Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 219–223, January 1998
Evolution

Relationship between "proto-splice sites" and intron phases: Evidence from dicodon analysis

(intron–exon structure/dicodon usage)

MANYUAN LONG*§, SANDRO J. DE SOUZA*, CARL ROSENBERG†, AND WALTER GILBERT*‡

*Department of Molecular and Cellular Biology, The Biological Laboratories, Harvard University, Cambridge, MA 02138; and ‡Falling Rain Genomics, Inc., Lincoln, MA 01773

Contributed by Walter Gilbert, November 13, 1997

ABSTRACT The coding sequence at the boundaries of exons flanking nuclear introns shows some degree of conservation. To the extent that such sequences might be recognized by the splicing machinery, this conservation may be a derived result of evolution for efficient splicing. Alternatively, such conserved sequences might be remnants of proto-splice sites, which might have existed early in eukaryotic genes and served as the targets for the insertion of introns, as has been proposed by the introns-late theory. The distribution of intron phases, the position of the intron within a codon, is biased with an over-representation of phase 0 introns. Could any distribution of proto-splice sites account for today's intron phase distribution? Here, we examine the dicodon usage in six model organisms, based on current sequences in the GenBank database, and predict the phase distribution that would be expected if introns had been inserted into proto-splice sites. However, these predictions differ between the various model organisms and disagree with the observed intron phase distributions. Thus, we reject the hypothesis that introns are inserted into hypothetical proto-splice sites. Finally, we analyze the sequences around the splice sites of introns in all six of the species to show that the actual conservation of sequence in exon regions near introns is very small and differs considerably between these species, which is inconsistent with a general proto-splice sites model.

The significance of any conservation of DNA sequences near the exon–intron boundaries is an open question. Within the introns, there is very high conservation at and near the boundaries: The GT..AG rule is obeyed very well, with a minor exception, the AT..AC signal, in a small class of nuclear introns (1–2). Within the exon, various groups have conjectured that coding sequences near the boundaries also are conserved, such as the hypothesis of a (C/A)AG¦G conservation in mammalian genes and an (A/G)¦N conservation in *Saccharomyces cerevisiae* (the symbol "¦" stands for the intron positions) (3–4).

There are two alternative scenarios to account for the origin of conserved exon sequences. One is that the conserved sequence is a splicing signal at the exon boundary that has evolved as a result of natural selection for efficient splicing because the small nuclear RNAs in the splicing apparatus necessarily interact with some of the exon sequence. One clear case of such pairing has been identified in *S. cerevisiae* (5–6). Alternatively, conserved exon sequences might be remnants of early sequences in the coding regions that served as recognition sites for the insertion of introns, as proposed by Dibb and Newman (7), who called such consensus sequences "proto-splice" sites. Such proto-splice sites have been used as a conceptual basis for introns-late theories (8–10).

Some authors (11–13) have shown that the distribution of intron phases is significantly biased toward the phase 0 introns. Although these authors have argued that this biased distribution was most likely to be a consequence of exon shuffling, an alternative hypothesis would be that the biased intron-phase distribution is a consequence of intron insertion into nonrandomly distributed proto-splice sites.

In this study, we tested such models of proto-splice site insertion by examining their predictions for the intron-phase distribution. We will show that the distribution of hypothetical proto-splice sites in the genomes of six model organisms fails to explain the actual distribution of intron phases. Furthermore, taking the extensive sequence comparisons now available, we can show that most of the conserved information is confined within the intron and that the conserved information content within the exon is very small, and different in different organisms, suggesting only a fragile basis for any proto-splice site model.

METHODS

General Approach. To analyze the distribution of proto-splice sequences, we analyzed the dicodon distribution in the coding sequence of intron-containing genes. We examined the true dicodon distribution rather than simply using the codon frequencies because the correlation between adjacent codons may affect the distribution of proto-splice sites that cross codons. We calculated the correlation of dicodon frequencies by information analysis and calculated the distribution of various hypothetical proto-splice sites. We then compared such distributions with the actual intron phase distribution.

Finally, to assess the validity of a proto-splice site analysis, which is based on sequence conservation within exons, we determined the distribution of sequences of both exons and introns and calculated the information content of each position, summarized by the logo analysis of Schneider *et al*. (14).

Exon Databases and Calculation of Intron Phase Proportions. We chose six model species, which have many sequenced genes and contain representatives of the major eukaryotic lineages. These organisms are *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Arabidopsis thaliana.*

We used computing methods similar to those of Long *et al*. (12) to develop intron–exon databases of the six model organisms from GenBank database release 96. We deleted the highly redundant genes, such as the Ig superfamily in humans or the Adh (alcohol dehydrogenase) sequences created for population genetics study in *Drosophila*. We purged the CDS

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

‡To whom reprint requests should be addressed. e-mail: gilbert@chromo.harvard.edu.

§Present address: The University of Chicago, Department of Ecology and Evolution, 1101 East 57th Street, Chicago, IL 60737.

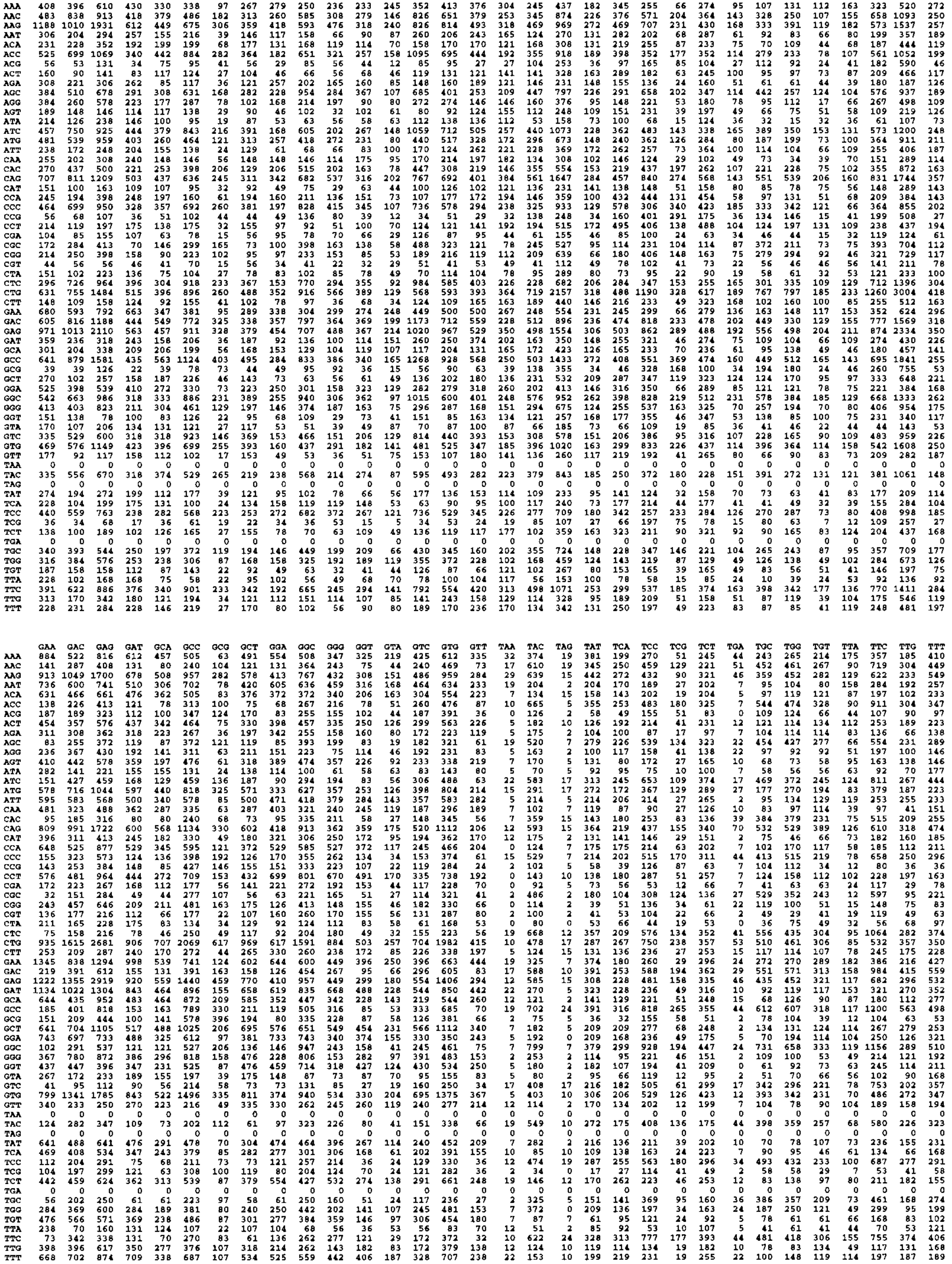
Table 1. Proportions of three intron phases

Species	Intron phases, %			Intron number	Gene number	*P*
	0	1	2			
A. thaliana	56	23	21	1342	300	2.93×10^{-68}
S. cerevisiae	39	35	26	163	160	0.11
S. pombe	45	29	26	796	364	1.46×10^{-11}
C. elegans	47	29	24	1096	195	1.23×10^{-21}
D. melanogaster	46	31	23	1145	385	4.73×10^{-21}
H. sapiens	44	36	20	7743	1301	1.99×10^{-151}

databases (a CDS is the composed coding sequence for an intron-containing gene) to a criterion of 80% to remove duplicates and closely related genes by using the GBPURGE program (12). (We also purged each database further to a criterion of 20% with GBPURGE and found that the proportions of intron phases and the other properties that we computed in various species did not change or changed insignificantly.) The *C. elegans* database contains many cosmid sequences that are analyzed by prediction by the GENEFINDER computer program. We purged these hypothetical genes from the *C. elegans* database for these calculations.

Proto-Splice Sites. Based on the previous analysis of conservation of coding sequences around the splicing sites (1, 4, 7), we chose four candidates for proto-splice sites: G¦G, AG¦G, AG¦GT, and (C/A)AG¦R (where the bar symbol indicates the site of the intron and R is purine A or G). [(C/A)AG¦R was proposed by Dibb and Newman (7).]

We used the information content measure of Schneider to evaluate the importance of any sequence conservation (14–



	AAA	AAC	AAG	AAT	ACA	ACC	ACG	ACT	AGA	AGC	AGG	AGT	ATA	ATC	ATG	ATT	CAA	CAC	CAG	CAT	CCA	CCC	CCG	CCT	CGA	CGC	CGG	CGT	CTA	CTC	CTG	CTT
AAA	408	396	610	430	330	338	97	267	279	250	236	233	245	352	413	376	304	245	437	182	345	255	66	274	95	107	131	112	163	323	520	272
AAC	483	838	913	418	379	486	182	313	260	585	308	279	146	826	651	379	253	345	874	226	376	571	204	364	143	328	250	107	155	658	1093	250
AAG	1188	1010	1931	612	449	675	306	359	418	593	476	318	240	826	814	493	318	469	969	272	469	707	231	430	168	333	391	119	182	573	1537	257
AAT	306	204	294	257	155	216	39	146	117	158	66	90	87	260	206	243	165	124	270	131	282	202	68	287	61	92	83	66	80	199	357	189
ACA	231	228	352	192	199	199	68	177	131	168	119	114	70	158	170	170	121	168	308	131	219	255	87	233	75	70	109	44	68	187	444	119
ACC	525	699	1069	340	442	884	282	364	182	651	321	257	158	1095	695	444	192	355	918	189	398	352	177	352	114	279	233	78	107	561	1052	199
ACG	56	53	131	34	75	95	41	56	29	85	56	44	12	85	95	27	27	104	253	36	97	165	85	104	27	112	92	24	41	182	590	46
ACT	160	90	141	83	117	124	27	104	46	66	56	68	46	119	131	121	141	141	328	163	289	182	63	245	100	95	97	73	87	209	466	117
AGA	308	221	306	262	85	117	36	121	257	202	165	160	85	148	160	189	121	146	231	148	155	136	24	160	51	61	61	44	39	180	187	126
AGC	384	510	678	291	308	631	168	282	228	954	284	367	107	685	401	253	209	447	797	226	291	658	202	347	114	442	257	124	104	576	937	189
AGG	384	260	578	223	177	287	78	102	168	214	197	90	80	272	274	146	146	160	376	95	148	221	53	180	78	95	112	17	66	267	498	109
AGT	189	148	146	114	117	138	29	90	46	102	32	102	61	80	92	124	155	112	248	109	151	231	39	197	49	66	75	51	58	109	219	126
ATA	214	126	238	146	100	95	19	87	53	63	56	58	63	112	138	136	112	53	158	73	100	68	15	124	36	32	15	32	36	61	107	73
ATC	457	750	925	444	379	843	216	391	168	605	202	267	148	1059	712	505	257	440	1073	228	362	483	143	338	165	389	350	153	131	573	1200	248
ATG	481	539	959	403	260	464	121	313	257	418	272	231	80	440	517	328	172	296	673	148	240	362	126	284	80	187	199	73	100	364	911	211
ATT	238	172	248	204	155	138	24	129	61	68	66	83	100	170	124	262	221	228	369	172	262	257	73	364	100	114	104	66	109	255	406	187
CAA	255	202	308	240	148	146	56	148	148	146	114	175	95	170	214	197	182	134	308	102	146	124	29	102	49	73	34	39	70	151	289	114
CAC	270	437	500	221	253	398	206	129	206	515	202	163	78	447	308	219	146	355	554	153	219	437	197	262	107	221	228	75	102	355	872	163
CAG	707	811	1209	503	437	636	245	311	342	682	537	316	202	767	692	401	384	561	1647	284	457	840	274	568	143	551	539	206	160	831	1744	357
CAT	151	100	163	109	107	95	32	92	49	75	29	63	44	100	126	102	121	136	231	141	138	148	51	158	80	85	78	75	56	148	289	143
CCA	245	194	398	248	197	160	61	194	160	211	136	151	73	107	177	172	194	146	359	100	432	444	131	454	58	97	131	51	68	209	384	143
CCC	464	699	950	328	357	692	260	381	197	828	415	345	107	736	578	294	238	325	933	129	578	306	340	423	185	333	342	121	66	364	855	202
CCG	56	68	107	36	51	102	44	44	49	136	80	39	12	34	51	29	32	138	248	34	160	401	291	175	36	134	146	15	41	199	508	27
CCT	214	119	197	175	138	175	32	155	97	92	51	100	70	124	121	141	192	194	515	172	495	406	138	488	104	124	197	131	109	238	437	194
CGA	104	85	155	107	63	78	15	56	95	78	70	66	29	126	87	95	44	61	155	46	85	100	24	63	34	46	44	15	32	119	124	61
CGC	172	284	413	70	146	299	165	73	100	398	163	138	58	488	323	121	78	245	527	95	114	231	104	114	87	372	211	73	75	393	704	112
CGG	214	250	398	158	90	223	102	95	97	233	153	85	53	189	216	119	112	209	639	66	180	406	148	163	75	279	294	92	46	321	729	117
CGT	44	56	56	46	41	70	15	56	34	41	22	32	29	51	41	53	49	41	112	49	78	102	41	73	22	56	46	46	56	141	211	78
CTA	151	102	223	136	75	104	27	78	83	102	85	78	49	70	114	104	78	95	289	80	73	95	22	90	19	58	61	32	53	121	233	100
CTC	296	726	964	396	304	918	233	367	153	770	294	355	92	984	585	403	226	228	682	206	284	347	153	255	165	301	335	109	129	712	1396	304
CTG	631	755	1484	515	396	896	260	488	352	916	566	389	129	568	593	393	364	719	2157	318	488	1190	328	617	189	767	797	185	233	1260	3004	418
CTT	148	109	158	124	92	155	41	102	78	97	36	68	34	124	109	165	163	189	440	146	216	233	49	323	168	102	160	100	85	255	512	163
GAA	680	593	792	663	347	381	95	289	338	304	299	274	248	449	500	500	267	248	554	231	245	299	66	279	136	163	148	117	153	352	624	296
GAC	605	816	1188	444	549	772	325	338	357	797	364	369	199	1173	712	559	228	512	896	236	474	818	233	478	202	449	330	129	155	777	1569	318
GAG	971	1013	2110	563	457	911	328	379	454	707	488	367	214	1020	967	529	350	498	1554	306	503	862	289	488	192	556	498	204	211	874	2334	350
GAT	359	236	318	243	158	206	36	187	92	136	100	114	151	260	250	374	202	163	350	148	255	321	46	274	75	109	104	66	109	274	430	226
GCA	301	204	338	209	206	199	56	168	153	129	104	119	107	117	204	131	165	172	423	126	165	233	70	236	61	95	138	49	46	180	457	141
GCC	641	879	1581	435	563	1124	403	495	284	833	386	340	165	1268	928	568	250	503	1433	272	408	551	369	474	160	449	512	165	143	695	1841	255
GCG	39	39	126	22	39	78	73	44	49	95	92	36	15	56	90	63	39	138	355	34	46	328	168	100	34	194	180	24	46	260	755	53
GCT	270	102	257	158	187	226	46	143	73	63	56	61	49	136	202	180	136	231	532	209	287	347	119	323	124	124	170	95	97	333	648	221
GGA	525	398	539	410	272	330	73	223	250	301	158	323	129	282	279	318	260	202	413	146	316	350	66	289	85	121	121	78	75	221	384	168
GGC	542	663	986	318	333	886	231	389	255	940	306	362	97	1015	600	401	248	576	952	262	398	828	219	512	231	578	384	185	129	668	1333	262
GGG	413	403	823	211	304	461	129	197	146	374	187	163	75	296	287	168	151	294	675	124	255	537	163	325	70	257	194	70	80	406	954	175
GGT	151	138	78	100	83	126	22	95	68	109	29	73	41	151	85	163	134	121	257	168	177	355	46	347	53	138	85	100	75	231	340	117
GTA	170	107	206	134	131	121	27	117	53	51	39	49	87	70	87	100	87	66	185	73	66	109	19	85	36	41	46	22	44	44	143	53
GTC	335	529	600	318	318	923	146	369	153	466	151	206	129	814	440	393	153	308	578	151	206	386	95	316	107	228	165	90	109	483	959	226
GTG	469	576	1149	423	396	699	255	393	160	437	291	182	141	481	525	347	185	396	1020	163	299	833	226	437	114	396	364	114	158	542	1608	250
GTT	177	92	117	158	112	102	17	153	49	53	36	51	75	153	107	180	141	136	260	117	219	192	41	265	80	66	90	83	73	209	282	187
TAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TAC	335	556	670	318	374	529	265	219	238	568	214	274	87	595	493	282	223	379	843	185	250	372	180	228	151	391	272	131	121	381	1061	148
TAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TAT	274	194	272	199	112	177	39	121	95	102	78	66	56	177	136	153	114	109	233	95	141	124	32	158	70	73	63	41	83	177	209	114
TCA	228	104	199	175	131	100	24	134	158	119	119	148	53	63	90	95	100	117	240	73	177	214	44	177	41	41	49	32	39	155	284	104
TCC	440	559	763	238	282	568	223	253	272	682	372	267	121	736	529	345	226	277	709	180	342	257	233	284	126	270	287	73	80	408	998	185
TCG	36	34	68	17	36	61	19	22	34	36	53	15	5	34	53	24	19	85	107	27	66	197	75	78	15	80	63	7	12	109	257	27
TCT	138	100	189	102	126	165	27	155	78	70	63	109	49	136	119	117	177	102	359	163	323	211	90	321	92	90	165	83	124	204	437	168
TGA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TGC	340	393	544	250	197	372	119	194	146	449	199	209	66	430	345	160	202	355	724	148	228	347	146	221	104	265	243	87	95	357	709	177
TGG	316	384	576	253	238	306	87	168	158	325	192	189	119	355	372	228	102	168	459	124	143	219	87	129	49	126	138	49	102	284	673	126
TGT	187	158	158	112	87	143	22	92	49	63	32	41	44	126	87	66	121	102	267	80	153	165	39	165	49	83	56	51	41	146	197	75
TTA	228	102	168	168	75	58	22	95	102	56	49	68	70	78	100	104	117	56	153	100	78	58	15	85	24	10	39	24	53	92	136	92
TTC	391	622	886	376	340	901	233	342	192	665	245	294	141	792	554	420	313	498	1071	253	299	537	185	374	163	398	342	177	136	770	1411	284
TTG	313	170	342	180	121	194	34	121	112	151	114	107	85	141	243	158	129	114	328	95	189	209	51	158	51	87	119	39	104	175	546	119
TTT	228	231	284	228	146	219	27	170	80	102	56	90	80	189	170	236	170	134	342	131	250	197	49	223	83	87	85	41	119	248	481	197

	GAA	GAC	GAG	GAT	GCA	GCC	GCG	GCT	GGA	GGC	GGG	GGT	GTA	GTC	GTG	GTT	TAA	TAC	TAG	TAT	TCA	TCC	TCG	TCT	TGA	TGC	TGG	TGT	TTA	TTC	TTG	TTT
AAA	884	522	816	612	457	505	63	491	554	508	347	325	219	425	612	335	32	374	19	381	199	270	51	245	44	243	265	214	175	357	185	410
AAC	141	287	408	131	80	240	104	121	131	364	243	75	44	240	469	73	17	610	19	345	250	459	129	221	51	452	461	267	90	719	304	449
AAG	913	1049	1700	678	508	957	282	578	413	767	432	308	151	486	959	284	29	639	15	442	272	432	90	321	46	359	452	282	129	622	233	549
AAT	736	600	741	510	306	702	78	420	605	636	459	316	168	464	634	233	19	204	2	204	170	189	27	202	7	95	104	80	158	284	192	257
ACA	631	466	661	476	362	505	83	376	372	372	340	206	163	304	554	223	7	134	15	158	143	202	19	204	5	97	119	121	87	197	102	233
ACC	138	226	413	121	78	313	100	75	68	267	216	78	51	260	476	87	10	665	5	355	253	483	180	325	7	544	474	328	90	911	304	347
ACG	187	189	323	112	100	347	124	170	83	255	155	102	44	187	391	36	0	126	2	58	49	155	51	83	0	109	124	66	44	107	90	97
ACT	454	357	576	437	342	464	75	330	398	457	335	250	126	299	563	226	5	182	10	126	192	214	41	231	12	121	114	134	112	253	189	223
AGA	311	308	362	318	223	267	36	197	342	255	158	160	80	172	223	119	5	175	2	104	100	87	17	97	7	104	114	114	83	136	66	138
AGC	83	255	372	119	87	372	121	119	85	393	199	83	19	182	321	61	19	520	7	279	226	539	134	323	22	454	427	277	66	554	231	289
AGG	236	367	430	192	141	311	63	211	151	223	75	114	46	192	231	83	5	163	2	100	117	158	41	138	22	97	92	92	51	197	100	146
AGT	410	442	578	359	197	476	61	318	389	474	357	226	92	233	338	219	7	170	5	131	80	172	27	165	10	68	73	58	95	163	138	146
ATA	282	141	221	155	155	131	24	138	114	100	61	58	63	83	143	80	5	70	5	92	95	75	10	100	7	58	56	56	63	92	70	177
ATC	151	427	459	168	129	459	136	187	90	294	194	83	56	306	488	63	22	583	17	313	245	653	109	374	17	469	372	245	124	811	267	444
ATG	578	716	1044	597	440	818	325	571	333	627	357	253	126	398	804	214	15	291	17	272	172	367	129	289	27	177	270	194	83	379	187	223
ATT	595	583	568	500	340	578	85	500	471	418	379	284	143	357	583	282	5	214	5	214	206	214	27	265	2	95	134	129	119	253	255	233
CAA	481	323	488	362	287	335	63	287	403	321	240	245	119	187	296	189	7	102	7	119	87	90	27	126	10	83	97	114	39	97	41	151
CAC	95	185	316	80	80	240	68	73	95	335	211	58	27	148	345	56	7	359	15	143	180	253	83	136	39	384	379	231	75	515	209	255
CAG	809	991	1722	600	568	1134	330	602	418	913	362	359	175	520	1112	206	12	593	15	364	219	437	155	340	70	532	529	389	126	610	318	474
CAT	396	311	413	245	182	330	49	180	321	306	250	172	95	194	362	170	12	175	2	131	141	146	29	151	2	75	46	66	73	182	160	185
CCA	648	525	877	529	345	595	121	372	529	585	527	372	117	245	466	204	0	124	7	175	175	214	63	202	7	102	170	117	58	185	112	211
CCC	155	323	573	124	136	398	192	126	170	355	262	134	34	153	374	61	15	529	7	214	202	515	170	311	44	413	515	219	78	658	250	296
CCG	143	253	384	148	85	427	146	155	151	333	223	107	22	119	284	24	2	102	5	58	39	126	87	63	7	104	112	34	12	80	36	36
CCT	576	481	964	444	272	709	153	432	699	801	670	491	170	335	738	192	0	143	10	138	180	287	51	257	7	124	158	112	102	228	197	163
CGA	172	223	267	168	112	177	56	141	221	272	192	153	44	117	228	70	0	92	5	73	56	53	12	66	7	41	63	63	24	117	29	78
CGC	32	151	284	49	44	277	107	56	63	221	165	51	27	114	321	41	2	486	2	180	104	308	124	136	27	529	352	243	12	597	95	221
CGG	243	457	646	209	211	481	163	175	126	413	148	155	46	182	330	66	0	114	2	39	51	136	34	61	22	119	100	51	15	148	75	83
CGT	136	177	216	112	66	177	22	107	160	260	170	155	56	131	287	80	2	100	2	41	53	104	22	66	2	49	29	41	19	119	49	63
CTA	211	165	228	175	83	134	34	129	92	124	112	83	58	61	168	53	0	80	0	53	66	44	19	53	0	36	75	49	32	56	68	97
CTC	75	158	216	78	46	250	49	117	92	204	180	49	32	155	223	56	19	668	12	357	209	576	134	352	41	556	435	304	95	1064	282	374
CTG	935	1615	2681	906	707	2069	617	969	617	1591	884	503	257	704	1982	415	10	478	17	287	267	750	238	357	53	510	461	306	85	532	357	350
CTT	253	209	287	240	170	272	44	265	330	260	238	172	85	226	338	197	5	124	15	131	136	236	27	253	15	117	114	107	78	245	175	228
GAA	1345	838	1294	998	539	741	124	602	644	600	449	396	250	396	663	444	19	325	7	374	180	260	29	296	24	272	270	289	182	386	216	427
GAC	219	391	612	155	131	391	163	158	126	454	267	95	66	296	605	83	17	588	10	391	253	588	194	362	29	551	571	313	158	984	415	559
GAG	1222	1355	2919	920	559	1440	459	770	410	957	449	299	180	554	1406	294	12	585	15	308	228	481	158	335	46	435	452	321	117	682	296	532
GAT	1134	1022	1304	843	464	896	155	658	619	835	668	488	228	544	850	442	22	270	5	323	228	236	49	316	10	92	119	117	153	321	270	352
GCA	644	435	952	483	464	872	209	585	352	447	342	228	143	219	544	260	12	121	2	141	129	221	51	248	15	68	126	90	87	180	112	277
GCC	185	401	818	153	163	789	330	211	119	505	316	85	53	333	685	70	19	702	24	391	316	818	265	355	44	612	607	318	117	1200	563	498
GCG	151	209	444	100	141	578	396	194	80	335	228	87	58	126	381	66	2	75	5	36	32	155	58	51	2	78	104	39	12	104	63	53
GCT	641	704	1105	517	488	1025	206	695	576	651	549	454	231	566	1112	340	7	182	5	209	209	277	68	248	2	134	131	124	114	267	279	253
GGA	743	697	733	488	325	612	97	381	733	743	340	374	155	330	350	243	5	192	0	209	168	236	49	175	5	70	194	114	104	250	126	321
GGC	102	291	537	121	121	527	206	136	146	947	243	158	41	245	461	75	7	799	7	379	299	928	194	447	24	731	658	333	119	1156	289	510
GGG	367	780	872	386	296	818	158	476	228	806	153	282	97	391	483	153	2	253	2	114	95	221	46	151	2	109	100	53	49	214	121	192
GGT	437	447	396	347	231	525	87	476	459	714	318	427	124	430	534	250	5	180	2	182	107	194	41	209	0	61	92	73	63	245	114	211
GTA	267	172	233	189	155	197	39	175	148	87	73	87	70	95	155	83	5	80	2	95	66	119	12	95	2	51	70	66	56	102	90	168
GTC	41	95	112	90	56	214	58	73	73	131	85	27	19	160	250	34	17	408	17	216	182	505	61	299	17	342	296	221	78	753	202	357
GTG	799	1341	1785	843	522	1496	335	811	374	940	534	330	204	695	1375	367	5	403	10	306	206	529	126	423	12	393	342	231	70	486	272	347
GTT	340	233	250	270	223	216	49	335	330	262	245	260	119	240	277	214	12	114	2	170	134	202	12	199	7	104	78	90	104	189	158	194
TAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TAC	124	282	347	109	73	202	112	61	97	323	226	80	41	151	338	66	19	549	10	272	175	408	136	175	44	398	359	257	68	580	226	323
TAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TAT	641	488	641	476	291	478	70	304	474	464	396	267	114	240	452	209	7	282	2	216	136	211	39	202	10	70	78	107	73	236	155	231
TCA	469	408	534	347	243	379	85	282	277	301	306	168	61	202	391	155	10	85	10	109	138	163	24	223	7	90	95	46	61	134	66	168
TCC	112	204	291	75	68	211	73	73	121	257	214	36	24	129	330	36	12	474	19	287	255	563	180	296	34	493	432	233	100	687	277	291
TCG	104	197	299	121	63	308	100	119	80	204	124	70	24	121	282	36	2	34	0	17	27	114	41	49	2	58	58	29	7	53	41	58
TCT	442	459	624	362	313	539	87	379	554	427	532	274	138	291	661	248	19	146	12	170	262	223	46	253	12	83	138	97	80	211	182	155
TGA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TGC	56	202	250	61	61	223	97	58	61	250	160	51	24	117	236	27	2	325	5	151	141	369	95	160	36	386	357	209	73	461	168	274
TGG	284	369	600	284	189	381	80	240	250	442	202	141	107	245	481	153	7	372	0	209	136	197	34	163	24	187	250	121	49	299	95	199
TGT	476	566	571	369	238	486	87	301	277	384	359	146	97	306	454	180	7	87	7	61	95	121	24	92	5	78	61	61	66	168	83	102
TTA	238	70	160	131	124	107	22	107	104	68	56	36	53	56	83	70	12	51	2	85	92	53	10	107	5	41	61	41	44	70	53	121
TTC	73	342	338	131	70	270	83	61	136	262	277	121	29	172	372	32	10	622	24	328	313	777	177	393	44	481	418	306	155	755	374	406
TTG	398	396	617	350	277	376	107	318	214	262	143	182	83	172	379	138	12	124	10	119	114	134	19	182	10	78	83	134	49	117	131	168
TTT	668	702	874	709	338	687	107	534	525	559	442	406	187	328	707	238	22	153	10	199	219	231	19	255	22	100	148	119	114	197	187	189

FIG. 1. Human dicodon table (10^6).

Table 2. Dicodon correlation (I)

A. thaliana	0.13
S. serevisiae	0.13
S. pombe	0.07
C. elegans	0.13
D. melanogaster	0.14
H. sapiens	0.14

15). The amount of information at each nucleotide site *i* is calculated by

$$Rs(i) = 2 - (H(i) + e(n))$$

where $H(i) = -\Sigma f_j(i)\log_2 f_j(i)$, $f_j(i)$ is the frequency of the base *j* at position *i*. Here $e(n) = 3/(2ln(2)n)$ is a correction for sample size (approximate calculation), where *n* is the number of introns. When there is no conservation, $Rs(i) = 0$; when frequency of a single base reaches 100% (maximum conservation), $Rs(i) = 2$.

Dicodon Correlation. We generated CDS databases for the six model species. We calculated the frequencies of the 64 × 64 dicodons from the CDS databases. We analyzed the frequencies of the dicodon types (stop codon)$N_4N_5N_6$ as a control for two possible errors: errors caused by any irregularity in the feature tables in GenBank from which the CDS databases were developed and errors arising by the inclusion of pseudogenes. These errors lead to non-zero frequencies of the dicodons of these types and were all removed.

We analyzed the information content in the dicodon sequences (in-frame hexamers) (16) by using the formula

$$I = \sum_{i=1}^{64}\sum_{j=1}^{64} P_{ij}\log_2\left(\frac{P_{ij}}{P_iP_j}\right)$$

where *Pij* is the frequency of the dicodon *i* and *j*; *Pi* and *Pj* are codon frequencies of *i* and *j*. When all codon pairs *i* and *j* are completely independent, $Pij = Pi.Pj$, and the information *I* is equal to 0. On the contrary, if one codon (*i* or *j*) completely determines the other (*j* or *i*), the information *I* will reach a maximum of six.

We then wrote a computer program to scan every dicodon for proto-splice sites in each phase. To avoid any repeated counting of proto-splice sites that have length equal to or shorter than 3, we counted the dicodons that contained the site only at the 5′ codons and the site across the codon. (Counting 3′ codons yields the same results.) If a dicodon sequence contains more than one proto-splice site, we counted its frequency for each site.

RESULTS

The Six Model Species Show Different Patterns of Intron Phases. Table 1 shows the distribution of intron phases for the six species. All of the species, except yeast because of a small sample size, showed a significant deviation from an equal–probable distribution (one-third) of intron phases and a preference for phase 0 introns. Furthermore, the species differed. Phase 0 introns ranged from 56% in *Arabidopsis* to 39% in *S. cerevisiae*, and phase 1 introns ranged from 36% in humans to 23% in *Arabidopsis*. This variation in the proportions of intron phases was not consistent with the model that proto-splice sites were used for targeting in the early stages of eukaryote evolution because this model would predict similar distributions of intron phase across the eukaryotic organisms. However, one might assume that the frequencies of proto-splice sites could have evolved differently in the different lineages. Can the distribution of proto-splice sites explain these peculiar distributions of intron phases?

Prediction of Intron Phases Based on Proto-Splice Sites and Dicodon Usage. Fig. 1 shows, as an example, the dicodon frequency of human genes by using the 80% purged database. (The dicodon frequencies of the other five species are available on request.) First, we analyzed the correlation between codons by calculating an information content by using a measure that ranges from 0 to 6. Table 2 shows that there were correlations between adjacent codons. However, the correlations differed by up to 2-fold. *D. melanogaster* and humans had the highest correlations ($I = 0.14$), and *S. pombe* had the lowest ($I = 0.07$).

We then calculated the expected frequencies of intron phase that would be determined by four hypothetical proto-splice sites [G¦G, AG¦G, AG¦GT, and (C/A)AG¦R] by using the dicodon frequencies for the six species. Table 3 lists these different patterns and compares them to the observed phases; the *P* values are given below each dicodon for all species.

The predicted intron phase frequencies were not consistent with the observed proportions of intron phases for many of the species. For example, for *A. thaliana*, no proto-splice sites or combinations of sites gave intron phase frequencies close to observation because the observed fraction of phase 0 introns was too high. Among 24 comparisons, only the *S. cerevisiae* phase pattern was very similar to the (C/A)AG¦R one; we think this single case is a random match to the 163 introns of *S. cerevisiae*. All of the other comparisons showed a significant difference between the observed and expected proportions in χ^2 tests.

Sequence Conservation at the Exon Side of the Splice Sites. We analyzed the distribution of 10 bases on both the intron and exon sides of each splice junction in our databases. The logos in Fig. 2 show the sequence conservation at the splice sites for the six species. Here, the total height of the stacked letters at each position is the total amount of information at that position, and the heights of individual letters reflect the proportion of the nucleotides.

Table 3. Intron phase distribution (%) predicted from dicodon frequency

Proto-splice site	G/G			AG/G			AG/GT			(C/A)AG/R			Observed		
intron phase	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
A. thaliana	36	43	21	41	37	22	36	48	16	41	42	17	56	23	21
P values		3×10^{-60}			3×10^{-32}			1×10^{-75}			2×10^{-44}				
S. cerevisiae	40	42	18	48	34	18	31	59	10	39	36	25	39	35	26
P values		0.02			0.01			6×10^{-14}			0.95				
S. pombe	40	41	19	47	36	17	35	52	13	32	52	16	45	29	26
P values		2×10^{-12}			2×10^{-11}			1×10^{-45}			3×10^{-38}				
C. elegans	38	50	12	51	44	5	57	38	5	42	45	13	47	29	24
P values		2×10^{-55}			7×10^{-185}			10×10^{-182}			8×10^{-38}				
D. melanogaster	55	32	13	75	14	12	77	16	7	71	19	10	46	31	23
P values		1×10^{-23}			6×10^{-105}			1×10^{-157}			2×10^{-83}				
H. sapiens	46	32	23	57	26	17	59	27	14	50	35	15	44	36	20
P values		3×10^{-17}			4×10^{-124}			2×10^{-158}			2×10^{-41}				

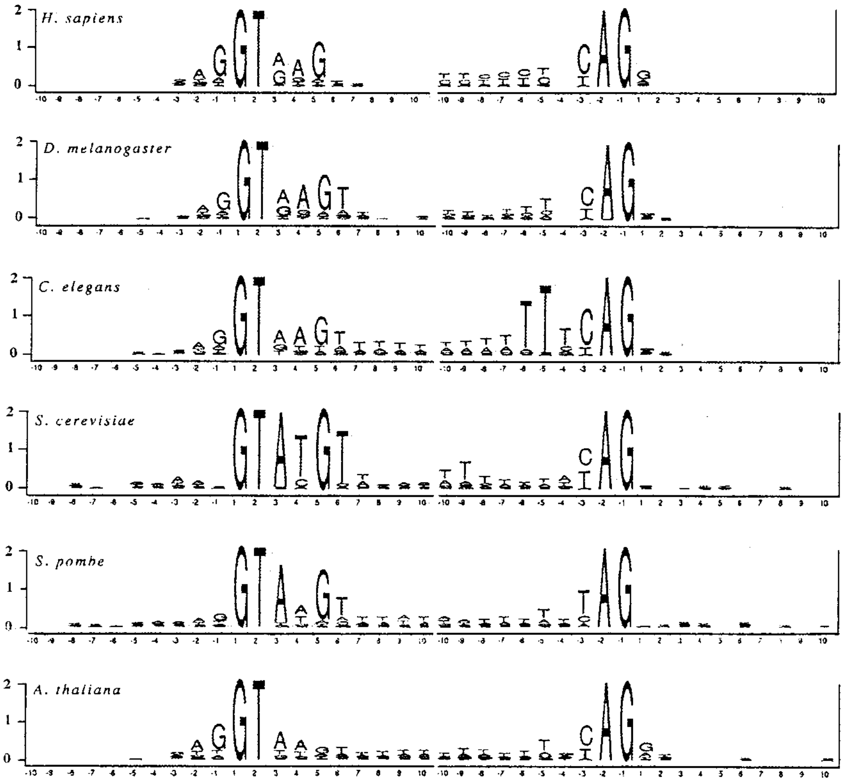


FIG. 2. Information content at each position in the 10 bases flanking the exon intron boundary and the intron–exon boundary. The total height at each position is given by the information content at that position. The height of each letter is proportional to the fraction that base is of the total at each position.

Fig. 2 shows that the information content is very uneven between exons and introns in all six species. The amount of information in the exons is very small; more than 90% of the information is contained within the intron. For the limited conservation within the exon sequences in *S. cerevisiae*, Long *et al.* (5) have argued that conservation may reflect a molecular role of pairing with the U5 small nuclear RNA rather than a signal for the insertion of introns.

The consensus sequences in the exon regions flanking introns varied among the six species. Taking at least 40% of the total to be the criterion for a consensus nucleotide, we found the sequence $A_{60}G_{75}$¦$G_{54}T_{42}$ for *A. Arabidopsis*; $A_{45}A_{55}A_{45}N$¦N for *S. cerevisiae*; $A_{42}A_{44}A_{47}G_{55}$¦N for *S. pombe*; $A_{54}G_{70}$¦G_{41} for *D. melanogaster*; $A_{40}A_{54}G_{65}$¦N for *C. elegans*; and $A_{61}G_{81}$¦G_{56} for *H. sapiens*. (The subscripts represent percentages of the consensus nucleotides.)

The limited and variable conservation in the exon sequences of different organisms suggests differential local requirements for the splicing processes and does not support the conception of a proto-splice site sequence preexisting in ancestor mRNAs that did not contain introns.

DISCUSSION

By analyzing dicodon frequencies from six model species (*A. thaliana*, *S. cerevisiae*, *S. pombe*, *D. melanogaster*, *C. elegans*, and *H. sapiens*), we have shown that the four candidate hypothetical proto-splice sites [G¦G, AG¦G, AG¦GT, and (C/A)AG¦R] cannot explain the actual intron phase distributions across all the species. For G¦G, AG¦G, and AG¦GT sites, there were no similar patterns in the pattern of the three phases in the data. The (C/A)AG¦R in *S. cerevisiae* and *H. sapiens* showed patterns similar to the intron phase distribution, but the actual differences between the prediction and the expectation were very significant for *H. sapiens*, for which there were a lot of data. For *S. cerevisiae*, for which there were only 163 introns in our purged database drawn from the entire genome, we think this coincidence was not significant but was a random match.

In general, the predictions of the proto-splice models change drastically with the species. This model, the most general model of the introns-late theory, fails to account for the uneven distribution of intron phases.

We extended the information content conservation analysis of Stephens and Schneider (15) to these six model species and observed only a very low information content in the coding regions that flank the splice sites. This finding weakens the argument for proto-splice sites.

To date, the only clear examples of the insertion of introns are the spliceosomal introns in the U2 and U6 small nuclear RNA genes in certain yeast species (17–19). However, the coding sequences flanking these introns are random (20). Hence, the example of real insertion of introns does not suggest any proto-splice sites.

A direct model that fits the data of intron phase distribution is the introns-early theory (21–23). The excess of phase 0 introns would be a consequence of the mini-gene nature of primordial exons and the use of exon shuffling. Moreover, the excess of symmetric exons in modern as well as ancient conserved genes supports an important role for exon shuffling both later and also in the early evolution of genes before the divergence of prokaryotes and eukaryotes.

We thank T. Schneider for valuable discussions on information analysis and S. Mount for the discussion about the relationship between splicing signals and intron phases. S.J.deS. was supported by the PEW–Latin American Program and Fundacao de Amparo a Pesquisa do Estado de Sao Paulo.

1. Mount, S. M. (1982) *Nucleic Acids Res.* **10,** 459–472.
2. Hall, S. L. & Padgett, R. A. (1996) *Science* **271,** 1690–1691.

3. Horowitz, D. S. & Krainer, A. R. (1994) *Trends Genet.* **10,** 100–106.
4. Csank, C., Taylor, F. M. & Martindale, D. W. (1990) *Nucleic Acids Res.* **18,** 5133–5141.
5. Long, M., de Souza, S. J. & Gilbert, W. (1997) *Cell*, in press.
6. Newman, A. J. & Norman, C. (1992) *Cell* **68,** 743–754.
7. Dibb, N. J. & Newman, A. J. (1989) *EMBO J.* **8,** 2015–2022.
8. Palmer, J. D. & Logsdon, J. M. (1991) *Curr. Opin. Genet. Dev.* **1,** 470–477.
9. Logsdon, J. M., Jr., Tyshenko, M. G., Dixon, C., Jafari, J. D., Walker, V. K. & Palmer, J. D. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 8507–8511.
10. Long, M. & Stoltzfus, A. (1997) *HMS Beagle: A BioMedNet Publication*. Issue 1 (Feb. 1). Available at http://hmsbeagle.com. Accessed November 24, 1997.
11. Fedorov, A., Suboch, G., Bujakov, M. & Fedorova, L. (1992) *Nucleic Acids Res.* **20,** 2553–2557.
12. Long, M., Rosenberg, C. & Gilbert, W. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 12495–12499.
13. Tomita, M., Shimuzu, N. & Brutlag, S. (1996) *Mol. Biol. Evol.* **13,** 11–15.
14. Schneider, T. D., Stormo., G. D., Gold. L. & Ehrenfeucht, A. (1986) *J. Mol. Biol.* **188,** 415–431.
15. Stephens, R. M. & Schneider, T. D. (1992) *J. Mol. Biol.* **228,** 1124–1136.
16. Farber, R., Lapeds, A. & Sirotkin, K. (1992) *J. Mol. Biol.* **226,** 471–479.
17. Takahashi, Y., Urushiyama, S., Tani, T. & Ohshima, Y. (1991) *Mol. Cell. Biol.* **5,** 1022–1231.
18. Tani, T. & Ohshima, Y. (1991) *Genes Dev.* **5,** 1022–1031.
19. Tani, T. & Ohshima, Y. (1989) *Nature (London)* **337,** 87–90.
20. Long, M., de Souza, S. J. & Gilbert, W. (1995) *Curr. Opin. Genet. Dev.* **5,** 774–778.
21. Doolittle, W. F. (1978) *Nature (London)* **272,** 581–582.
22. Gilbert, W. (1978) *Nature (London)* **271,** 501.
23. Gilbert, W. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52,** 901–905.