis of the nine positions from the chicken gene. However, none of the seven additional positions known from *Aspergillus* and potato is closer than 48 codons to this site.

(the root-mean-square deviation from the center of mass). The diameter metric is suggested by Gilbert's proposal, based on TPI (*4*), that primordial exon-encoded peptides would be compactly folded peptides "circumscribed by a sphere 28 Å in diameter." Indeed, the average diameter of exon-encoded peptides for the set of 12 inferred ancestral (*30*) TPI exons is 26.7 Å. However, the average diameter for reference (*31*) sets of exons is 26.8 Å, not significantly different ($P = 0.48$) (Table 3). In fact, no significant correlations are apparent for any of the four examples when extensity is quantified by either the diameter or the radius of gyration (Table 3). Exon-encoded peptides are only as compact as one might expect by chance.

Gilbert and Glynias (*10*) described results of the first statistical analysis of the extensity of exon-encoded peptides while the present study was under review. They quantified extensity by tallying (for each exon-encoded peptide) the total number of pairwise Cα–Cα distances greater than an arbitrarily chosen value of 28 Å, comparing the observed set of exon sizes to randomly permuted sets. Gilbert and Glynias (*10*) report a probability of 0.039 for TPI with these methods. With the same metric and reference model, we obtained a similar result, $P = 0.057$ (95% confidence interval of ± 0.013); the slight discrepancy between these two values is perhaps attributable to a difference in accounting of intron positions.

The ultimate evolutionary importance of this (or any other) metric depends on its power to reveal a correlation that applies to ancient proteins in general. Accordingly, tests with this metric have been performed on three other proteins, revealing no significant correlation for ADH ($P = 0.49$), globins ($P = 0.70$), or PK ($P = 0.48$) (all tests based on 1000 simulations). Thus, there is no evidence that the relatively low P value for TPI represents a general significant tendency for exons to encode compact peptides.

Table 1. Correspondence of intron positions with secondary structural elements. The distance score for a set of introns is the average of the individual scores, where the individual score is the distance (base pairs) to the nearest interelement region (0 bp if the intron is within an interelement region). Mean reference scores (Ref.) are based on 1000 reference sets of N randomly generated intron positions, where N is the number of observed intron positions (the SD of the reference mean is also given). The P value represents the probability that a set of introns generated by the random reference model (*29*) would avoid secondary structure as well or better than the observed set.

Gene	Introns	Distance score (bp)		SD	P
		Observed	Ref.		
ADH	20	4.35	3.87	1.09	0.70
Globins	12	12.1	10.2	2.72	0.76
PK	16	5.06	4.86	1.58	0.59
TPI	14	5.86	6.54	1.80	0.38

Table 2. Tendency of introns to map to central regions of globular domains. The centrality score for a set of introns is the average of individual scores, where the individual score is the distance (angstroms) from the residue associated with the intron (the residue encoded by a codon that contains the intron, or that is bounded on its 5′ end by the intron) to the center of mass of the domain (estimated as the average position of all Cα atoms). Mean reference scores (Ref.) are based on 1000 reference sets of N randomly generated intron positions, where N is the number of observed intron positions (the SD of the reference mean is also given). The P value represents the probability that a set of introns generated by the random reference model (*29*) would be closer to the centers of globular domains than the observed set.

Gene	Introns	Centrality score (Å)		SD	P
		Observed	Ref.		
ADH	20	14.8	14.5	1.06	0.58
Globins	12	12.9	14.4	1.28	0.12
PK	16	14.7	14.6	1.25	0.53
TPI	14	16.1	16.3	1.26	0.42

Analysis of Hypothetical Optimized Gene Structures

Our analyses of ADH, globins, PK, and TPI reveal no significant signs of modular assembly from exon-encoded units. These negative results stimulated us to evaluate our methods with idealized genes: hypothetical genes intentionally designed to have properties expected from the exon theory of genes. Algorithms have been developed for finding an optimized N-exon gene structure, where N is a whole number and the quantity to be minimized is a measure of the extensity of exon-encoded peptides (*32*). Hypothetical PK and TPI genes that were optimized by the size-weighted average radius of gyration of exon-encoded peptides (that is, weighted by the length of each exon) are highly significant when scored by the radius of gyration, and they are also highly significant when scored by the diameter ($P \leq 0.001$, on the basis of tests of optimized PK genes for N values of 10, 15, 20, 25, and 30, and optimized TPI genes for N values 10 and 15) (*33*).

Introns in these idealized genes are clearly nonrandom with regard to other structural elements. The optimized five-exon gene for PK has introns abutting codons 42, 115, 224, and 386, positions that correspond closely ($P = 0.001$) to the four interdomain boundaries shown in Fig. 3, confirming that these classical globular domains, identified by Muirhead *et al.* (*28*),

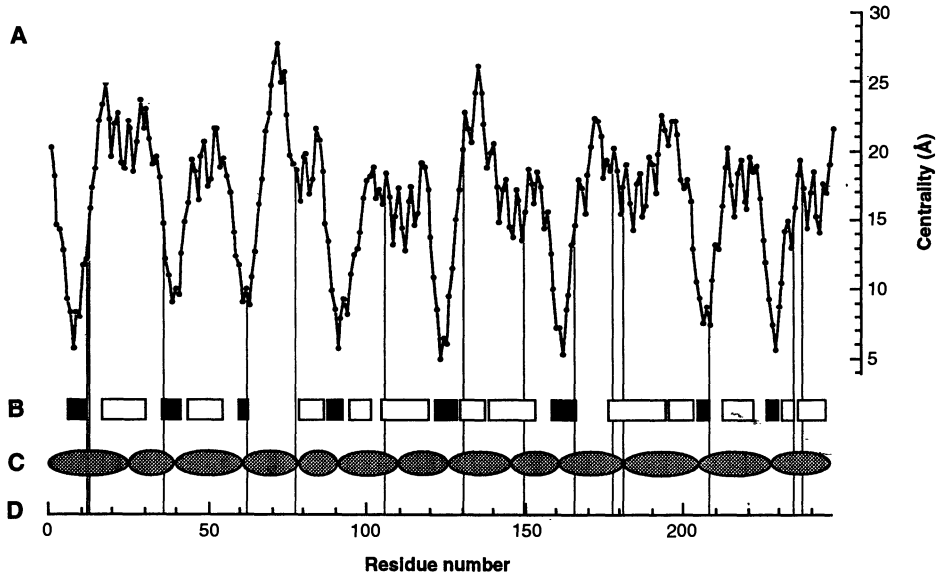

Fig. 4. Intron positions of TPI genes in relation to structural features of the 247-residue chicken muscle enzyme (*28*). (**A**) The centrality plot (average centrality, 16.2 Å) reveals the regularity of the β-barrel domain; the eight troughs represent the eight β strands that pass near the center, whereas the zigzagging segments between the troughs show the course of the peptide backbone as it winds through the peripheral α helices [see also domain A of PK (Fig. 3)]. (**B**) Elements of secondary structure. (**C**) Modules proposed by Gō and Nosaka (*15*). The 14 known intron positions (**D**) are represented in four genomic sequences (*27*) as follows: chicken, 37–1, 78–2, 107–0, 151–1, 180–0, 209–1; maize, 14–0, 37–1, 78–2, 107–0, 151–1, 183–0, 209–1, 237–0; *Aspergillus*, 13–2, 107–0, 132–0, 167–2, 239–1; mosquito, 64–0. Other conventions are as indicated for Fig. 1.

are compactly folded units (P value based on 1000 reference sets of four introns each). Furthermore, introns in hypothetical optimized PK and TPI genes consistently show significant ($P < 0.05$) centrality scores, on the basis of tests of 25- and 30-exon PK optimizations, and 10- and 15-exon TPI optimizations (*33*). These results validate the logic behind Gō's (*13*) original suggestion that, if exons encode compact units, introns will tend to divide extended segments that pass near the center of a domain. Finally, optimized gene structures consistently show a weaker inverse correlation with regard to secondary structure: P values ranged from 0.90 to 0.99 in tests of the tendency for introns to fall between secondary structural elements, indicating that the introns actually tend to interrupt secondary structures (*33*). This result suggests that exons are unlikely to encode both compact modules and discrete pieces of secondary structure simultaneously, as has been proposed for TPI (*9–11*).

Evolutionary Implications

A quantitative analysis of 62 intron positions in genes for four ancient proteins has revealed no evidence of a significant tendency for introns to avoid interrupting secondary structures, modules, or globular domains. Likewise, no set of introns exhibits a significant tendency to be centrally located in a globular domain. No significant result was obtained with two nonarbitrary measures of the extensity of exon-encoded peptides. A third metric of extensity, developed by Gilbert and Glynias (*10*) for TPI, yields a P value in the range of 0.05 when applied to TPI but fails to reveal a significant correspondence in the three other proteins. In short, objective methods applied uniformly to four examples provide no significant evidence that ancient proteins were assembled from exon-encoded modules of structure.

The importance of these observations can be clarified as follows. First, our results pertain to ancient proteins and have no bearing on the possibility of correspondences in protein-coding genes assembled recently by exon shuffling (*3, 6, 19*). Second, the absence of evidence for several specific types of correspondences does not imply that introns are positioned entirely randomly with regard to protein structure. On the contrary, a general nonrandom relation between intron positions and protein structure is a necessary consequence of the facts that (i) introns are nonrandomly positioned with regard to flanking exonic nucleotide sequences (*34*); (ii) exonic nucleotide sequences are correlated (as described by the genetic code) with amino acids; and (iii) amino acids are nonrandomly distributed with regard to protein structures. This same line of reasoning explains why a nonrandom relation of split gene structure to protein structure is expected if introns are mobile elements that either insert preferentially at a target site, or insert randomly but are selectively retained only in specific sequence contexts. Correlations that arise by such routes may not follow any of the patterns sought here. Indeed, the significant tendency for residues associated with introns to map to surface-accessible regions of proteins (*35*) is not a logical implication of Blake's (*7*) general conjecture that exons should encode discrete structural units, but is perhaps attributable to an association of introns with a few codons for hydrophilic amino acids.

Finally, the negative results presented here suggest a reappraisal of the status of the exon theory of genes. The growing mass of data on the phylogenetic distribution of intron-containing genomes suggests that spliceosomal introns arose and spread in eukaryotic nuclear genomes (*1*). Furthermore, the number of distinct intron positions known for a gene usually exceeds the expectations of the exon theory of genes as soon as homologs of the gene are sequenced from diverse eukaryotic phyla, as has happened in the case of globin (Fig. 2), glyceraldehyde-3-phosphate dehydrogenase (*21*), and other genes. The restricted phylogenetic distribution of most individual intron positions also argues for addition of introns (*36*). Proposals to the effect that the spliceosomal introns in eukaryotic nuclear genes of bacterial origin are ancient (*21*), beg the question of why the bacterial genes that have supposedly retained their ancient spliceosomal introns are inevitably found in the nucleus, not in organelles or in bacteria—exactly the pattern expected if introns are mobile elements that have propagated only in the eukaryotic nucleus.

Thus, several lines of evidence favor a late insertional origin of introns, whereas the only empirical evidence specifically favoring the exon theory of genes has been the proposal that exons encode meaningful units of protein structure. Given that such a correlation does not appear to exist, we suggest that the exon theory of genes is untenable.

Table 3. Correspondences based on extensity of exon-encoded peptides for inferred ancestral exons. The score for a set of exons is the average of the individual exon scores, where the diameter is the maximum of all Cα–Cα distances (angstroms) in the exon-encoded peptide, and the radius of gyration is the root-mean-square distance (angstroms) of Cα atoms from the center of mass of the exon-encoded peptide. The reference scores (Ref.) are based on 1000 sets of N randomly generated exons, where N is the number of exons in the inferred ancestral set. The P value represents the probability that a set of exons generated by the random reference model (*31*) would encode peptides as compact or more compact than the ancestral set inferred (*30*) from observed intron positions.

Gene	Inferred ancestral exons	Extensity rule	Extensity score (Å)		SD	P
			Inferred ancestral	Ref.		
ADH	20	Diameter	22.1	22.3	1.19	0.43
		Radius of gyration	8.11	8.05	0.389	0.54
Globins	11	Diameter	15.7	16.5	1.19	0.26
		Radius of gyration	5.61	5.92	0.337	0.18
PK	15	Diameter	32.6	33.1	1.44	0.34
		Radius of gyration	11.1	11.2	0.480	0.37
TPI	12	Diameter	26.7	26.8	1.35	0.48
		Radius of gyration	9.23	9.25	0.432	0.49

REFERENCES AND NOTES

1. J. Palmer and J. Logsdon, *Curr. Opin. Genet. Dev.* **1**, 470 (1991).
2. J. E. Darnell, *Science* **202**, 1257 (1978); W. F. Doolittle, *Nature* **272**, 581 (1978).
3. W. F. Doolittle, *Am. Nat.* **130**, 915 (1987).
4. W. Gilbert, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 901 (1987).
5. T. Cavalier-Smith, *Trends Genet.* **7**, 145 (1991).
6. D. A. Hickey, B. F. Benkel, S. M. Abukashawa, *J. Theoret. Biol.* **137**, 41 (1989).
7. C. C. F. Blake, *Nature* **273**, 267 (1978).
8. ———, *ibid.* **306**, 535 (1983).
9. D. Straus and W. Gilbert, *Mol. Cell. Biol.* **5**, 3497 (1985); W. Gilbert, M. Marchionni, G. McKnight, *Cell* **46**, 151 (1986).
10. W. Gilbert and M. Glynias, *Gene* **135**, 137 (1994).
11. N. Lonberg and W. Gilbert, *Cell* **40**, 81 (1985).
12. E. M. Stone, K. N. Rothblum, R. J. Schwartz, *Nature* **313**, 498 (1985); G. Duester, H. Jornvall, G. W. Hatfield, *Nucleic Acids Res.* **14**, 1931 (1986); A. M. Michelson, C. C. F. Blake, S. T. Evans, S. H. Orkin, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 6965 (1985).
13. M. Gō, *Nature* **291**, 90 (1981).
14. ———, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1964 (1983).
15. Module boundaries are defined as the positional averages of profile extrema listed in table 1 of M. Gō and M. Nosaka [*Cold Spring Harbor Symp. Quant. Biol.* **52**, 915 (1987)].
16. R. L. Dorit, L. Schoenbach, W. Gilbert, *Science* **250**, 1377 (1990).
17. R. Doolittle, *ibid.* **253**, 677 (1991); L. Patthy, *Bioessays* **13**, 187 (1991).
18. W. Gilbert, *Nature* **271**, 501 (1978).
19. R. F. Doolittle, *Trends Biochem. Sci.* **10**, 233 (1985); L. Patthy, *Curr. Opin. Struct. Biol.* **1**, 351 (1991).
20. J. D. Watson, N. H. Hopkins, J. W. Roberts, J. A. Steitz, A. Weiner, *Molecular Biology of the Gene* (Benjamin/Cummings, Menlo Park, CA, 1987), pp. 1145–1146.

21. R. Kersanach *et al.*, *Nature* **367**, 387 (1994).
22. A. Stoltzfus and W. F. Doolittle, *Curr. Biol.* **3**, 215 (1993).
23. No new intron positions for hen-type lysozyme genes have appeared in GenBank, leaving Gō's (*14*) prediction uncomfirmed; intron positions identified in goose-type lysozyme genes conflict with predictions extrapolated from Gō's model for hen-type lysozyme [T. Nakano and T. Graf, *Biochim. Biophys. Acta* **1090**, 273 (1991)].
24. J. Burke *et al.*, *Proteins: Structure, Function and Genetics* **2**, 177 (1987).
25. C. Tittiger, S. Whyard, V. K. Walker, *Nature* **361**, 470 (1993); W. F. Doolittle and A. Stoltzfus, *ibid.*, p. 403.
26. Intron positions for ADH and PK were gathered from all known complete eukaryotic genomic sequences (26 ADH and 13 PK sequences). The minimal set of sequences representing all known distinct intron positions for ADH consists of the genes from human [G. Duester, M. Smith, V. Bilanchone, G. W. Hatfield, *J. Biol. Chem.* **261**, 2027 (1986)], rat [D. W. Crabb *et al.*, *Genomics* **5**, 906 (1989)], *Aspergillus* [D. I. Gwynne *et al.*, *Gene* **51**, 205 (1987)], and maize [E. S. Dennis *et al.*, *Nucleic Acids Res.* **12**, 3983 (1984)]. The minimal set for PK consists of the genes from chicken (*16*), *Aspergillus* [L. de Graff and J. Visser, *Curr. Genet.* **14**, 553 (1988)], and potato [K. P. Cole, S. D. Blakeley, D. T. Dennis, *Gene* **122**, 255 (1992)]. Inferred amino acid sequences were aligned with each other and the reference sequence with the use of Clustal V [D. G. Higgins and P. M. Sharp, *ibid.* **73**, 237 (1988); *Comp. Appl. Biosci.* **5**, 151 (1989)] with default settings.
27. The current complete set of TPI intron positions is that used by Tittiger *et al.* (*25*). Sources of globin intron positions can be found in B. Pohajdak and B. Dixon [*FEBS Lett.* **320**, 281 (1993)] and L. Moens *et al.* (*ibid.*, p. 284). Three new intron positions have been discovered in a *Chlamydomonas eugametos* globin gene (EMBL database entry CELI637) [M. Couture, H. Chamberland, B. St-Pierre, J. L. Lafontaine, M. Guertin, *Mol. Gen. Genet.* **243**, 185 (1994)]. Because a small number of diverse globins cannot be aligned reliably, we derived our alignment from a larger alignment of 91 globins and globin-related proteins with the use of Multalin, a multiple alignment program [F. Corpet, *Nucleic Acids Res.* **16**, 10881 (1988)].
28. The four reference proteins are horse liver ADH, subunit A [H. Eklund, J.-P. Samama, T. A. Jones, *Biochemistry* **23**, 5982 (1984)], horse β-deoxyhemoglobin [W. Bolton and M. F. Perutz, *Nature* **228**, 551 (1970)], feline muscle PK [H. Muirhead *et al.*, *EMBO J.* **5**, 475 (1986)], and chicken muscle TPI, α subunit [D. W. Banner *et al.*, *Nature* **255**, 609 (1975)]. Refined Cα coordinates for feline muscle PK were obtained from H. Muirhead (personal communication), who checked our list of secondary structural boundaries inferred from figure 2 of Muirhead *et al.* (cited above). Other structural data were obtained from the files pdb6adh.ent (ADH), pdb2dhb.ent (globin), and pdb1tim.ent (TPI) at the Brookhaven Protein Data Bank [F. C. Bernstein *et al.*, *J. Mol. Biol.* **112**, 535 (1977); E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, J. Weng, Protein Data Bank, in *Crystallographic Databases—Information Content, Software Systems, Scientific Applications*, F. H. Allen, G. Bergerhoff, R. Sievers, Eds. (Data Commission of the International Union of Crystallography, Cambridge, 1987), pp. 107–132].
29. Two reference models for intron positions have been used. In the uniform model, each reference set consists of *N* nonidentical intron positions drawn randomly with uniform probabilities from the set of all possible sites, where *N* is the number of intron positions in the observed set. In the *PIID* (permuted interintronic distances) model, reference sets are generated by randomly permuting the distances between the intron positions in the observed set. Results presented here are from the uniform model, but both models give similar results, including an absence of *P* values below the 1% or 5% critical level. Algorithms for reference models relied on a long-period pseudorandom number generator with a shuffle, the ran2 generator [W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes in C* (Cambridge Univ. Press, London, ed. 2, 1992), p. 282.
30. Proponents of the hypothesis that exon-encoded peptides are compact maintain that nonidentical intron positions from different copies of a gene actually represent the same ancestral intron. The difference in position is attributed to a hypothetical process of intron movement. For the purpose of testing extensity, we assume that intron positions within an arbitrarily chosen limit of three codons represent the same ancestral intron, whose position is defined as the average of the extant positions. With this rule, we infer a set of 20 ancestral exons for ADH, 11 for globins, 15 for PK, and 12 for TPI [the last apparently similar to the set of 12 TPI exons inferred by others (*10, 25*)].
31. Two reference models for exon sizes have been used. In the lognormal model, a reference set consists of *N* exon sizes (in integral numbers of codons) with the same sum and lognormal mean and SD as the inferred ancestral set of *N* exons [a lognormal distribution is consistent with the actual distribution; a Kolmogorov-Smirnov one-sample test applied to the 58 inferred ancestral exon sizes in base pairs gives a maximum difference of 0.097 ($P = 0.18$, not significant)]. In the permutation model, a reference set consists of a random permutation of the order of inferred ancestral exon sizes (in integral numbers of codons). Results presented here are from the lognormal model, but both models give similar results, including an absence of *P* values below the 1% or 5% critical level.
32. M. Zuker, unpublished results (programs available over the Internet by anonymous FTP from the "pub" directory at "nrcbsa.bio.nrc.ca").
33. _____ and A. Stoltzfus, unpublished results.
34. R. M. Stephens and T. D. Schneider, *J. Mol. Biol.* **228**, 1124 (1992).
35. C. S. Craik, W. J. Rutter, R. Fletterick, *Science* **220**, 1125 (1983).
36. N. J. Dibb and A. J. Newman, *EMBO J.* **8**, 2015 (1989). These authors introduced a notation by which *X*–0, *X*–1, and *X*–2 designate intron positions between codon *X* and the preceding codon, after the first nucleotide of codon *X*, and after the second nucleotide of codon *X*, respectively.
37. C. Branden and J. Tooze, *Introduction to Protein Structure* (Garland, New York, 1991), pp. 142–146.

38. We thank H. Muirhead (University of Bristol, UK) for providing refined Cα coordinates for feline muscle PK, M. Guertin for providing unpublished data on *Chlamydomonas* globin genes, and W. Blanchard, B. Dixon, R. Doolittle, D. Hickey, A. Ellington, D. Mark, J. Parrish, K. Robison, and C. Wallace for their encouragement, advice, and comments. Supported by the Canadian Institute for Advanced Research (Program in Evolutionary Biology), the Medical Research Council of Canada, the National Research Council of Canada, and NIH grant GM-35087 (to J. D. Palmer for J.M.L.).