

RESEARCH ARTICLES

A Sequence-Based Model Accounts Largely for the Relationship of Intron Positions to Protein Structural Features

Danny W. De Kee, Vivek Gopalan, and Arlin Stoltzfus*

Center for Advanced Research in Biotechnology, Rockville, MD; and *Biochemical Science Division, National Institute of Standards and Technology, Gaithersburg, MD

Claims of intron-structure correlations have played a major role in debates surrounding split gene origins. In the formative (as opposed to disruptive or “insertional”) model of split gene origins, introns represent the scars of chimaeric gene assembly. When analyzed retrospectively, formative introns should tend to fall between modular units, if such units exist, or at least to exhibit a preference for sites favorable to chimaera formation. However, there is another possible source of preferences: under a disruptive model of split gene origins, fortuitous intron-structure correlations may arise because the gain of introns is biased with respect to flanking nucleotide sequences. To investigate the extent to which a sequence-biased intron gain model may account for the present-day distribution of introns, data on over 10,000 introns in eukaryotic protein-coding genes were integrated with structural data from a set of 1,851 nonredundant protein chains. The positions of introns with respect to secondary structures, solvent accessibility, and so-called “modules” were evaluated relative to the expectations of a null model, a disruptive model based on amino acid frequencies at splice junctions, and a formative model defined relative to these. The null model can be excluded for most structural features and is highly improbable when intron sites are grouped by reading frame phase. Phase-dependent correlations with secondary structure and side-chain surface accessibility are particularly strong. However, these phase-dependent correlations are explained largely by the sequence-based disruptive model.

Introduction

The question of what factors determine the positions of introns has been given much attention over the last 25 years (Blake 1978; Go 1981; Tittiger et al. 1993; Stoltzfus et al. 1994; Logsdon et al. 1995; de Souza et al. 1996; Rzhetsky et al. 1997; Liu et al. 2005). Though a variety of factors might be invoked to account for the positions of introns, explanations often fall into 2 categories (Jellie et al. 1996; Qiu et al. 2004). In the “formative” category of explanation, the positions of introns reflect events of gene formation, as in the “exon theory of genes” (Gilbert 1987). In the “disruptive” or “insertional” interpretation, the positions of introns reflect evolutionary addition of introns to a gene region that was not split previously (Cavalier-Smith 1991; Palmer and Logsdon 1991).

A key implication of the formative view, as first recognized by Blake (1978), is that exons in protein-coding genes will tend to correspond to structural or functional units of proteins. A correspondence of exons to globular domains was anticipated originally (Blake 1978), but it soon became clear (Campbell and Porter 1983; Go 1983) that exons in animal genes are generally too short for such a correspondence. Later studies focused on spatially compact subdomain regions (Go 1981) referred to as “modules,” surface-accessible regions (Craik et al. 1983) and secondary structure motifs (Duester et al. 1986; Gilbert et al. 1986). These various claims led to a consensus of opinion that intron-structure correlations were the strongest evidence for a formative view of introns (Doolittle 1987), until a statistical analysis revealed that no single claim could be justified when applied consistently to multiple cases (Stoltzfus et al. 1994).

Key words: intron evolution, secondary structure, sequence preferences, splice site.

E-mail: stoltzfu@umbi.umd.edu.

Mol. Biol. Evol. 24(10):2158–2168. 2007

doi:10.1093/molbev/msm151

Advance Access publication July 23, 2007

Published by Oxford University Press 2007.

Subsequent work on this issue has not resolved clearly the question of how introns correlate with structural features or what are the causes of this relationship. On the one hand, analysis of cases of exon shuffling (Patthy 1991) provides a clear proof of a formative role for introns and reveals a highly nonrandom tendency for introns to be located in interdomain regions (Liu and Altman 2003). On the other hand, with respect to estimating the extent to which the formative model accounts for the totality of intron positions in extant genes, the situation is much more confusing. Initially, advocates of the formative view (Gilbert 1987) tended to assume that each and every intron had a special structural meaning, its position—plus or minus a few nucleotides due to putative “sliding”—reflecting some formative event of primordial assembly or subsequent exon shuffling. In the late 1990s, a weaker formative theory emerged in a series of papers that continued to claim structural evidence for primordial formation of genes from exons, but no longer cited the totality of evidence from intron positions, instead focusing only on a subset of phase 0 introns that were claimed to fall in the boundaries between compact modules (de Souza et al. 1996, 1998; Roy et al. 1999; Fedorov et al. 2001). Others continue to argue the formative view from a proposed correlation with secondary structure (Contreras-Moreira et al. 2003; Barik 2004).

Meanwhile, it has become clear that the positions of introns could exhibit nonrandomness for a completely unrelated reason: the evolutionary process of intron gain exhibits nucleotide preferences, favoring the pattern MAG \wedge GT, where “ \wedge ” is the site of gain (Qiu et al. 2004; Sverdlov et al. 2004). This preference may be responsible for much of the nonuniformity in intron phases because the “target site” is nonuniformly distributed among phases due to biased codon usage (Ruvinsky et al. 2005; Nguyen et al. 2006). This same preference is also relevant to understanding intron–protein correlations. Long before Qiu et al. showed that the MAG \wedge GT pattern flanking introns is largely due to

intron gain, Fichant (1992) had shown that the MAG^GT pattern flanking introns largely explains biased amino acid composition at intron junctions, for example, a tendency for phase 1 introns to fall in a glycine codon (i.e., G^GN). Obviously, because protein features correlate with amino acid sequence features, biased amino acid composition at a set of sites may result in biased structural properties for that same set of sites. Thus, nucleotide sequence preferences for intron gain are expected to generate intron-structure correlations, and it is possible that such preferences are largely or wholly responsible for any observed intron-structure correlations, as suggested by Stoltzfus et al. (1994).

In this study, we combine data on intron locations and protein structural features in order to assess the evidence for, and the causes of, any relation that might exist between intron positions and structural features of proteins. In principle, these data could be used to assess 3 models for such a relationship: a null model in which intron sites are a random sample of all sites, a disruptive model in which intron sites are biased due only to sequence preferences of intron gain, and a formative model in which intron sites are biased according to chimaerogenic potential. In practice, the formative model is poorly specified due to the lack of reliable measures of chimaerogenic potential, whereas the sequence-biased gain model can yield precise predictions due to the clear relationship between nucleotide and protein sequences (i.e., the genetic code) and the availability of data on the structural propensities of amino acids.

This analysis yields 2 conclusions. First, the null model can be rejected: there are systematic correlations between positions of introns and structural features, particularly when the introns are separated according to phase. Second, the model of sequence-biased intron gain largely explains the deviations from randomness. For instance, phase 1 intron positions tend to avoid helices and to favor coils, but these preferences are explained quantitatively by a glycine preference consistent with the MAG^GT nucleotide preference for intron gain. These conclusions apply across vertebrate, invertebrate, plant, and fungal sets of genes, and they do not depend on excluding animal-specific (AS) gene families, which show a similar pattern. In general, the formative model does not have an important role in accounting for the overall pattern observed in present-day genes, though it might become important to account for some subset of introns defined by phylogenetic or other restrictions, an issue that is not resolved by the results reported here.

Methods

Approach

The main aim of this work is to determine the extent to which a sequence-biased gain model accounts for any observed correspondence between intron positions and protein structural features. To do so, we define a null model, H_0 , and a disruptive model, H_D ; in some cases, it is possible to define a formative model, H_F , relative to H_D or H_0 . In the approach used here, these models generate quantitative knowledge-based predictions by incorporating observed frequencies and structural propensities of amino acids, as described below.

Null Model

In the null model, H_0 , intron locations are chosen without respect to sequence or structure so that all possible intron sites are equally probable. For instance, under the null model, if 30% of residues are assigned to α -helices, then 30% of intron sites are expected to fall in α -helices, or more generally:

$$P_{\mathbb{S}} = k_{\mathbb{S}}, \quad (1)$$

where $P_{\mathbb{S}}$ is the probability of an intron position that, relative to the encoded protein structure, maps to a site assigned to secondary structural element \mathbb{S} (helix, strand, and coil), and $k_{\mathbb{S}}$ is the fraction of that same type of secondary structural element in a large nonredundant set of proteins. To distinguish the null hypothesis from H_D , we may specify further that the null model is not sensitive to phase

$$P_{\phi, \mathbb{S}} = k_{\mathbb{S}}, \quad (2)$$

where $P_{\phi, \mathbb{S}}$ is the probability of a phase ϕ intron position that maps to a secondary structural element of type \mathbb{S} .

Disruptive Model

In the disruptive model, H_D , any structural correlations with intron sites are due solely to biases in amino acid composition that reflect the nucleotide preferences for intron gain (Qiu et al. 2004; Sverdlov et al. 2004). Qiu et al. conclude specifically that these preferences are taxon specific (at least at the level of kingdoms) and that, even within a taxon, the intron gain preferences cannot represent a single target sequence such as AG^G but must reflect a mixture or statistical profile. Therefore, it is not possible to compute precise quantitative expectations of the disruptive model without an equally precise statistical profile of kingdom-specific preferences for intron gain. Unfortunately, no such precise profile is available.

To circumvent this difficulty, we assume that amino acid frequencies at splice junctions can be used to reflect the effects of nucleotide preferences independent of structural effects. Thus, we evaluate H_D in terms of an amino acid frequency model, in which the correlations of intron positions with structural features are predictable entirely based on the observed frequencies of amino acids near splice junctions. For instance, in the case of secondary structure:

$$P_{\phi, \mathbb{S}} = \sum_i f_{i, \phi} q_{i, \mathbb{S}}, \quad (3)$$

where $P_{\phi, \mathbb{S}}$ is defined as before, $f_{i, \phi}$ is the frequency with which an amino acid i is associated with a phase ϕ intron position and $q_{i, \mathbb{S}}$ is the propensity of any amino acid i in a protein, to map to a secondary structural element, \mathbb{S} .

Note 2 complications of this model. First, given that prior work (Qiu et al. 2004) revealed taxonomic differences in nucleotide preferences for intron gain, amino acid frequencies near splice junctions must be computed taxon specifically (however, we do not define the structural propensities of amino acids taxon specifically). Second, given that significant nucleotide preferences for intron gain extend from nucleotide sites -3 to $+2$ relative to the intron

site (Qiu et al. 2004), multiple amino acid sites may be affected even when considering a single intron phase, and each site would exhibit a different biased composition. However, because the preferences are weak except from -2 to $+1$, we assume that such effects can be ignored for phase 1 and 2 sites. For phase 0 introns, we consider 2 kinds of amino acid sites, the upstream or “phase 3” site and the downstream or “phase 0” site.

Refined (di-amino acid) Disruptive Model

The amino acid frequency model assumes that amino acids are independent. However, phase 0 introns lie between 2 codons that are both simultaneously affected by the MAG \wedge GT preference for intron gain. Thus, under the disruptive model, the CAG-encoded glutamines upstream of a phase 0 intron are not a random sample of glutamines, but are a biased sample in that they tend to be followed by valine residues encoded by GTN. The di-amino acid disruptive model is similar to the disruptive model except that 2 adjacent sites are treated together and are weighted by the frequencies of pairs of amino acids.

$$P_{ij,q} = \sum_{i,j} f_{ij} q_{ij}, \quad (4)$$

where $P_{ij,q}$ is the probability of a pair of amino acids (one upstream and one downstream of the intron) in a secondary structural element, f_{ij} is the frequency of a pair of amino acids, and q_{ij} is the propensity of a pair of adjacent amino acids to be in a particular structural secondary assignment, where there are 9 such assignments ($\alpha\alpha$, $\alpha-$, $\alpha\beta$, $\beta\alpha$, $\beta-$, $\beta\beta$, $-\alpha$, $--$, $-\beta$).

Formative Model

The formative model, H_F , is a correlation in excess of the null model, or in excess of H_D , and in a direction consistent with increased chimaerogenic potential. In the case of secondary structure and surface accessibility, H_F is poorly defined. Though it is often assumed that secondary structures are modules to some extent, the available evidence that might support this assumption is not strong (e.g., DuBose and Hartl 1989); thus, one does not know for certain how chimaerogenic potential is distributed relative to the secondary structure map. If modularity is dominated by secondary structure, then intersecondary structure sites will tend to be intermodule sites, but it also may be that the main chimaerogenic units are tertiary modules with boundaries that tend to interrupt secondary structures. With respect to surface accessibility, there is some evidence that it is easier to add to a protein by inserting fragments in surface-accessible loops (Benner et al. 1997); therefore, H_F implies increased surface accessibility.

In the case of the “modules” invoked by Gilbert and colleagues, it might seem obvious that the formative model predicts an excess of between-module introns relative to the expectations of H_0 and, if applicable, H_D . Indeed, other things being equal, this seems a reasonable expectation, and we follow it here. However, one should not assume this uncritically because it has never been shown that bound-

aries between these alleged modules are more likely sites for chimaera formation, though such a demonstration might be possible using available data (Voigt et al. 2002).

Structure Analysis

The representative sets of proteins are derived from the Protein Data Bank (PDB) database (Berman et al. 2000). The data set was filtered to remove redundancy at 30% sequence identity level by using sequence clustering program Blastclust (Altschul et al. 1990).

Secondary structural definitions are based on the DSSP program (Kabsch and Sander 1983). The resulting frequencies of helix, sheet, and coil are similar to those found by other researchers in large sets of data (Martin et al. 2005). Residues are classified as being associated with a helix, strand, or coil, using the entire set of 4,659 prokaryotic and eukaryotic genes.

The surface accessibility was determined using NACCESS (Hubbard and Thornton 1993) based on the method of Lee and Richards (1971). This program computes surface accessibility in either absolute (\AA^2) or relative (% relative to standard condition Ala-Xxx-Ala) terms, summing over all atoms, side-chain only or backbone only. In the absence of a clear prior rationale for preferring one of the resulting 6 measures, we chose the most residue-sensitive measure (the most likely to be influenced under hypothesis H_D), which is absolute side-chain accessibility, and the least residue-sensitive measure (the most likely to be influenced under hypothesis H_F), which is relative backbone accessibility.

The procedure for defining “modules” used by de Souza and colleagues was implemented in Perl, using a diameter value of 27.6 \AA , chosen based on their previous work reporting correlations between introns and “modules” of 27.6 \AA (de Souza et al. 1996). The Perl implementation was validated by a direct side-by-side comparison of the output with that of the original INTERMODULE program (de Souza et al. 1996).

Eukaryotic Gene Database

The eukaryotic gene information is obtained by a protocol modified from that used to construct the Intron Database (IDB) (Schisler and Palmer 2000). This new version of IDB (Hladish T, Schisler NJ, and Stoltzfus A, unpublished data) is a GenBank-based (version 142) eukaryotic protein-encoding gene database and is cross-referenced to SwissProt in order to remove redundancy. Entries with incomplete sequences or with noncanonical splice sites are removed for the analysis. The National Center for Biotechnology Information taxonomy database was used for classification of the IDB entries into 4 taxonomic groups, vertebrates, invertebrates, plants, and fungi. The current IDB contains 165,451 full-length entries.

Mapping of Intron Positions on Protein structures

The representative protein chains are aligned against the intron-containing subset of the IDB by utilizing BlastP.

Table 1
The Number of PDB Chains and Introns for the Four Taxonomic Groups

Taxonomic Group	Number of PDB Chains	Number of Introns
Vertebrates	765	4,247
Invertebrates	708	2,069
Plants	648	4,050
Fungi	229	593

A cutoff of 30% identity is used for identifying homologs. The top hit in each taxonomic group is used for this study, for example, the plant data set consists of the top plants hits to 648 PDB chains and includes 4,050 introns. Representative chains (1,354) had at least one intron mapped to a protein structure. The results of these alignments are summarized in table 1.

Removal of AS Gene Families

Pathy (1999) has shown clear evidence that exon shuffling events in mosaic proteins are specific to a metazoan origin. The intron positions in these proteins are present in protein domain boundaries and may have biased structural properties. Hence, the nonredundant protein structures were classified into 2 categories: animal-specific (AS) genes and without-animal-specific (WAS) genes. These classifications are based on the Blast homologues; the AS data set contains genes whose homologs are animal genes only and the WAS data set contains all other families.

Results

Taxonomically Defined Groups of Intron-Containing Homologs

Sequences for an initial set of 6,444 nonredundant structures of protein chains were pruned by removing sequences less than 90 residues and then aligned against the intron-containing subset of the IDB to identify homologs with at least 30% identity, yielding 4,659 protein chains (1,240,720 residues). For each of 4 taxonomic groups (vertebrates, invertebrates, plants, and fungi), the top hit of at least 30% was retained, resulting in the 4 sets of data described in table 1.

Each of these taxonomically defined sets was analyzed using the same methods, with results that exhibited an overall similarity in spite of some differences. To shorten the presentation of results, here we present complete results only for the vertebrates. Complete results for the other taxonomic groups are presented as Supplementary Material (see Supplementary Material online) and are cited in the text where there are important similarities or significant differences.

Null Model

The vertebrate set of observed intron positions mapped onto protein structures has a secondary structure distribution of 46.0% coil, 38.4% helix, and 15.6% sheet (table

Table 2
Observed and Expected Secondary Structure and Module Distributions for All Intron Positions According to the Null Model (95% CI)

Type	Subtype	Expected %	Observed %	<i>P</i> value
Secondary Structure	Coil	45.0	46.0 ± 1.9	0.71
	α-Helix	35.6	38.4 ± 1.8	0.06
	β-Strand	19.4	15.6 ± 1.4	<0.001
Module	27.6 Å	33.1	33.1 ± 1.8	1.00

NOTE.—The *P* value is based on the χ^2 statistic.

2; see Supplementary Material table S2 online for other data sets). As indicated in table 2, the expected distribution under the null model, which is just the distribution of secondary structure assignments for all amino acid sites (see Methods), is very similar: 45.0%, 35.6%, and 19.4% for coil, helix, and sheet, respectively (similar to the values observed in other large data sets: see Martin et al. 2005). Although the observed distribution deviates only slightly from the null model, the observed frequency of intron sites that map to β-strands, 15.6 ± 1.4%, is significantly lower than the expected value of 19.4%, that is, introns tend to avoid β-strands.

Table 2 also displays the frequency with which intron sites map to linker regions between “modules” as defined by the algorithm of de Souza et al. (1996). The observed value of 33.1% corresponds closely with the value expected under the null hypothesis (table 2).

None of the deviations are significant for the other taxonomic groups (table S2, Supplementary Material online).

With respect to surface accessibility, various measures are used commonly and none is definitive. Here we use absolute side-chain accessibility (in Å²) as a residue-sensitive measure suitable for detecting the distinctive implications of the disruptive model, and relative backbone accessibility (relative to the accessibility in an extended Ala-Xxx-Ala tripeptide) as a measure that is more indicative of protein secondary and tertiary structure, and thus more suitable for detecting the distinctive implications of the formative hypothesis. The observed intron positions mapped on the protein structures have a mean relative backbone accessibility of 27.0 ± 1.5% and a side-chain absolute accessibility of 32.6 ± 1.6 Å², as compared with expected values of 25.9% (*P* value = 0.06, based on *Z* scores) and 35.0 Å² (*P* value = 0.001). The backbone and side-chain measures are predicted very well for invertebrates (*P* values are 0.14 and 0.48, respectively), whereas there are significant differences among the plants and fungi (data not shown).

A more detailed implication of the null hypothesis *H*₀, as distinct from the disruptive model, is that the structural properties associated with intron sites should be insensitive to phase. For this purpose, although there are only 3 possible intron phases, we may consider 4 types of sites: “phase 1” sites where the amino acid is encoded by a codon interrupted by a phase 1 intron, “phase 2” sites corresponding to a codon interrupted by a phase 2 intron, and “phase 3” and “phase 0” sites representing the upstream and downstream (respectively) codons flanking an intron between 2 codons. It is important to distinguish phase 3 and phase 0 sites because, although they are equivalent

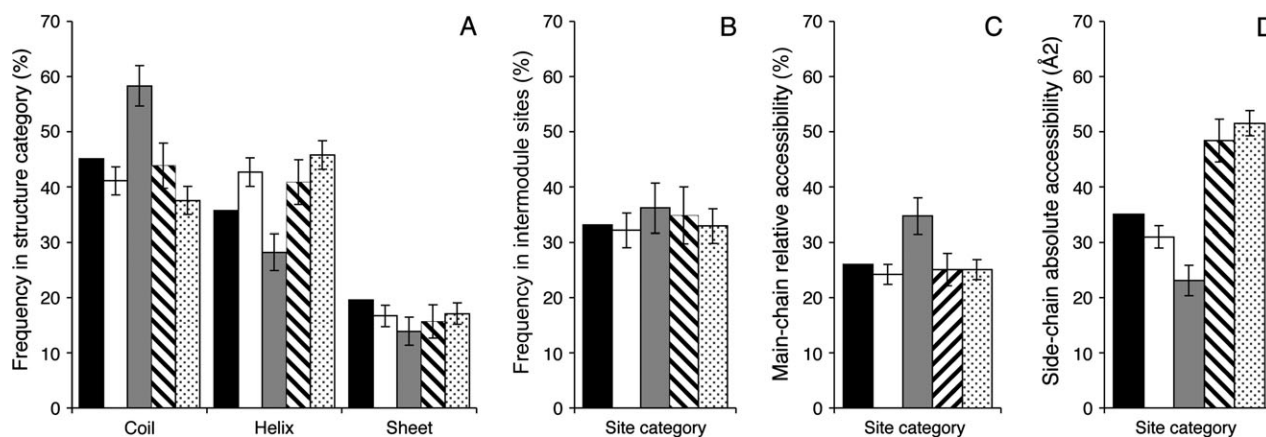


FIG. 1.—Intron-associated sites in proteins have phase-specific structural features. Observed values (error bars, 95% CI) for structural correspondences are shown in comparison to expected values (black bars) for 4 categories of intron-associated sites from the vertebrate data set (for other data sets, see supplementary fig. S1, Supplementary Material online): phase 0 (open bars), phase 1 (gray bars), phase 2 (striped bars), and phase 3 (stippled bars). The structural features are (A) frequency with which a protein site is assigned to 1 of 3 secondary structure categories, (B) frequency with which the site falls in a linker region between modules, (C) the main-chain relative surface accessibility at the site, and (D) the side-chain absolute surface accessibility for the residue at that site.

under the null model and the formative model, they are not equivalent under the disruptive model (as explained in Methods).

Figure 1 shows the results of dividing the data into these 4 phases of sites. Overall, these results reveal large and significant deviations from the null expectation of no effect. For instance, for phase 1 and 2 sites in the vertebrate data set (fig. 1A), the observed frequencies with which sites map to helices are 7.4% lower and 5.1% higher, respectively, than the null expectation of 35.7%. In general, among diverse taxa (fig. S1A, Supplementary Material online), phase 0, 2, and 3 sites are underpredicted in helical regions and overpredicted in coils, whereas this trend is reversed for phase 1 sites, which are observed to be much more common in coils than expected. Thus, the null model is excluded clearly for the case of secondary structure.

The locations relative to “modules” shown for vertebrate data in figure 1B do not exhibit significant deviations; however, there are small but significant deviations for the plant data at phase 3 sites and for invertebrate data at phase 2 sites (supplementary fig. S1B, see Supplementary Material online), with the observed intermodule frequencies being slightly higher than expected. For surface accessibility, shown for the vertebrate data in figure 1C and D, the 2 different measures reveal somewhat different effects. For backbone relative accessibility (the measure that should be more sensitive to H_F), the main effect shown for vertebrates in figure 1C is an excess accessibility of phase 1 sites; the same effect is seen for invertebrates and plants (fig. S1C, Supplementary Material online; the effect in fungi is in the same direction but is not statistically significant). For side-chain absolute accessibility (the measure that should be more sensitive to H_D) shown in figure 1D, accessibility is overpredicted for phase 0 and 1 sites and underpredicted for phase 2 and 3 sites; the same pattern of significant over- and underprediction is seen in invertebrates and plants (fig. S1D, Supplementary Material online; for fungi, phase 0 and 1 sites are underpredicted, but phase 2 and 3 sites do not differ significantly from expected values).

Thus, the intron-structure relationship is radically phase-dependent, as predicted under the disruptive model but not under the null or formative model. This phase dependence is strongest for secondary structure and solvent accessibility and is relatively minor for modules.

Disruptive Model

In the disruptive model (see Methods), structural preferences arise solely from amino acid preferences that are interpreted to reflect nucleotide preferences of MAG \wedge GT at sites of intron gain (Qiu et al. 2004; Sverdlov et al. 2004). Prior to this work, it has been established clearly that the distribution of amino acids at intron sites is nonuniform and phase dependent in a manner consistent with a nucleotide signal (Fichant 1992; Whamond and Thornton 2006). Table 3 summarizes the amino acid preferences at intron sites for the vertebrate data, revealing a pattern qualitatively similar to those reported earlier by others. From the Supplementary Material (fig. S2), it may be noted that the pattern of preferences for intron sites in invertebrates, plants, and fungi is qualitatively similar to that seen for vertebrate data, but the pattern is much less pronounced for fungi, for example, the entropy of amino acid composition at intron-associated sites is 3.9 for fungi, 3.3 for plants, 3.4 for vertebrates, and 3.7 for invertebrates. This difference is not due to a difference in the background level of amino acid composition bias, which has an entropy of 4.2 regardless of taxonomic source.

Figure 2 shows the observed distribution of the structural attributes for vertebrate data relative to the expectations of the disruptive model, that is, assuming that the observed distribution of structural attributes is entirely a side effect, not a cause, of amino acid composition biases that themselves reflect sequence-biased intron gain. From figure 2A–D, showing secondary structure data 4 classes of sites, it is clear that the disruptive model largely accounts for the phase-dependent nonuniformity seen earlier (fig. 1A). For instance, while the distribution of phase 1 sites in secondary

Table 3
Amino Acid Frequency Distribution at Intron Positions

	Residue	Codons	Phase 0	Phase 1	Phase 2	Phase 3	All Phases
Ala	A	GCN	9.50	8.62	0.70	2.15	7.42
Cys	C	TGY	1.18	1.11	1.92	0.21	1.32
Asp	D	GAY	7.84	10.57	0.87	1.80	7.09
Glu	E	GAR	7.56	6.82	2.61	15.33	6.33
Phe	F	TTY	2.91	0.28	0.87	0.49	1.79
Gly	G	GGN	9.15	51.60	9.06	1.73	20.29
His	H	CAY	1.18	0.42	0.87	0.83	0.91
Ile	I	ATH	7.49	1.53	0.70	0.62	4.50
Lys	K	AAR	3.12	0.83	9.76	27.74	3.91
Leu	L	CTN, TTR	7.42	0.70	2.61	5.41	4.64
Met	M	ATG	2.22	0.14	1.74	2.50	1.57
Asn	N	AAV	3.05	1.11	3.48	1.94	2.63
Pro	P	CCN	2.43	0.28	1.74	2.57	1.72
Gln	Q	CAR	1.18	0.14	2.96	24.76	1.28
Arg	R	CGN, AGR	2.01	1.25	36.41	3.74	9.03
Ser	S	TCN, AGY	3.68	2.78	14.63	2.08	5.74
Thr	T	CAN	3.40	1.11	1.57	2.50	2.41
Val	V	GTN	22.33	9.46	1.74	2.36	14.63
Trp	W	TGG	0.90	0.42	5.23	0.62	1.68
Tyr	Y	TAY	1.46	0.83	0.52	0.62	1.10

NOTE.—For instance, the codon upstream of an intron, that is, what we call the “phase 3” amino acid site, tends to be either lysine (AAG) or glutamine (CAG), whereas the downstream “phase 0” site has a preference for valine (GTN).

structure deviates radically from the null expectation (fig. 1A), this same distribution does not deviate significantly from the expectation of the disruptive model for vertebrates (fig. 2B; for the invertebrate, plant, and fungal data sets: fig. S3B, Supplementary Material online).

Although the disruptive model is an improvement for the vertebrate data set and for the other taxonomic groups (see Supplementary Material online), it does not fully account for secondary structure, as shown by the small but significant deviations in figure 2A, C and D. Helices are

underpredicted at phase 0 and phase 3 sites, whereas coils are overpredicted at phase 3 sites.

Likewise, the disruptive model does not fully account for the distribution of intron sites relative to “modules”. Although there are not significant deviations for the vertebrate data shown in figure 2E, the small but significant deviations noted earlier for the null model at phase 3 sites in plants and at phase 2 sites in invertebrate still remain and are not explained by the disruptive model (fig. S3E, Supplementary Material online).

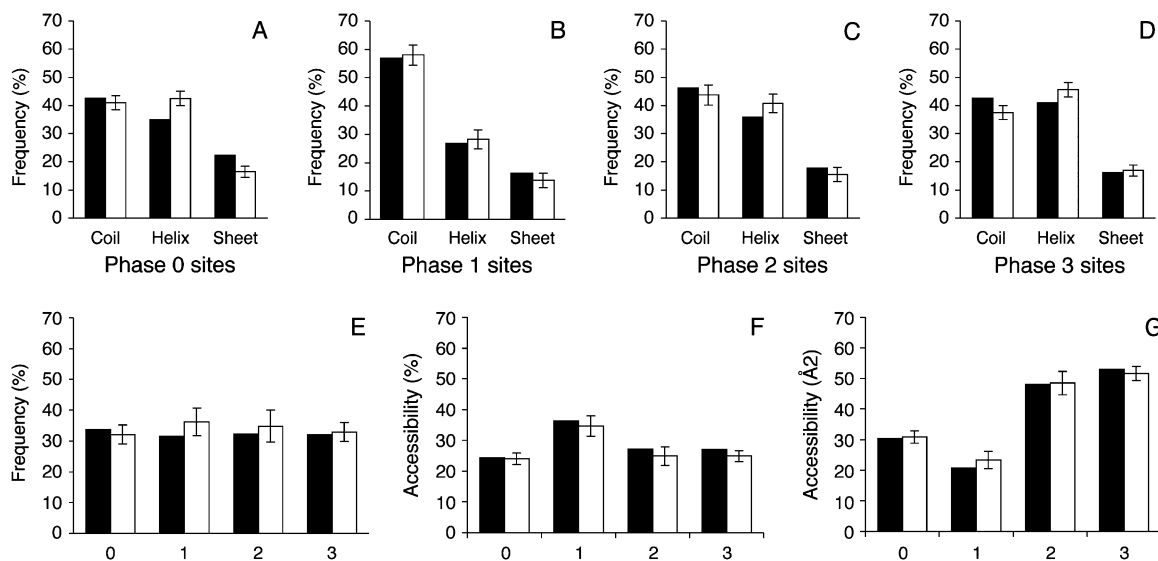


FIG. 2.—The amino acid model accounts largely for phase-associated nonuniformity in structural properties of intron-associated sites. Observed values (open bars; error bars, 95% CI) for structural correspondences are shown in comparison to expected values (black bars) from the disruptive model based on amino acid frequencies. Results shown here are for the vertebrate data set (for other data sets, see supplementary fig. S2, Supplementary Material online). Panels (A) through (D) show results for secondary structure for phase categories 0, 1, 2, and 3, respectively; panel (E) shows results for all 4 phase categories relative to intermodule linkers; and panels (F) and (G) show main-chain relative surface accessibility and side-chain absolute accessibility, respectively, for all 4 phase categories.

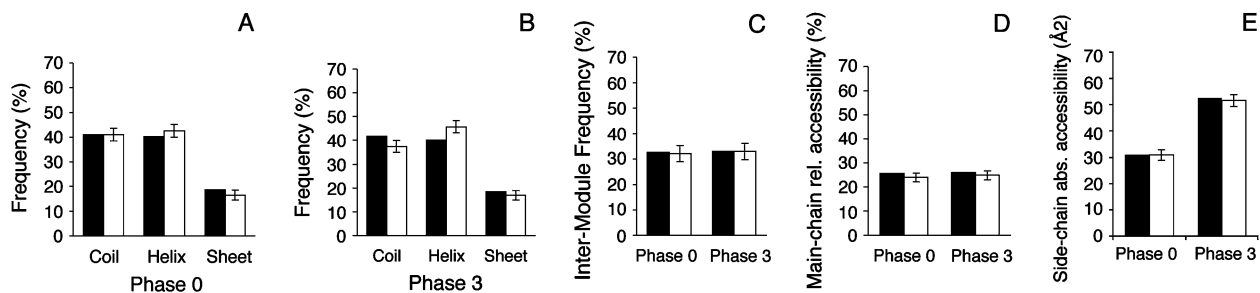


FIG. 3.—The di-amino acid model accounts for structural correspondences of intron sites slightly better than the amino acid model. Graphic conventions are as in figure 2, except that, because the di-amino acid model only differs from the amino acid model for phase 0 and phase 3 intron-associated sites, results are shown only for these sites. Panels (A) and (B) show results for secondary structure, panel (C) shows results for intermodule linkers, and panels (D) and (E) show main-chain relative surface accessibility and side-chain absolute accessibility, respectively. Results are for the vertebrate data set (for other data sets, see supplementary fig. S3, Supplementary Material online).

Finally, the disruptive model accounts largely for the dramatic deviations from null expectations in regard to surface accessibility. For absolute side-chain accessibility, the disruptive model accounts for the nonuniformity among sites observed in the vertebrate data shown in figure 2G and that observed in the other taxonomically defined data sets (fig. S3G, Supplementary Material online), with the exception that accessibility is overpredicted for phase 0 and phase 3 sites in plants. For relative backbone accessibility, the disruptive model accounts for the vertebrate (fig. 2F) and invertebrate results (supplementary fig. S3F, see Supplementary Material online), whereas for plants and fungi, accessibility is overpredicted consistently although the disruptive model correctly predicts the order of site classes (phase 1 > phase 2 > phase 3 > phase 0 sites for plants and phase 1 > phase 3 > phase 2 > phase 0 sites for fungi; fig. S3F, Supplementary Material online).

How much better is the disruptive model than the null model? One way to assess the difference is simply to count significant deviations. For the vertebrate data, out of 24 comparisons with the null expectation there are 16 deviations (11 for secondary structure, 0 for “modules,” 1 for backbone relative accessibility, and 4 deviations for side-chain absolute accessibility). By contrast, for the vertebrate data, the disruptive model leaves only 5 deviations (2 for phase 0 secondary structure, 1 each for phases 2 and 3 secondary structure, and 1 for the backbone relative accessibility). For all data (not just vertebrates), there are 48 deviations out of 96 for the null model and 30 deviations for the disruptive model.

Revised (di-amino acid) Disruptive Model

An obvious means to improve the analysis above would be to take into account the nonindependence of phase 3 and phase 0 sites expected under the disruptive model. Nonindependence is expected because the nucleotide preferences for intron gain affect both sites at once. Thus, the valine residues encoded (by GTN codons) just downstream of introns are not a random sample of all valines in proteins because they tend to be preceded by the upstream residue glutamine (CAG) or lysine (AAG).

To account for this anticipated effect of nonindependence, the structural properties of intron positions were predicted based on the structural propensities of pairs of amino

acids, weighted by the paired frequencies observed at intron sites (see Methods), with results shown in figure 3. As can be seen, this di-amino acid model improves the fit with observed values for phase 0 sites but not for phase 3 sites, for example, the observed value of 45.6 ± 2.6 for helix at phase 3 sites is still significantly higher than the expected value of 40.1 (P value < 0.001). For all taxonomic groups, the tendency at phase 3 sites is for helices to be underpredicted and coils to be overpredicted, and these deviations are significant for the plant and vertebrate data sets. In general, the di-amino acid model represents a significant but only modest improvement over the original model.

Formative Model

The formative model suggests that, other things being equal, intron sites should avoid α -helices and β -strands and instead should fall in coil regions (Lonberg and Gilbert 1985). In fact, when intron sites are not divided by phase, there is no such significant tendency in vertebrates, as shown in table 2 (nor in the other taxonomic groups: table S2, Supplementary Material online). When sites are divided into phases 0, 1, 2, and 3, only phase 1 sites correlate with coils (in vertebrates, as well as for invertebrate and plant data sets), whereas other sites show a deficit; this deficit is significant for phase 0 and 3 sites (see fig. 1A and fig. S1A, Supplementary Material online). However, as noted above, the pattern at phase 1 sites is explained by the disruptive model (fig. 2B and fig. S2B, Supplementary Material online). Thus, there is nothing for the formative model to explain.

For the case of “modules,” again, there is not a significant overall tendency for vertebrate introns to fall in intermodule regions (table 2; likewise for other taxonomic groups: table S2, Supplementary Material online). When sites are divided by phase, there is no significant correlation for the vertebrate data, but for the invertebrate data there is a phase-2-specific excess of “intermodule” introns, and for the plant data an excess of intermodule introns at phase 0 and 3 sites (fig. S2E, Supplementary Material online). As noted above, these idiosyncratic deviations are not explained by the disruptive model. Thus, because both these are excesses (instead of deficits), they suggest the formative model.

The formative model suggests that, other things being equal, intron sites should be more surface accessible.

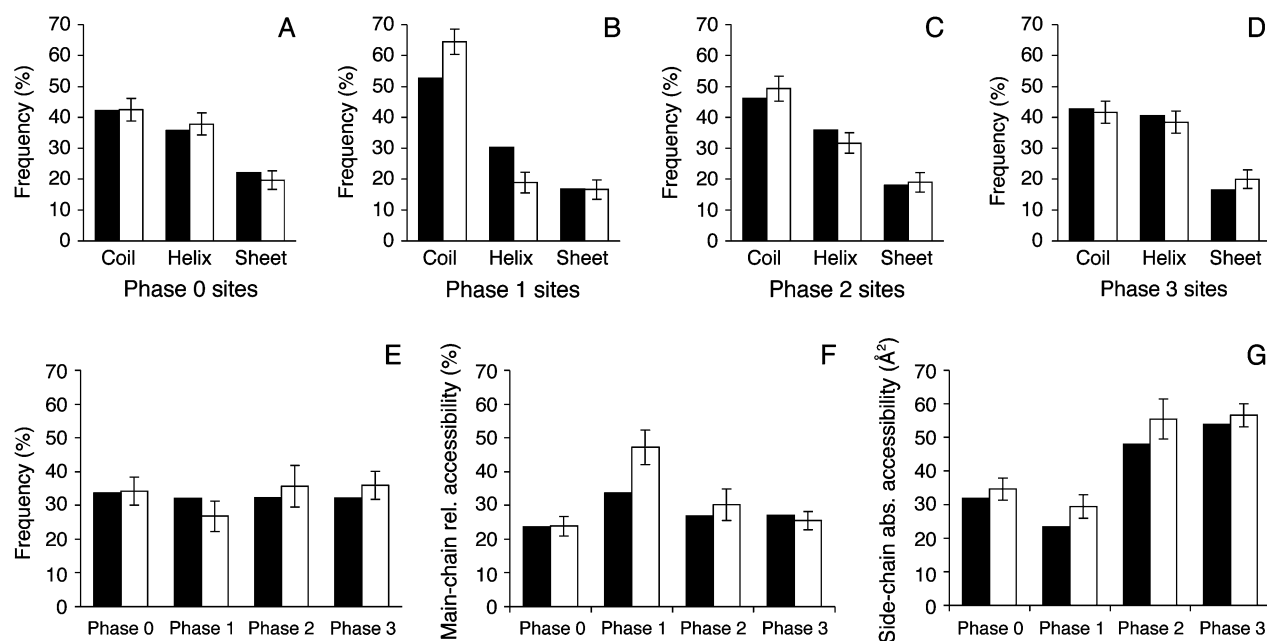


FIG. 4.—Correspondences for the AS data set are slightly different but are explained largely by the di-amino acid model. Observed values (open bars; error bars, 95% CI) for structural correspondences are shown in comparison to expected values (black bars) from the di-amino acid disruptive model, and for vertebrate data from AS gene families (for the invertebrate data from AS genes, see supplementary fig. S4, Supplementary Material online). As in figure 2, panels (A) through (D) show results for secondary structure, panel (E) for intermodule regions, and panels (F) and (G) for surface accessibility.

However, the opposite effect is seen when introns are not divided by phase. Side-chain absolute accessibility at intron sites is significantly lower than expected for vertebrate data (also for plant and fungal data; supplementary table S2, see Supplementary Material online). When sites are divided by phase, correlations occur in both directions: for side-chain absolute accessibility—phase 0 and 1 sites typically are significantly lower than expected in accessibility, whereas phase 2 and 3 sites are higher; for main-chain relative accessibility, phase 1 sites are significantly more accessible than expected, whereas the other sites are slightly (typically insignificantly) less accessible (table S2.5, Supplementary Material online). As noted earlier, the disruptive model accounts for these results; thus, there is nothing for the formative model to explain.

AS Genes Data Set

For the WAS genes data set described above, the secondary structural distribution at intron positions can be predicted with considerable accuracy using a disruptive model based only on observed amino acid frequencies. This same approach can be extended to an AS data set. As with the WAS data set, the null model is strongly excluded when phase is taken into account (data not shown; fig. 4 from the Supplementary Material online illustrates 12 deviations out of 24 comparisons). Figure 4S summarizes the results of applying the disruptive model to the AS data set, using the di-amino acid version of the model for phases 3 and 0. As with the WAS data set, the disruptive model accounts for most of the pattern of nonrandomness seen among phases in the AS data set, but is not perfect. In particular, the model

underpredicts coil at phase 1 sites, which at 64.4% have a frequency 6.4% higher in the AS data set than in the WAS data set.

Discussion

Two main results follow from a quantitative analysis of data on 11,334 intron positions in relation to a null model of no preferences, a disruptive model based on amino acid preferences, and a formative model. First, it is possible to exclude the null model for a variety of types of possible correlation, for example, the observed percentage distribution of β -strand is 15.6 ± 1.4 (95% confidence interval), whereas the expected percentage distribution is 19.4. Second, the disruptive model accounts largely for the most distinctive patterns of nonrandomness in the distribution of intron sites in relation to structure, which are in regard to secondary structure and solvent accessibility, whereas the formative model has little to explain.

To demonstrate this point clearly and to show that it applies across taxonomic groups (not merely to the vertebrate data presented in Results), a summary for all taxonomic groups is shown in figure 5. The strongest correlations with structural features are in regard to secondary structure (A) and side-chain absolute accessibility (D), while any tendency to fall between “modules” (B) is weak or insignificant. The greatest deviations are explained largely by the disruptive model. That is, although some significant deviations remain, the large deviations from expected values under the null model fade to being insignificant or only marginally significant under the disruptive model. The di-amino acid model offers a slight

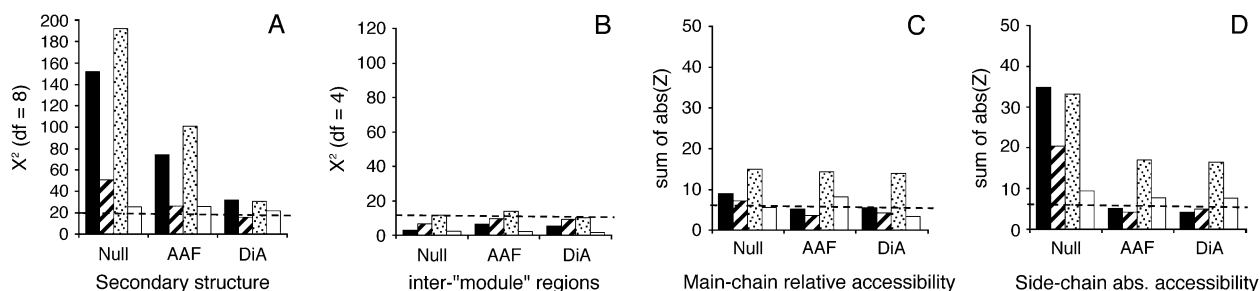


FIG. 5.—The disruptive model accounts largely for phase-specific correspondences observed in diverse taxonomically defined data sets. Each histogram shows, for each taxonomically defined data set, the combined deviation from expected values for 3 models (null, disruptive, di-amino acid) for the case of secondary structure (A), intermodule location (B), main-chain relative solvent accessibility (C), and side-chain absolute solvent accessibility (D). Each cluster of 4 bars represents results for vertebrates (black), invertebrates (gray), plants (light gray), and fungi (white), with the height of each bar being a measure of deviation combining 4 position-specific deviations (results for AS genes data sets are given in supplementary fig. S5, see Supplementary Material online). For (A) and (B), these deviations are χ^2 values (note that the degrees of freedom are 8 in A and 4 in B); for (C) and (D), the deviations are sums of absolute Z scores. To aid in comparisons, each panel has a dashed line representing $P = 0.01$. For the strongest correlations, shown in panels A and D, the correlation is explained largely by the disruptive model; in panel C, the correlation is weak and taxon specific, and the disruptive model offers only a modest improvement; in panel B, the correlation is very weak and taxon specific, and the disruptive model offers no overall improvement.

improvement. The overall trend of improvement in models is similar for the AS data sets, but the improvement is not as dramatic (fig. S5, Supplementary Material online).

This conclusion is stronger than that allowed from the results of a recent analysis by Whamond and Thornton (2006), who also addressed the relationship of intron positions to secondary structure within a disruptive model. In that study, the mean Euclidean distance between observed and expected frequency distributions (for locations of introns with respect to helix, sheet, and coil) decreased from 0.12 (average for 3 phase-specific distances) for the null model to 0.10 for the model of nucleotide-biased intron sites, whereas in the present study, the comparable values are 0.11 (null), 0.064 (disruptive), and 0.047 (di-amino acid disruptive) for vertebrate data (values for plant and invertebrate data are comparable; for the fungal data, the improvement is trivial, as is apparent from fig. 5; table S4, Supplementary Material online). Thus, the disruptive model used by Whamond and Thornton improves only slightly on the null model, in contrast to the major improvement found here (fig. 5). Possible reasons for this difference are that Whamond and Thornton began with nucleotide biases rather than amino acid biases (i.e., a more theoretical, less empirical prediction model, thus a more demanding prediction model), they combined data from diverse taxonomic groups (whereas we treat the groups separately to allow for heterogeneity), and they used unusual secondary structure assignments that yield background frequencies outside the range found here or in other studies (e.g., Martin et al. 2005).

The conclusions of this analysis may appear to contradict some statements of Gilbert and colleagues arguing for a formative “introns-early” view (in which primordial protein-coding genes emerge from fusing separate exon minigenes) based on the claim that phase 0 introns shared between different taxonomic kingdoms show an unusual tendency to fall between modules (de Souza et al. 1996, 1998; Roy et al. 1999; Fedorov et al. 2001, 2003).

However, the appearance of a contradiction mainly reflects divergent aims. This study aims to assess how well one can account for the distribution of introns in present-day genes using a quantitative predictive model that has

some theoretical content (a theory of intron evolution) as well as some ad hoc empirical content (amino acid frequencies, residue-averaged structure propensities); we focus on the big picture and do not attempt to track down all possible sources of minor deviations. By comparison, Gilbert and colleagues are focused on detecting a specific subtle signal that they interpret as a formative signal of introns-early; they do not attempt to assess the extent to which this signal accounts for nonrandomness observed in present-day gene structures. Such an assessment would be impossible given that the formative theory is not sufficiently well specified to make quantitative predictions. That is, there is no theoretical basis under the formative model for specifying precisely where introns are to be found relative to protein structural features or to phases. Instead, any predictions related to these quantities are relative (i.e., a value higher or lower than some null expectation) and conditional: under the formative theory, introns, or a subset of introns that one might discover (e.g., phase 0 introns or ancient introns), will tend to be found in positions that have a higher chimaerogenic potential than expected, which might correspond (or might not) to sites between secondary structures (e.g., Lonberg and Gilbert 1985) or sites between the modules of de Souza et al. (1998).

As a result of this difference in aims, the scope and methods of this analysis are different in several important ways. First, when we report that the disruptive model accounts largely for intron-structure correlations and, particularly for the strongest correlations, we are including aspects of structure—namely secondary structure and surface accessibility—that 1) are recognized universally to be important aspects of protein structure but 2) are not addressed by de Souza et al. and Fedorov et al. Because various past claims for a formative role of introns (e.g., Lonberg and Gilbert 1985) relied on arguments from secondary structure subsequently found to be weak (Stoltzfus et al. 1994), the decision by de Souza et al. and Fedorov et al. not to address secondary structure arises, not because secondary structure is irrelevant to the formative model but because an analysis of secondary structure is *ex posteriori* unlikely to satisfy their aim of finding a signal for introns-early.

Second, for similar reasons, we do not focus our interpretation of results on phase 0 introns but consider a sample representing the totality of introns in present-day genes. Third, because our main focus is on accounting for present-day genes, we do not consider whether patterns might be different for a small set of introns that might be old.

With these differences in mind, finally, one may ask whether results reported here in regard to modules are consistent with prior claims, a comparison made difficult by the fact that prior claims are not consistent with each other. Although previous arguments had emphasized 28 Å modules, de Souza et al. (de Souza et al. 1996, 1998) argued that the relevant “module” correlation involved 3 different sizes—21, 28, and 33 Å—and that the signal was limited to phase 0 introns in ancient conserved regions (ACRs) and for unknown reasons was not found in vertebrate genes. In subsequent more extensive analyses (Fedorov et al. 2001, 2003), the results for phase 0 introns in ACRs revealed a single peak of significance involving 25 Å “modules,” while the tendency in regard to 21, 28, and 33 Å “modules” was insignificant or marginal, and comparable in scale to other patterns that did not figure in the formative interpretation offered by the authors, for example, a tendency for phase 2 introns to fall between 30 Å “modules” (fig. 2 of Fedorov et al.).

Because the present study treats taxonomic groups separately and combines data from ACRs and non-ACRs, there is no direct comparison with these earlier studies. However, for the sake of making some comparison, one may hope that results reported here for 27.6 Å modules (the size chosen for this study, following the initial claims for the modules algorithm by [de Souza et al. 1996]) would fall somewhere between the values for ACRs and non-ACRs. For ACRs, the relative excess (the deviation divided by the expected value) of phase 0 introns between 27.6 Å “modules” is roughly 4%, with $\chi^2 \approx 3$ (fig. 2A of Fedorov et al., top row), whereas for non-ACRs, the excess is about 2% with $\chi^2 < 1$ (fig. 2B of Fedorov et al., top row). That is, the correspondence is small and insignificant. By comparison, the phase 0–ACR relation with 25 Å “modules” that is the basis of the authors’ formative interpretation represents a relative excess of roughly 5–8%, corresponding to χ^2 values typically from 4 to 7 (fig. 2A of Fedorov et al. 2001, top row, or fig. 1 of Fedorov et al. 2003), that is, significant for the $\alpha = 0.05$ critical level but not for $\alpha = 0.005$. Thus, from the work of Fedorov et al., we expect that any tendency for phase 0 introns to fall between 27.6 Å “modules” will be small, with a relative deviation of less than a few percent. However, given the low apparent repeatability of this type of analysis, reflected in the contradiction between de Souza et al. and Fedorov et al., one should allow for some volatility.

The results presented here are consistent with this very modest expectation. When all phases are combined, we find no correspondence with “modules” for the vertebrate data (consistent with de Souza et al. 1998) and a slight but insignificant excess of “intermodule introns” for the other taxonomic groups where the relative excess is about 3%. When sites are divided by phase, there are 2 significant excesses of “intermodules” sites relative to the null model (or

relative to the disruptive model, which is nearly the same): an excess of phase 2 sites in invertebrate data and of phase 3 sites in plant data. The latter correspondence implicates phase 0 sites and represents a relative excess of about 8%, consistent with prior claims. The correspondence involving phase 2 invertebrate sites is greater in degree (a 21% excess) but less statistically significant. When data are combined across taxonomic groups, the overall excess of phase 0 sites between 27.6 Å “modules” is about 3.6% and is not statistically significant, also consistent with prior claims. If the pattern uncovered by Fedorov et al. 2003 holds, then the statistical excess is due to intron sites shared between taxa and presumed, on this basis, to be old. However, this issue is not addressed here.

Supplementary Material

Supplementary tables and figures are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

This work was supported by National Institutes of Health grant R01-LM007218 to A.S. The identification of specific commercial software products is for the purpose of specifying a protocol and does not imply a recommendation or endorsement by the National Institute of Standards and Technology.

Literature Cited

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215: 403–410.
- Barik S. 2004. When proteome meets genome: the alpha helix and the beta strand of proteins are eschewed by mRNA splice junctions and may define the minimal indivisible modules of protein architecture. *J Biosci.* 29:261–273.
- Benner SA, Cannarozzi G, Gerloff D, Turcotte M, Chelvanayagam G. 1997. Bonafide predictions of protein secondary structure using transparent analyses of multiple sequence alignments. *Chem Rev.* 97:2725–2844.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.
- Blake CCF. 1978. Do genes-in-pieces imply proteins-in-pieces? *Nature.* 273:267.
- Campbell RD, Porter RR. 1983. Molecular cloning and characterization of the gene coding for human complement protein factor B. *Proc Natl Acad Sci USA.* 80:4464–4468.
- Cavalier-Smith T. 1991. Intron phylogeny: a new hypothesis. *Trends Genet.* 7:145–148.
- Contreras-Moreira B, Jonsson PF, Bates PA. 2003. Structural context of exons in protein domains: implications for protein modelling and design. *J Mol Biol.* 333:1045–1059.
- Craik CS, Rutter WJ, Fletterick R. 1983. Splice junctions: association with variation in protein structure. *Science.* 220: 1125–1129.

- de Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W. 1998. Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. *Proc Natl Acad Sci USA*. 95:5094–5099.
- de Souza SJ, Long M, Schoenbach L, Roy SW, Gilbert W. 1996. Intron positions correlate with module boundaries in ancient proteins. *Proc Natl Acad Sci USA*. 93:14632–14636.
- Doolittle WF. 1987. What introns have to tell us: hierarchy in genome evolution. *Cold Spring Harb Symp Quant Biol*. 52:907–913.
- DuBose RF, Hartl DL. 1989. An experimental approach to testing modular evolution: directed replacement of alpha-helices in a bacterial protein. *Proc Natl Acad Sci USA*. 86:9966–9970.
- Duester G, Jornvall H, Hatfield GW. 1986. Intron-dependent evolution of the nucleotide-binding domains within alcohol dehydrogenase and related enzymes. *Nucleic Acids Res*. 14:1931–1941.
- Fedorov A, Cao X, Saxonov S, de Souza SJ, Roy SW, Gilbert W. 2001. Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. *Proc Natl Acad Sci USA*. 98:13177–13182.
- Fedorov A, Roy S, Cao X, Gilbert W. 2003. Phylogenetically older introns strongly correlate with module boundaries in ancient proteins. *Genome Res*. 13:1155–1157.
- Fichant GA. 1992. Constraints acting on the exon positions of the splice site sequences and local amino acid composition of the protein. *Hum Mol Genet*. 1:259–267.
- Gilbert W. 1987. The exon theory of genes. *Cold Spring Harb Symp Quant Biol*. 52:901–905.
- Gilbert W, Marchionni M, McKnight G. 1986. On the antiquity of introns. *Cell*. 46:151–153.
- Go M. 1981. Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature*. 291:90–92.
- Go M. 1983. Modular structural units, exons, and function in chicken lysozyme. *Proc Natl Acad Sci USA*. 80:1964–1968.
- Hubbard SJ, Thornton JM. 1993. NACCESS. Department of Biochemistry and Molecular Biology, University College London.
- Jellie AM, Tate WP, Trotman CN. 1996. Evolutionary history of introns in a multidomain globin gene. *J Mol Evol*. 42:641–647.
- Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 22:2577–2637.
- Lee B, Richards FM. 1971. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*. 55:379–400.
- Liu M, Walch H, Wu S, Grigoriev A. 2005. Significant expansion of exon-bordering protein domains during animal proteome evolution. *Nucleic Acids Res*. 33:95–105.
- Liu S, Altman RB. 2003. Large scale study of protein domain distribution in the context of alternative splicing. *Nucleic Acids Res*. 31:4828–4835.
- Logsdon JM Jr, Tyshenko MG, Dixon C, Jafari JD, Walker VK, Palmer JD. 1995. Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc Natl Acad Sci USA*. 92:8507–8511.
- Lonberg N, Gilbert W. 1985. Intron/exon structure of the chicken pyruvate kinase gene. *Cell*. 40:81–90.
- Martin J, Letellier G, Marin A, Taly JF, de Brevern AG, Gibrat JF. 2005. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Struct Biol*. 5:17.
- Nguyen HD, Yoshihama M, Kenmochi N. 2006. Phase distribution of spliceosomal introns: implications for intron origin. *BMC Evol Biol*. 6:69.
- Palmer JD, Logsdon JM Jr. 1991. The recent origins of introns. *Curr Opin Genet Dev*. 1:470–477.
- Pathy L. 1991. Modular exchange principles in proteins. *Curr Opin Struct Biol*. 1:351–361.
- Pathy L. 1999. Genome evolution and the evolution of exon-shuffling—a review. *Gene*. 238:103–114.
- Qiu WG, Schisler N, Stoltzfus A. 2004. The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol Biol Evol*. 21:1252–1263.
- Roy SW, Nosaka M, de Souza SJ, Gilbert W. 1999. Centripetal modules and ancient introns. *Gene*. 238:85–91.
- Ruvinsky A, Eskesen ST, Eskesen FN, Hurst LD. 2005. Can codon usage bias explain intron phase distributions and exon symmetry? *J Mol Evol*. 60:99–104.
- Rzhetsky A, Ayala FJ, Hsu LC, Chang C, Yoshida A. 1997. Exon/intron structure of aldehyde dehydrogenase genes supports the “introns-late” theory. *Proc Natl Acad Sci USA*. 94:6820–6825.
- Schisler NJ, Palmer JD. 2000. The IDB and IEDB: intron sequence and evolution databases. *Nucleic Acids Res*. 28:181–184.
- Stoltzfus A, Spencer DF, Zuker M, Logsdon JM Jr, Doolittle WF. 1994. Testing the exon theory of genes: the evidence from protein structure. *Science*. 265:202–207.
- Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV. 2004. Reconstruction of ancestral protosplice sites. *Curr Biol*. 14:1505–1508.
- Tittiger C, Whyard S, Walker VK. 1993. A novel intron site in the triosephosphate isomerase gene from the mosquito *Culex tarsalis*. *Nature*. 361:470–472.
- Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH. 2002. Protein building blocks preserved by recombination. *Nat Struct Biol*. 9:553–558.
- Whamond GS, Thornton JM. 2006. An analysis of intron positions in relation to nucleotides, amino acids, and protein secondary structure. *J Mol Biol*. 359:238–247.

William Martin, Associate Editor

Accepted June 25, 2007