Intron Framing Exonic Nucleotides: A Compromise Between Protein Coding and Splicing Constraints

A. Ruvinsky*,[1] and W. Ward[1,2]

[1]*Institute for Genetics and Bioinformatics, University of New England, Armidale 2350 NSW, Australia*

[2]*School of Mathematics, Statistics and Computer Sciences, University of New England, Armidale 2350 NSW, Australia*

Abstract: Introns in eukaryotic genes are located either between codons (phase 0) or within codons (phase 1 and 2). Phase 0 introns are more frequent. Several factors might contribute to this phenomenon with codon usage bias playing a significant role. The nucleotides located at the very ends of intermediate exons are involved not only in protein coding but also in splicing regulation. This study indicates that phase 0 introns create more flexibility for protein coding without affecting splicing sensitive exonic nucleotides than the other two intron types. The canonic **AG↓G** site, for instance, is particularly frequent around phase 0 introns. In humans the observed frequency of **AG↓G** sites framing phase 0 introns is at least 2 to 3 times higher than in phase 1 and 2 introns. It is possible that the higher flexibility of exonic nucleotides surrounding phase 0 introns may serve as a driving force increasing frequencies of sites like **AG↓G** and this could lead to more stable or efficient splicing without compromising protein coding. If so, this type of selection might also contribute to higher frequency of phase 0 introns.

Keywords: Exon, intron, splicing, codon usage frequency, information content.

INTRODUCTION

Introns are located either between codons (phase 0) or within codons, after the first nucleotide (phase 1) and after the second (phase 2). Phase 0 introns are more common and phase 2 introns are rare [1]. Several hypotheses were proposed to explain the distribution of intron phases, including exon shuffling [1], sequence conservation of splice signals in exons [2], correlation with regions of amino acid conservation for phase 1 & 2 introns [3], specific elements of protein structure [4] and intron sliding [5]. Recent simulations showed that species specific codon usage frequencies might affect intron phase distribution [6] and the observed bias in the intron phase distribution is, at least in part, caused by codon usage frequencies.

The nucleotides located at the very ends of intermediate exons are involved not only in protein coding but also in splicing regulation. Composition of such exonic nucleotides framing introns is diverse and even the most typical **AG↓G** site does not exceed 30-35% around phase 0 introns. The frequency of **AG↓G** site around phase 1 and 2 introns is significantly lower. This study indicates that phase 0 introns provides more flexibility for protein coding without affecting splicing sensitive exonic nucleotides. Higher flexibility of exonic sites around phase 0 may also support mutation driven increase in frequencies of few exonic sites, like **AG↓G**, which could improve stability or efficiency of splicing without compromising protein coding. The two other intron types follow a similar path but with more limitations depending on species.

Splicing is strongly influenced by factors such as nearly absolute uniformity of intronic splicing sites [7], as well as existence of enhancers and silencers located at some distances from the intron-exon junctions [8]. It can be hypothesized that initially exonic nucleotides surrounding introns might not be critical for securing correct splicing. However, some sites, like **AG↓G**, might become more suitable than others and, if so, this would lead to a positive selection and eventual increase of **AG↓G** frequency, as well as few other desirable sites. Phase 0 introns possibly created better opportunities for such mutation driven shift in favour of “canonic” sites like **AG↓G**.

MATERIALS AND METHODS

Gene Data

Information relevant to *A. thaliana*, *C. elegans*, *D. melanogaster* and *H. sapiens*, was extracted from the exon-intron database (EID, version 112), which was compiled in the W. Gilbert laboratory, Department of Molecular and Cellular Biology, Harvard University [9]. The initial database was extensively purged by J. Chamary (University of Bath, UK). The removal of potential duplicates was done after performing an all-against-all BLAST, with an expected value of $P < 0.001$ [10], and creating clusters of duplicated genes. The longest of the duplicate genes were left in the database. This procedure was based on the assumption that, in the case of alternative transcripts, the longest is the constitutive form. Even if this is not the case, it is just an arbitrary way of selecting one sequence. Then one from the 'longest' duplicates, if several are of the same length, was randomly selected. The authors will provide the dataset on request. The final dataset consisted of 2033 genes of *H. sapiens* (18465 exons), 8502 genes from *D. melanogaster* (37856 exons), 10312 genes from *C. elegans* (69180 exons) and 9914 genes from 9914 (65766 exons).

DB and Software

From the purged databases; four separate exon databases, for each of the model species {*Hs* (*Homo sapiens*), *Dm*

*Address correspondence to this author at the Institute for Genetics and Bioinformatics, University of New England, Armidale NSW 2351, Australia; Tel: (61) 267 73 3900; Fax: (61) 267 73 3275;
E-mail: aruvinsk@une.edu.au

1874-4044/08 2008 Bentham Science Publishers Ltd.

(*Drosophila melanogaster*), *Ce* (*Caenorhabditis elegans*) and *At* (*Arabidopsis thaliana*)}, were created using a Perl script. The sequences were read into objects using modules from the BioPerl toolkit and regular expressions were used to iterate through the gene sequences identifying each exon. In addition to the exon sequences, the exon length, nucleotide number for the starting position of each exon, its 5' and 3' phases, and the gene ID numbers were extracted and included as fields in the databases. The primary keys were created by appending the exon number to the ID number of each gene ID. Outputs from the program were loaded into the Postgres DBMS, and SQL and Perl scripts which were used to generate the data.

Expected Frequencies of AG↓G and Statistical Analysis

Expected frequencies of **AG↓G** for model species were calculated for phase 0 as follows. Frequency of **AG** representing two last nucleotides of codons preceding the following introns: $Fr_{\mathbf{AG}}$ = (Σ freq. of N**AG** codons)/(1-(Σ freq of stop codons). Frequency of **G** as the first nucleotide of codons which follow the introns: $Fr_{\mathbf{G}}$ = (Σ freq. of **G**NN codons)/(1-Σ (freq of stop codons). Finally $Fr_{\mathbf{AG|G}} = Fr_{\mathbf{AG}} * Fr_{\mathbf{G}}$. Similar logic was applied to calculation of **AG↓G** for phase 1 and 2 with relevant modifications. Codon usage frequencies for a particular species were taken from the database located at http://www.kazusa.or.jp/codon/

Statistical comparisons were performed using common methods like contingency χ^2 test and Fisher-Yates φ test.

Information Content of Intron Framing Exonic Nucleotides

The information content (**IC**) of an exonic site can be calculated using the standard formula $\mathbf{IC} = \sum \{\mathbf{P(x_i)} * -\mathbf{log_4}\ \mathbf{P(x_i)}\}$, where $\mathbf{x_i}$ is a frequency of one of four nucleotides [11]. If four possible nucleotides at a particular position have equal frequencies (0.25), **IC** reaches maximum value of 1. If only one nucleotide can be found at a position, then **IC** reaches minimum value of 0. Essentially **IC** measures similarity of nucleotides frequencies at a position. A lower **IC** indicates more disparity between frequencies of nucleotides and preference for particular nucleotide(s), a higher **IC** on the contrary indicates more similarity between nucleotides frequencies. Using nucleotides frequencies presented in Table **S1**, we calculated information content for each nucleotide at four exonic positions framing introns (-2,-1↓1,2). Then average **IC** values were calculated for each phase in four model species.

RESULTS

Observed and Expected Frequencies of AG↓G Sites

The exonic nucleotides immediately surrounding introns should satisfy at least two independent requirements: protein coding and splicing. In some instances a conflict between these competing requirements may occur. For instance, a strong preference in favour of a particular nucleotide, which might be beneficial for stable splicing can limit possible changes in coding sequence and thus slow down or prevent certain changes in respective proteins. As introns are inserted in one of three phases (0, 1 and 2), the conflict has three potential resolutions. One of these resolutions is possibly more flexible as it may provide a greater number of codon combinations with higher frequencies, which will not necessarily affect preferable splicing combination of exonic nucleotides. If so, this might create mutation driven bias in favor of preferable exonic nucleotides framing introns, like **AG↓G**. Different intron phases might not be equal in supporting such process.

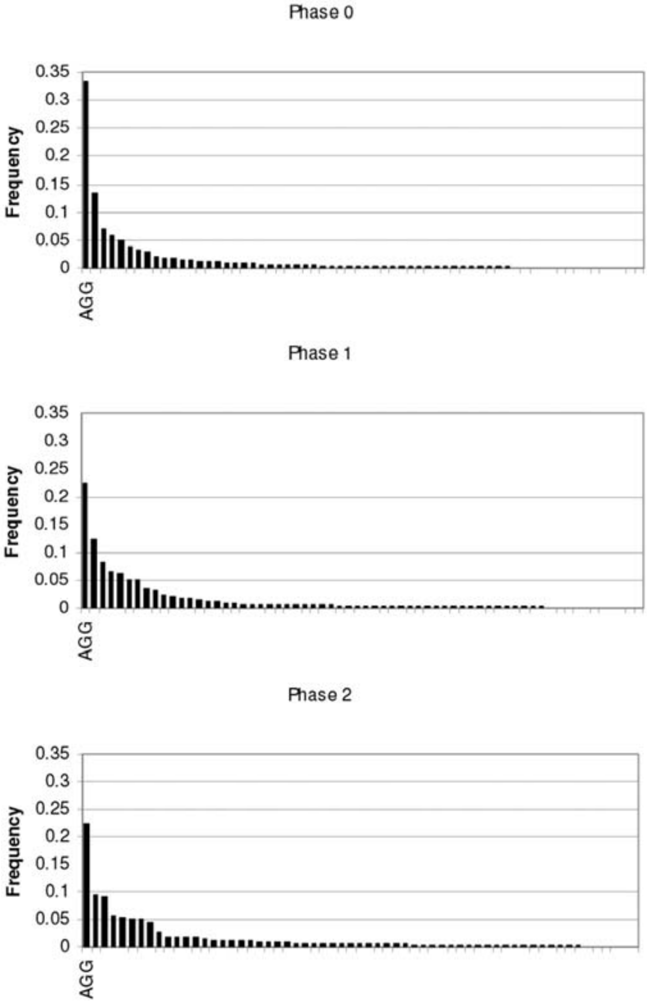
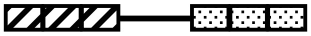
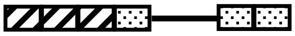
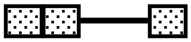


Fig. (1). Observed frequencies of intron framing exonic sites {-2;-1 ↓ +1} in *Homo sapiens*. Position of **AG↓G** can be seen at the X axis; all other nucleotide combinations are not shown. Frequencies of the exonic sites add up to 100% for each phase.

One of the most frequent exonic splice sites is **AG↓G** (Fig. **1**), which is considered as a canonic sequence for humans and other species [12]. While there are numerous signals located in introns and exons that strongly affect splicing [8], here we shall concentrate attention only on exonic nucleotides framing introns and **AG↓G** in particular. Table **1** demonstrates that the number of codons or combination of codons, which allow the **AG↓G** sequence, differs significantly in different phases. There is only one codon (namely AGG) which satisfies the condition for phase 2, while phases 1 and 0 have much greater flexibility because the number of possible codon combinations creating the **AG↓G** sequence is large (Table **1**). This fact alone indicates that phase 2 introns provide fewer opportunities for a reasonable resolution of the potential conflict between coding and splicing requirements. It is also essential that not only a number of codons or codon combinations but also codon usage affect

Table 1. Observed and Expected Frequencies of Intron Framing AG↓G Exonic Sites in Model Species

3’ A G G 5’	Phase	Number of codons or codon combinations which allow AG↓G as intron framing exonic site	Observed frequency of AG↓G exonic sites in: *				Expected frequency of AG↓G exonic sites in: **			
			Hs	*Dm*	*Ce*	*At*	*Hs*	*Dm*	*Ce*	*At*
	0	3*16 = 48	0.149	0.073	0.067	0.191	0.034	0.037	0.019	0.026
	1	14*4 = 56	0.076	0.028	0.025	0.045	0.017	0.016	0.015	0.018
	2	1	0.047	0.016	0.009	0.04	0.012	0.006	0.004	0.011

***AG↓G** is not the most frequent intron framing exonic site in *Ce* (all phases) and in *Dm* (phase 2) (data not shown). Frequencies were calculated by dividing the observed number of **AG↓G** in a certain phase to the total number of introns.
Expected frequencies of **AG↓G were calculated using codon usage frequencies, as described in Materials and Methods.

the frequencies of **AG↓G** or other intron framing exonic sequences.

Observed and expected frequencies of **AG↓G** in different phases are also shown in Table **1**. The expected frequencies were calculated using species specific codon usage frequencies as described in Material and Methods. Phase 0 introns are characterized by the highest frequencies of **AG↓G** in all studied species. In humans phase 0 introns framed by **AG↓G** are much more common than phase 1 introns; 0.149 against 0.076. Contingency χ^2 test shows significant differences between frequencies of exonic nucleotides framing phase 0 and 1 introns (χ^2 = 106.5; P<0.00001) as well as phase 0 and 2 introns (χ^2 = 73.8, P<0.00001). Both expected and observed frequencies of **AG↓G** were much lower in phase 1 than in phase 0 despite a large number of possible codon combinations leading to **AG↓G** sequence around phase 1 introns. This fact most likely indicates that frequencies of the relevant codons used to frame phase 1 introns are not high enough. Other factors should not be ruled out. For instance, Fedorov *et al.* [13] found anticorrelation of phase 1 intron positions with boundaries of protein modules, which might affect frequency of phase 1 introns.

The observed values of **AG↓G** for all intron types are significantly higher than expected frequencies. A significant bias in favour of observed **AG↓G** frequencies inevitably causes low frequencies of other exonic sites. For example, one of the rare sites **GC↓C** is dramatically underrepresented comparing with the expectation (the data are not shown). The same conclusion is correct for many other exonic sites which have low observed frequencies.

Phase 0 Introns Are More Likely Framed by AG↓G Sequence than Other Types of Introns

While the **AG↓G** sequence is the most common in *Homo sapiens*, *Arabidopsis thaliana*, *Drosophila melanogaster* and is among the most common in *Cearnohabditis elegans*, it is by no means the only one (Fig. **1**). Essentially all other possible combinations of nucleotides were observed and their total frequency varies from phase to phase and between species from 0.66 to 0.8. The first three most frequent exonic combinations in humans comprise 53.85% in phase 0 (AG↓G, AG↓A and AG↓C); 42.97% in phase 1 (AG↓G, AG↓A and TG↓G) and 41.10% in phase 2 (AG↓G, AG↓C and AG↓A). Thus frequencies of nucleotides at the exonic ends differ between phases and the nucleotide combinations similar to the canonic exonic site **AG↓G** are more common in phase 0.

Earlier published observations as well our own (supplementary Table **S1)** provide additional evidence that **AG↓G** is more common for phase 0 introns. Indeed, exonic nucleotides framing phase 0 show the best match to the canonic sequence **AG↓G** than either phase 1 or phase 2. For instance, in humans at the -2 position, **A** is the most common nucleotide in phase 0 (0.663); in phase 1 it equals 0.509 and phase 2 is 0.613. The frequency of nucleotide **G** at the -1 position has similar values in phase 0 and 1 (0.829 and 0.837) but phase 2 has a lower value of 0.697. Nucleotide **G** at the +1 position in phase 0 is equal to 0.558 and in phase 1 and 2 is lower (0.442 and 0.519). As mentioned earlier in humans the probability of **AG↓G** occurrence in phase 0 is higher then in phase 1 or 2. All statistical comparisons (contingency χ^2 test) of the corresponding positions between Ph0 and Ph1, Ph0 and Ph2, Ph1 and Ph2 demonstrate highly significant differences (P<0.00001).

Quite similar tendencies can also be seen in other model species. Frequencies of nucleotides in phase 0 are more consistent in all model species while phases 1 and 2 show more fluctuations (Table **S1**). A conclusion can be drawn that phase 0 is characterized by frequencies of intron framing exonic nucleotides that are significantly closer to the canonic **AG↓G** sequences than in the two other phases.

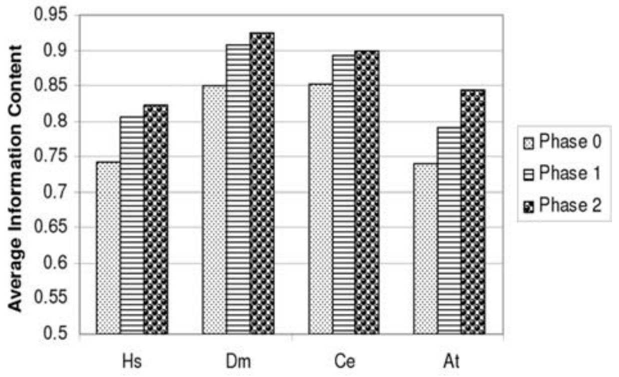


Fig. (2). Average information content per four intron framing nucleotides (positions -2,-1↓1,2) in model species. The differences between phase 0 and phase 1 & 2 are significant (P<0.00001) in all species.

Average Information Content of Intron Framing Exonic Nucleotides

The average **IC** values were calculated for each phase in four model species using the approach described in Material and Methods. Figure **2** shows that in all studied species average **IC** of four exonic sites framing phase 0 are significantly lower than in two other phases. In Hs the difference between phase 0 and phase 1 is significant (F= 61, P<0.00001), as well as the difference between phase 0 and phase 2 (F= 85.7, P<0.00001). In other species the differences are also statistically highly significant. This observation reaffirms other data described in this paper that the exonic sites framing phase 0 introns are more biased toward canonic combinations of nucleotides. Phase 1 and 2 introns are surrounded by less biased combinations of nucleotides. As mentioned earlier phase 0 introns create less coding restrictions and hence more flexibility in the surrounding codons and nucleotides. This may explain lower average IC values for exonic nucleotides framing phase 0 introns.

DISCUSSION

Exonic nucleotides, which frame introns, play at least two very different roles. One of them is obviously protein coding and another is splicing regulation even though other stronger splicing signals exist elsewhere [8]. It is well known, and our data support this, that there are just a few combinations of exonic nucleotides framing introns, like **AG↓G,** which are very common. All other numerous nucleotide combinations are much less frequent. This fact may indicate that such "canonic" sites are preferred in the splicing process. Phase 0 introns, as shown here, are more often framed by such few "canonic" sites than other intron types. Phase 1 and particularly phase 2 introns have more diverse framing with a smaller proportion of "canonic" sites.

Superior flexibility of exonic sites surrounding phase 0 introns is the likely explanation for the higher frequency of **AG↓G** and other "canonic" combinations. The essence of such flexibility is ability to reconcile a broad combination of codons surrounding phase 0 introns with specific splicing constrains applied to the exonic nucleotides adjacent to introns. Similarly to what was suggested by Fichant [14] we believe that composition of the exonic nucleotides surrounding introns might reflect a balance of competing constrains caused by protein coding and splicing requirements. It seems that phase 0 introns provide a better balance and more opportunities for the evolution of involved codons and splicing signals. The cause for less strict constrains imposed by phase 0 introns, in turn, is related to a number of codon combinations (and their frequencies) more compatible with the splicing requirements. Using different methods Long and Deutsch [2] also demonstrated that phase distribution of spliceosomal introns has strong correlation with the conservation of splice signals in exons. The most frequent phase 0 introns are associated with the highest conservation of surrounding exonic sites. Phase 2 introns, on the contrary, have less conserved sequences around them.

Earlier we have shown [6] that codon usage frequencies alone offer a satisfactory explanation to the observed distribution of intron phases in model species. It looks possible that species specific distributions of intron phases might be generally settled a long time ago. However, evolution of exonic sites framing introns was a lengthy process, and may still be evolving [15]. It is plausible that exonic nucleotides framing phase 0 introns were able to accumulate more splicing sites, like **AG↓G**, which provide more stable or efficient splicing. If so, such sites might represent the highest selection point and further mutational changes in these sites could be limited. In the course of evolution, sites like **AG↓G** became common at least in some eukaryotic genomes and should be less prone to further mutational changes. Our preliminary data show that substitution rate per **AG↓G** sites among ~1000 splicing sites from sets of 10 mammalian genes (0.130±0.008 substitutions per **AG↓G**) was significantly lower (t=2.35; P=0.015) than in all other possible exonic sites (0.207±0.004 substitutions per site). If this limited analysis were confirmed by the future studies, it could indicate that the compromise between coding and splicing requirements reaches its maximum in **AG↓G** sites. On the contrary, around phase 2 introns frequency of **AG↓G** is restricted by only one available arginine codon **AGG;** which is relatively rare in eukaryotic genomes. Thus, quite expectedly, **AG↓G** has low frequency in phase 2 and its further increase might be significantly restricted by the protein coding requirements. Assuming that more combinations of codons with higher frequencies can lead to "canonic" splicing sites framing phase 0 introns than in two other intron phases, one can speculate that the higher protein coding flexibility is an important cause for the phase 0 introns advantage. In other words phase 0 introns provide more flexibility for codon substitutions around the exon-intron junction without changing splicing sensitive "canonic" sites.

As intron positions during evolution remain very stable [16] and intron slidings, which could affect the phase distribution are rare [17], it seems likely that proportion of "canonic" sites, like **AG↓G,** steadily increased since the time of intron invasion particularly around phase 0 introns. Much higher observed frequencies of **AG↓G** comparing with the expected frequencies (Table **1**) support this view. If it is indeed correct that phase 0 introns present a better opportunity for resolving a conflict between coding and splicing requirements, this may constitute the essence of their selective advantage and may lead to steady increase of phase 0 introns at the expense of phase 2 introns and, at a lesser extent phase 1 introns. There are indications that novel introns at least in *C. elegans* are enriched by phase 0 introns [18]. It is not clear whether the frequency of phase 0 introns is higher among novel introns than among old introns. As far as we aware, there are no data showing that phase 2 introns were more frequently lost during evolution as hypothesized by Long and Deutsch [2]. Hopefully future investigations will address this and other relevant questions.

Finally, if one of the conflicting constraints (for instance protein coding) is not highly relevant to a particular gene, then the second requirement (for instance more stable or efficient splicing) could become more dominant and vice versa. It may mean that in different genes the balance between coding and splicing requirements imposed on intron framing exonic sites may vary significantly, as well as the proportion of "canonic" sites and frequencies of intron phases.

ACKNOWLEDGEMENT

The authors are grateful to J-V. Chamary for purging the databases used in this study; as well as L.D. Hurst, N.J. Dibb and J. Nolan for the useful discussion.

REFERENCES

[1] Fedorov, A.; Suboch, G.; Bujakov, M.; Fedorova, L. Analysis of nonuniformity in intron phase distribution. *Nucleic Acids Res*., **1992**, *20*: 2553-2557.

[2] Long, M.; Deutsch, M. Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. *Mol. Biol. Evol*., **1999**, *16*: 1528-1534.

[3] Endo, T.; Fedorov, A.; de Souza, S.J.; Gilbert, W. Do introns favor or avoid regions of amino acid conservation? *Mol. Biol. Evol*., **2002**, *19*: 521-525.

[4] Gilbert, W.; de Souza SJ, Long, M. Origin of genes. *Proc. Natl. Acad. Sci. USA*, **1997**, *94*: 7698-7703.

[5] Lynch, M. Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. USA*, **2002**, *99*: 6118-6123.

[6] Ruvinsky, A.; Eskesen, S.T.; Eskesen, F.N.; Hurst, L.D. Can codon usage bias explain intron phase distributions and exon symmetry? *J. Mol. Evol*., **2005**, *60*: 99-104.

[7] Burset, M.; Seledtsov, I.A.; Solovyev, V.V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res*., **2000**, *28*: 4364-4375.

[8] Maniatis, T.; Tasic, B. Alternative splicing pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **2002**, *418*: 236-243.

[9] Saxonov, S.; Daizadeh, I.; Fedorov, A.; Gilbert, W. EID: the Exon-Intron Database – an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res*., **2000**, *28*: 185-190.

[10] Lercher, M.J.; Chamary, J.V.; Hurst, L.D. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Res*., **2004**, *14*: 1002-1013.

[11] Shannon, C.E. A mathematical theory of communication. *Bell System Tech. J*., **1948**, *27*: 379-423.

[12] Rogozin, I.B.; Sverdlov, A.V.; Babenko, V.N.; Koonin, E.V. Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief Bioinform*., **2005**, *6*: 118-134.

[13] Fedorov, A.; Cao, X.; Saxonov, S.; de Souza, S.J.; Roy, S.W.; Gilbert, W. Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. *Proc. Natl. Acad. Sci. USA*, **2001**, *98*: 13177–13182.

[14] Fichant, G.A. Constrains acting on the exon positions of the splice site sequence and local amino acid composition of the protein. *Hum. Mol. Genet*., **1992**, *1*: 259-267.

[15] Parmley, J.L.; Chamary, J.V.; Hurst, L.D. Evidence for purifying selection against mutations on mammalian exonic splicing enhancers. *Mol. Biol. Evol*., **2006**, *23*: 301-309.

[16] Roy, S.W.; Gilbert, W. Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc. Natl. Acad. Sci. USA*, **2005**, *102*: 5773-5778.

[17] Rogozin, I.B.; Lyons-Weiler, J.; Koonin, E.V. Intron sliding in conserved gene families. *Trends Genet*., **2000**, *16*: 430-432.

[18] Coglan, A.; Wolfe, K.H. Origins of recently gained introns in *Caenorhabditis*. *Proc. Natl. Acad. Sci. USA*, **2004**, *101*: 11362–11367.

Received: December 14, 2007 Revised: January 14, 2008 Accepted: January 15, 2008

© Ruvinsky and Ward; Licensee *Bentham Open*.

This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/license/by/2.5/), which permits unrestrictive use, distribution, and reproduction in any medium, provided the original work is properly cited.

SUPPLEMENTARY

Intron Framing Exonic Nucleotides: A Compromise Between Protein Coding and Splicing Requirements

A. Ruvinsky and W. Ward.

Table S1. Frequencies of Nucleotides at Three Exonic Positions Surrounding Introns*

		Frequencies**									χ^2 contingency comparisons***
		Phase 0			Phase 1			Phase 2			
		-2	-1	+1	-2	-1	+1	-2	-1	+1	
Homo sapiens	A	**0.663**	0.064	0.228	**0.509**	0.081	0.282	**0.613**	0.174	0.187	All comparisons of corresponding positions between Ph0 and Ph1, Ph0 and Ph2, Ph1 and Ph 2 have P<0.00001
	C	0.112	0.026	0.124	0.140	0.033	0.154	0.136	0.050	0.179	
	T	0.138	0.081	0.090	0.188	0.049	0.122	0.090	0.079	0.115	
	G	0.087	**0.829**	**0.558**	0.163	**0.837**	**0.442**	0.161	**0.697**	**0.519**	
Drosophila melanogaster	A	**0.573**	0.081	0.306	**0.373**	0.164	**0.346**	**0.526**	0.302	**0.313**	All comparisons of corresponding positions between Ph0 and Ph1, Ph0 and Ph2, Ph1 and Ph 2 have P<0.00001
	C	0.141	0.083	0.193	0.180	0.078	0.203	0.182	0.091	0.276	
	T	0.172	0.142	0.137	0.205	0.102	0.173	0.099	0.133	0.172	
	G	0.114	**0.694**	**0.364**	0.242	**0.656**	0.278	0.193	**0.474**	0.239	
Cearnohabditis elegans	A	**0.576**	0.149	**0.395**	**0.445**	0.211	**0.394**	**0.560**	0.300	**0.534**	All comparisons of corresponding positions between Ph0 and Ph1, Ph0 and Ph2, Ph1 and Ph 2 have P<0.00001
	C	0.165	0.044	0.144	0.139	0.084	0.181	0.145	0.108	0.149	
	T	0.173	0.197	0.112	0.316	0.115	0.163	0.128	0.156	0.152	
	G	0.086	**0.610**	0.349	0.100	**0.590**	0.262	0.167	**0.436**	0.165	
Arabidopsis thaliana	A	**0.684**	0.079	0.244	**0.521**	0.098	0.293	**0.611**	0.165	0.240	All comparisons of corresponding positions between Ph0 and Ph1, Ph0 and Ph2, Ph1 and Ph 2 have P<0.00001
	C	0.103	0.029	0.094	0.126	0.034	0.140	0.096	0.060	0.118	
	T	0.147	0.114	0.098	0.282	0.068	0.139	0.135	0.123	0.183	
	G	0.066	**0.778**	**0.564**	0.071	**0.800**	**0.428**	0.158	**0.652**	**0.459**	

*Positions -2;-1 (3' end of an exon preceding the intron) and +1 (5' end of an exon following the intron) represent locations of exonic nucleotides framing introns.
**The most frequent nucleotides at each position are shown in bold.
*** Observed numbers rather than the frequencies shown in the table were used for χ^2 contingency tests. Degree of freedom is always 3.