J Mol Evol (2005) 61:341–350
DOI: 10.1007/s00239-004-0318-y

JOURNAL OF MOLECULAR EVOLUTION

Signs of Ancient and Modern Exon-Shuffling Are Correlated to the Distribution of Ancient and Modern Domains Along Proteins

Maria Dulcetti Vibranovski,[1,2] Noboru Jo Sakabe,[1,2] Rodrigo Soares de Oliveira,[1] Sandro José de Souza[1]

[1] Ludwig Institute for Cancer Research, Sao Paulo Branch, Rua Prof. Antonio Prudente 109 4° andar, CEP 01509-010, São Paulo, SP, Brazil
[2] Ph.D Program, Departmento de Bioquímica, Universidade de São Paulo, Av. Prof. Lineu Prestes, 748, Bloco 03 superior, sala 351, CEP: 05508-900, Cidade Universitária, São Paulo, SP, Brazil

Received: 10 November 2004 / Accepted: 11 March 2005

Abstract. Exon-shuffling is an important mechanism accounting for the origin of many new proteins in eukaryotes. However, its role in the creation of proteins in the ancestor of prokaryotes and eukaryotes is still debatable. Excess of symmetric exons is thought to represent evidence for exon-shuffling since the exchange of exons flanked by introns of the same phase does not disrupt the reading frame of the host gene. In this report, we found that there is a significant correlation between symmetric units of shuffling and the age of protein domains. Ancient domains, present in both prokaryotes and eukaryotes, are more frequently bounded by phase 0 introns and their distribution is biased towards the central part of proteins. Modern domains are more frequently bounded by phase 1 introns and are present predominantly at the ends of proteins. We propose a model in which shuffling of ancient domains mainly flanked by phase 0 introns was important in the ancestor of eukaryotes and prokaryotes, during the creation of the central part of proteins. Shuffling of modern domains, predominantly flanked by phase 1 introns, accounted for the origin of the extremities of proteins during eukaryotic evolution.

Key words: Exon-shuffling — Introns-early — Introns-late — Protein domains

Correspondence to: Sandro José de Souza; *email:* sandro@compbio.ludwig.org.br

Introduction

Several studies in the last decade have allowed the emergence of a synthetic theory of intron evolution (de Souza et al. 1998; de Souza 2003; Roy et al. 1999; Roy 2003). This theory represents an attempt to reconcile different aspects of two divergent theories on intron evolution; the ''introns-early'' (Blake 1978; Doolittle 1978; Gilbert 1978, 1987) and the ''introns-late'' theories (Cavalier-Smith 1991; Palmer and Logsdon 1991; Stoltzfus et al. 1994). The major focus of disagreement between these two theories is the date of origin of the first introns, before or after the divergence of eukaryotes and prokaryotes. As a consequence, ''introns-early'' advocates postulate that exon-shuffling (intron-mediated recombination between unrelated genes) contributed to the building of proteins in the progenote (the ancestor of prokaryotes and eukaryotes) while this process would be restricted to eukaryotes according to proponents of the ''introns-late'' theory. The synthetic theory of intron evolution attempts to reconcile these opposing views by postulating that a fraction of present-day phase 0 introns (those present between codons) antedates the eukaryote/prokaryote divergence while the remaining introns are acquisitions by the eukaryotes (de Souza et al. 1998; de Souza 2003; Roy et al. 1999; Roy 2003). Support for this theory comes from the following observations: (1) the higher frequency of phase 0 introns in almost all eukaryotes (Fedorov et al. 1992; Long et al. 1995); (2) the even

higher frequency of phase 0 introns in genes whose origins antedate the divergence of eukaryotes and prokaryotes (Long et al. 1995); (3) the correlation between phase 0 introns and the boundaries of protein modules in a set of ancient proteins (de Souza et al 1998; Fedorov et al. 2001); (4) the even stronger correlation between phase 0 introns that are conserved in several kingdoms and the boundaries of protein modules (Fedorov et al. 2003; Roy et al. 1999); and (5) the growing amount of evidence indicating intron insertion as a frequent phenomenon in the eukaryotic evolution (Palmer and Logsdon 1991; Stoltzfus et al. 1997). Evidence for ancient exon-shuffling (occurring in the progenote) comes from the above correlation between phase 0 introns and the boundaries of modules in ancient proteins. More important, however, is the excess of symmetric exons (flanked by introns of same phase) shown to exist in a set of ancient proteins (de Souza et al. 1998; Long et al. 1995). An excess of symmetric exons is evidence for exon-shuffling because of the deleterious effect of the shuffling of genetic units with a length not a multiple of 3 nucleotides, which could cause a frame-shift in the host gene (Patthy 1987, 1991).

Recently, Kaessmann et al. (2002) investigated the role of exon-shuffling in the construction of human proteins. They found that introns present at the boundaries of protein domains show a highly significant excess of symmetrical combinations. Internal introns present in the same domains do not show the same pattern. More recently, a correspondence between protein domains and exon boundaries was found for other organisms as well (Liu and Grigoriev 2004). The most important observation from Kaessmann et al. (2002) regarding the evolution of introns is an overrepresentation of 0-0 symmetrical domains in a set of ancient domains (domains present in both eukaryotic and prokaryotic sequences). Furthermore, they also confirmed that 1-1 symmetrical domains are significantly overrepresented in modern domains, an observation made earlier by Patthy (1991, 1999, 2003). Although the excess of 0-0 ancient domains was not statistically significant, it was suggestive of primordial domain shuffling in the progenote.

The data from Kaessmann et al. (2002), together with all the remaining evidence for the synthetic theory of intron evolution discussed above, motivated us to further explore the distribution of ancient and modern domains along proteins. We found that the excess of symmetric domains was not homogeneously distributed and that it was correlated to the age of the domains. Ancient domains flanked by phase 0 introns were predominantly located at the central region of proteins, while 1-1 modern domains occurred more frequently at the extremities of proteins. Based on this, we propose a model that describes the contribution of ancient and modern exon-shuffling to the construction of new proteins. According to this model, shuffling of domains flanked by phase 0 introns predominated in the progenote and was important during the evolution of the core of proteins, while recent domain shuffling, involving predominantly phase 1 introns, occurred at the extremities of proteins. This model represents a theoretical advancement regarding the evolution of exon-shuffling and gives further support to both the synthetic theory of intron evolution and to the ancient nature of a fraction of present-day phase 0 introns.

Materials and Methods

Databases

In order to analyze the position of introns in relation to proteins, we used the ExInt database release 133 (Sakharkar et al. 2002). This database is composed of 134,442 sequences from GenBank containing annotated introns with their position and phase indexed with respect to the amino acid sequences. Filtered databases were generated from ExInt to eliminate GenBank redundancy as previously explained (Long et al. 1995). In short, all sequences were aligned to each other using FASTA (Pearson and Lipman 1988). Then, we calculated the identity of all pairwise alignments by multiplying the alignment identity (match percentage) by its length and dividing this product by the size of the shortest sequence. Sequences were merged into single clusters based on the degree of their identity, as calculated by the procedure described above (here called recalculated identity). Each filtered database was generated by selecting one representative sequence (the one with most exons) from each cluster. Two databases were generated: one with sequences presenting 99% or more of recalculated identity, used to eliminate duplicate and partial sequences yielding 99,583 proteins, and another with sequences presenting 20% or more of recalculated identity, used to eliminate homologous sequences yielding 27,708 proteins.

To select ancient proteins, the proportion of the proteins similar to prokaryotic genes was assessed by aligning amino acids sequences from ExInt to the bacterial sequences from the SWISS-PROT database (release 41; Boeckmann et al. 2003) using Blastp with *E*-value threshold $\leq 10^{-4}$ (Altschul et al. 1997). In addition, 8755 sequences from ExInt homologous to 276 ancient proteins previously used by Fedorov et al. (2001) were selected (using the same criteria as above) as an independent set of ancient proteins.

Protein domains were selected from Pfam (Bateman et al. 2002) and were grouped by age using taxonomic information following criteria similar to Kaessmann et al. (2002). We considered domains as ancient when they were shared by eukaryotes and prokaryotes whereas modern domains were defined as those exclusive to eukaryotes.

Frequency of Intron Phase and Symmetric Exons

All protein sequences from ExInt were analyzed with respect to their intron phase distribution and symmetric and non-symmetric exons as previously described (Long et al. 1995). The three classes of intron phase were counted, their distributions obtained, and the differences were statistically tested by chi-squared tests. The nine classes of exons (with respect to their intron phase correlation) were counted and their distributions were analyzed in comparison to the

expected values, calculated based on the frequency of their flanking introns.

Protein Regions for Analysis of Symmetric Exons

The distributions of intron phases and symmetric exons were analyzed along the sequences corresponding to protein domains. Every sequence was divided in three equal parts (0–33, 34–66, 67–100%) where exon frequencies were analyzed separately. Exons were only counted if both their flanking introns were located inside the protein region analyzed. Since exon-shuffling events may involve more than one protein domain, we decided to count as a single domain region those domains that were separated by either five or fewer amino acids or by less than 10% of the protein's length (but not by more than 50 amino acids). This procedure was only used for this analysis and allowed the identification of exons encompassing more than one domain.

Domain Distribution

The domain content of protein sequences was analyzed using Pfam 12 (Bateman et al. 2002). The positions in the protein of the domain hits with E-value ≤ 0.1 were annotated. The positions in amino acid coordinates were converted to percentages of the protein length. The plots shown in Figs. 1, 2, and 4 present the summed relative frequency of each percent position of all proteins in a given dataset, as a function of protein length (1–100%).

Simulation of Exon Distribution

Simulations were performed to verify if the difference in the distribution of the excess of symmetric exons between the different protein regions was significant. Five thousand iterations for each analysis were performed in which only the distribution of exons varied. The number of sequences, number of introns, their positions, the frequency of intron phases, the frequency of the different types of exons (symmetric and non-symmetric), and the regions where ancient and modern domains were located in the simulated data were the same as in the observed data. The instances in which the distributions of excess of symmetric exons were similar to the observed distributions were counted. Distributions were considered similar when they presented an excess of exons 1-1 at the extremities of the proteins at least 3.5% higher than the excess observed in the central part and an excess of exons 0-0 in the central part of the protein 6.0% higher than the excess observed at the extremities. These values correspond to half of the smallest differences between those regions in the real data.

Analysis of Domains Flanked by Introns

To characterize the domains that corresponded to exons or to sets of exons in the ExInt dataset, we identified domains that were bounded by introns. We used the criteria developed by Kaessmann et al. (2002) to determine the relationship between domain boundaries and the position of flanking introns: $|\,dn - in\,| + |\,dc - ic\,| < 0.1 \cdot (dc - dn)$. dn and dc represent the amino and carboxy terminal coordinates of the domain, respectively, whereas in and ic are the analogous coordinates of the flanking introns. To define a domain as flanked by introns, the sum of the differences between the domain boundaries and the positions of flanking introns had to be less than 10% of the domain length. If there were more than one intron located in the flanking region, we chose the closest one.

Observed and expected numbers of symmetric domains (0-0 and 1-1) were compared to assess the deviation of their distribution in ancient and modern domains. Expected values were calculated based on the frequencies of 0-0 and 1-1 domains using both datasets (ancient and modern) together (Kaessmann et al. 2002). Simulations of 10,000 iterations were performed to evaluate significant deviations. Each iteration selects 0-0 and 1-1 domains with frequencies calculated independently of the age of the domain. Cases where the number of 0-0 ancient and 1-1 modern domains were equal or higher as well as the number of 0-0 modern and 1-1 ancient domains were equal or higher than the observed value were counted to measure the significance of the data.

Results and Discussion

The Distribution of Ancient and Modern Domains Along Proteins Depends on Their Age

Domains were classified as either ancient or modern according to Pfam's taxonomy (see Materials and Methods). Ancient and modern domains can be present in proteins that contain only either ancient or modern domains (here called "simple" proteins) and in proteins with both ancient and modern domains (here called "hybrid" proteins). Figure 1 shows the distribution of ancient and modern domains along "simple" and "hybrid" proteins in the ExInt database. For the size of the different datasets, see Table 1. The distribution of these domains was not homogeneous along proteins. Ancient and modern domains presented the same central distribution in the dataset of "simple" proteins (Fig. 1C and D). In the dataset of "hybrid" proteins, modern domains were preferentially located at their extremities (Fig. 1A) while ancient domains were located more centrally (Fig. 1B). These data suggested a very simple mode of protein evolution where modern domains were acquired at the extremities of ancient proteins, probably because these regions were more tolerant to the addition of new protein domains. Interestingly, the distribution of ancient domains in "hybrid" proteins, although predominantly central, was slightly shifted towards the carboxy end (Fig. 1B). In agreement with that, the peak of modern domains in the same set of proteins was much higher at the amino portion of the proteins, suggesting that new domains were more frequently acquired by proteins at their amino end.

The set of "hybrid" proteins was probably composed of both ancient proteins (originated before the divergence between prokaryotes and eukaryotes) and modern proteins. Since our criterion for the inclusion of a protein in this dataset was simply the presence of ancient and modern domains in the same molecule, it also contained modern proteins that acquired an ancient domain by modern domain shuffling. In the set of "hybrid" proteins, we identified those in which more than 40% of their length was similar to bacterial

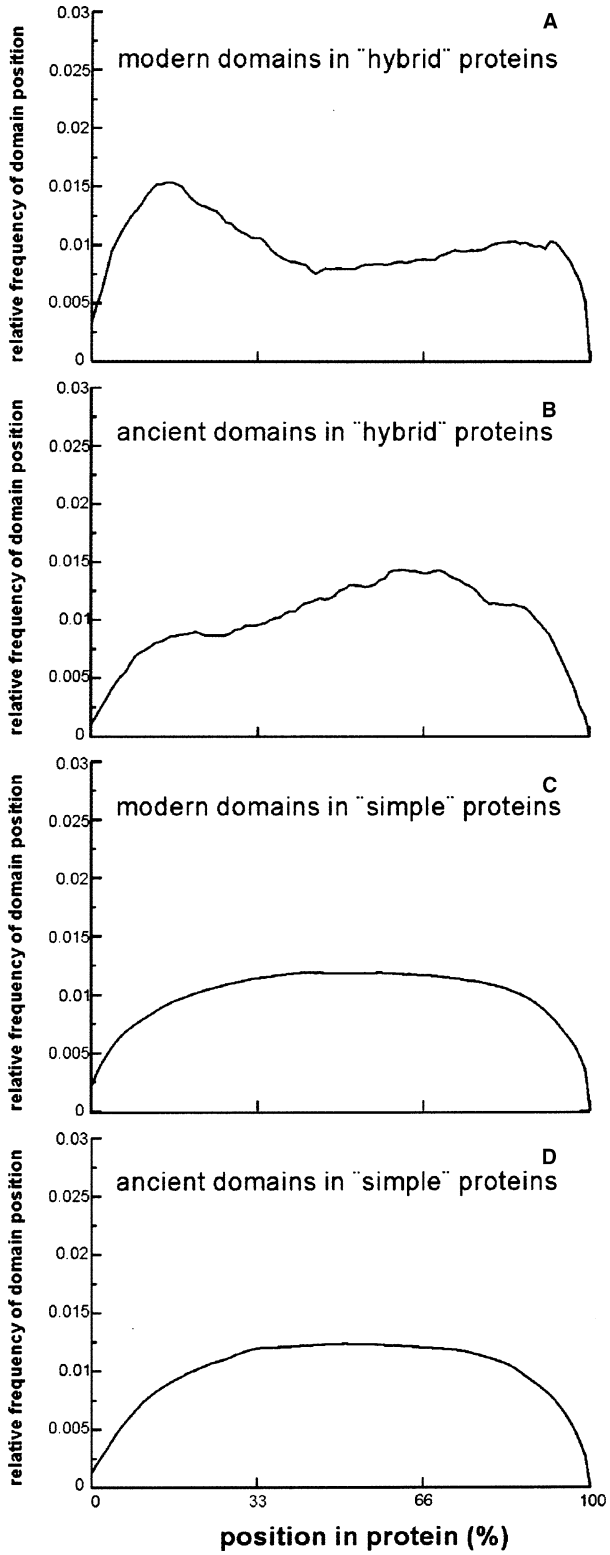

Fig. 1. Domain distributions along proteins from the ExInt database purged with 99% of identity (see Materials and Methods). Distribution of modern domains in "hybrid" (**A**) and "simple" (**C**) proteins. The distribution of ancient domains in "hybrid" and "simple" proteins is shown in (**B**) and (**D**), respectively. The plots present the summed relative frequency of each percent position of all proteins in a given dataset, as a function of protein length (1–100%).

Table 1. The number of ancient and modern domains in the ExInt purged database (99%, see Materials and Methods)

Dataset	Ancient domains	Modern domains
Proteins w/ only ancient domains	55,106 (34,881)	—
Proteins w/ only modern domains	—	25,370 (18,776)
Proteins w/ ancient and modern domains	8045 (4326)	6547 (4326)

Note: The number of proteins found in each dataset is noted in parentheses

sequences (see Materials and Methods for details), corresponding to the fraction of ancient proteins within this set. Figure 2A shows that the pattern observed in "hybrid" proteins in Fig. 1, with ancient domains in the center and modern domains at the extremities, was enhanced when only ancient proteins were considered. The same pattern was also observed in an independent set of ExInt sequences homologous to the 276 ancient proteins identified by Fedorov et al. (2001) (Fig. 2B). What are the mechanisms responsible for the acquisition of protein domains? Kaessmann et al. (2002) and Liu and Grigoriev (2004) observed a high excess of introns of the same phase flanking domains, suggesting that exon-shuffling is the most important mechanism responsible for domain shuffling. Patthy (1996, 1999, 2003) has extensively discussed that exon-shuffling has been widely involved in the creation of new proteins in metazoa through domain shuffling. Based on these results, we evaluated the distribution of symmetric exons along the protein domains analyzed in this study.

The Distribution of Symmetric Exons Is Not Homogeneous Within Protein Domains

Intron phase and symmetric exon distributions were calculated using the ExInt database (Sakharkar et al. 2002). In a dataset of 537,987 introns, we observed a higher frequency of phase zero (49%) followed by phase one (28%) and phase two (23%) introns (χ^2 = 65,873; *d.f.* = 2; $p < 10^{-53}$). There was a significant excess of symmetric exons as well: 10% of excess for 0–0 exons, 23% of excess for 1-1 exons, and 12% of excess for 2–2 exons (χ^2 = 6,037; *d.f.* = 8; $p < 10^{-53}$). These data are in agreement with data previously obtained (Long et al. 1995).

The distribution of the excesses of 0–0 and 1–1 exons was evaluated since those excesses are in the center of several evolutionary arguments about exon-shuffling and also due to the small size of the dataset of 2-2 exons, which makes it more sensitive to statistical fluctuations. To render this result comparable

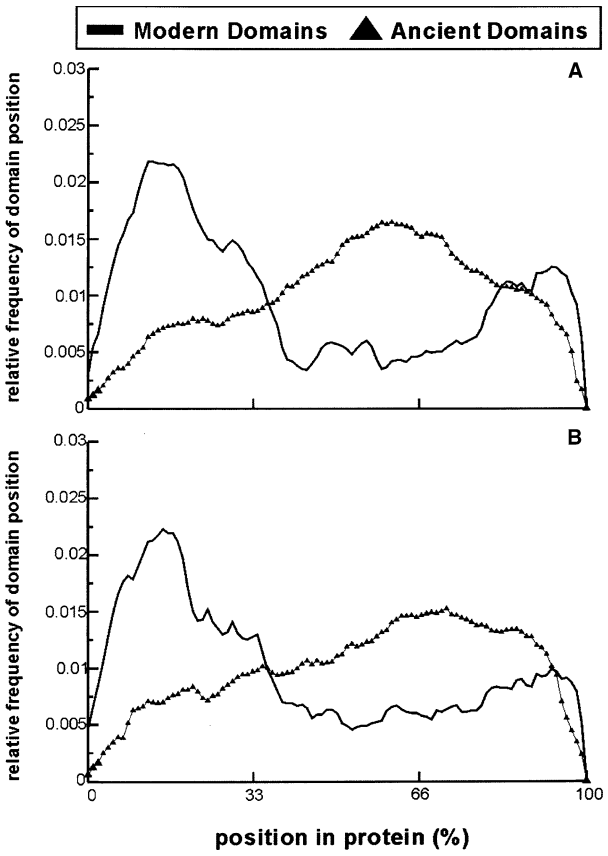

Fig. 2. Domain distributions along "hybrid" proteins from the ExInt database purged with 99% of identity. (**A**) Proteins with sequence similarity to prokaryotic proteins ≥40%. (**B**) Proteins homologous to 276 proteins selected by Fedorov et al. (2001).

to the one presented in Fig. 1, we only evaluated those exons that matched known protein domains from Pfam in the set of "hybrid" proteins. Figure 3A shows that the excess of 0-0 and 1-1 exons was not equally distributed along all proteins from ExInt. The excess of 0-0 exons was predominantly located in the central portion of proteins while the excess of 1-1 exons was larger at the amino and carboxy ends of the proteins. Simulations using random sets of introns showed that the pattern observed in Fig. 3A was observed in only one out of 5000 simulated sets ($p = 0.0002$) (see Materials and Methods). An even sharper pattern was observed in the set of ancient proteins defined by us (Fig. 3B) and in the set of ancient proteins defined by Fedorov et al. (2001) (Fig. 3C). The pattern observed in these two last datasets is rather qualitative as they both present a small number of exons, especially for 1-1 exons.

The data from Figs. 1–3 suggested that shuffling of 0-0 exons might be involved in the evolution of ancient domains, while shuffling of 1-1 exons might be more predominantly involved in the origin of modern domains. Some independent observations give support to this scenario. First, only phase 0 introns are correlated with the boundaries of ancient protein modules (de Souza et al. 1998; Fedorov et al. 2001). Furthermore, 1-1 exons are known to be involved in recent exon-shuffling as shown for modern proteins of multicellular organisms (Patthy 1991, 1996, 1999). Finally, Kaessmann et al. (2002) have shown that ancient domains bounded by introns present an excess of 0-0 exons and a depletion of 1-1 exons while the opposite was found for modern domains. Based on these arguments, we evaluated the distribution of ancient and modern domains flanked by introns of the same phase (for the sake of simplicity they will be called symmetric domains) along proteins.

Distribution of Ancient and Modern Symmetric Domains Along Proteins

We directly determined the distribution of symmetric domains along proteins. For this analysis, we again only used the set of "hybrid" proteins. Six thousand and ten (6010) domains, out of 95,068 (6%) domains found in 99% purged ExInt database, were bounded by introns. This percentage is in agreement with the results found by Kaessmann et al. (2002) using only human genes. However, the number of sequences analyzed is drastically reduced in comparison to the whole domain dataset in our study, affecting the intensity of the observed signals and its level of statistical significance. Both 0-0 and 1-1 modern domains had distributions that were biased towards the amino end of proteins (Fig. 4). This indicates that new domains, independent of being 0-0 or 1-1, are acquired mainly at the amino portion of proteins. Symmetric 0-0 ancient domains were centrally distributed (Fig. 4A), while the distribution of 1-1 ancient domains was predominantly homogeneous throughout the length of the protein, although slightly higher at the extremities (Fig. 4B). These distributions are in accordance with the model proposed above in which modern domains would be acquired at the extremities of proteins by exon-shuffling mainly mediated by 1-1 exons. The central portion of proteins, enriched with ancient domains, would be constructed during evolution mainly through the shuffling of 0-0 domains.

However, the interpretation of data from Fig. 4 is more complex. de Souza et al. (1998) argued that a small fraction of phase 1 introns existed in the progenote and, therefore, could be involved in the formation of ancient domains. Indeed, Long et al. (1995) and de Souza et al. (1998) found a significant excess of 1-1 exons in a set of ancient conserved regions. Our observation of a homogeneous distribution of ancient 1-1 domains is in accordance with these previous observations, since a small fraction of them could have participated in the formation of the core of proteins.

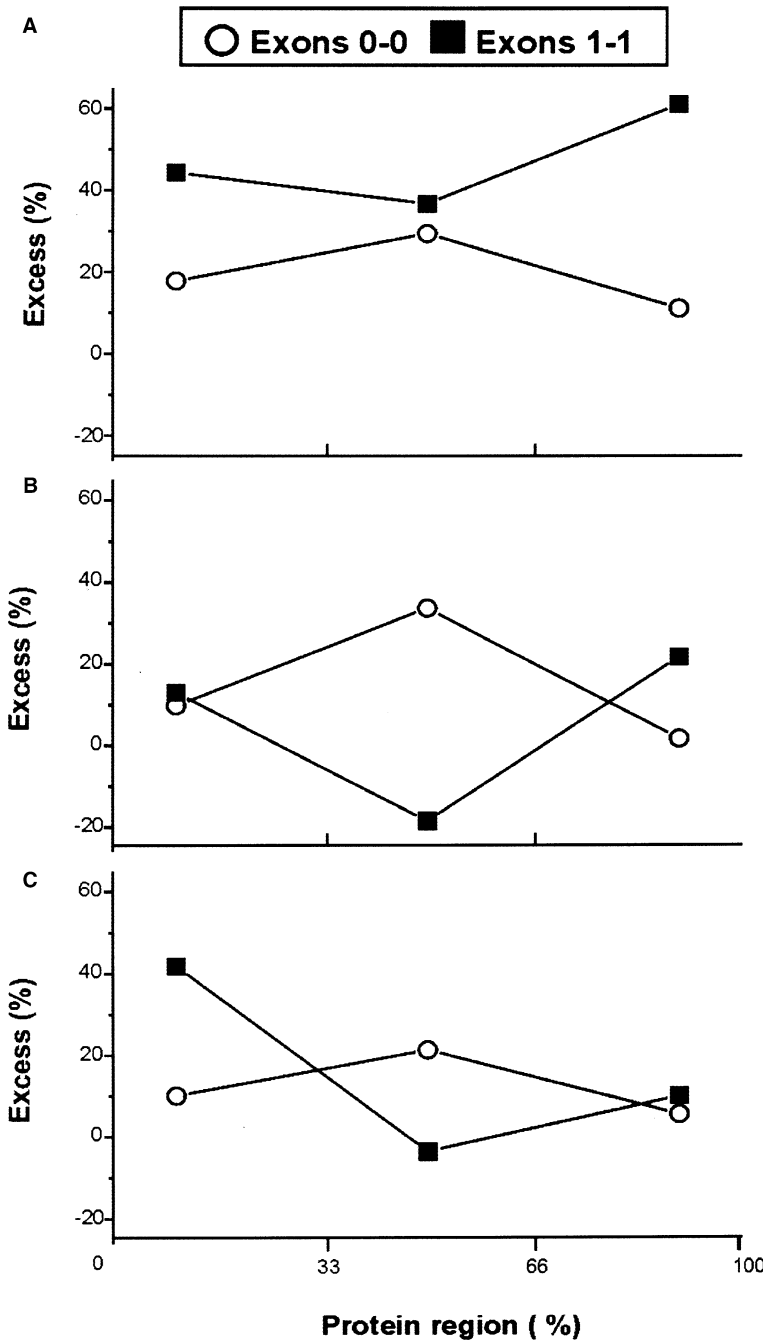

Fig. 3. Distribution of excess of symmetric exons in three equally divided regions (0–33, 34–66, 67–100%) of the proteins from ExInt database (purged with 99% of identity). The analyses were performed only in the regions where domains aligned. (**A**) "Hybrid" proteins (total number of exons observed in all regions; 0-0: 3,537; 1-1: 1,020). (**B**) "Hybrid" proteins with sequence similarity to prokaryotic proteins ≥40% (total number of exons observed in all regions; 0-0: 633; 1-1: 91). (**C**) "Hybrid" proteins homologous to 276 ancient proteins selected by Fedorov et al. (2001) (total number of exons observed in all regions, 0-0: 475; 1-1: 148).

This is further supported by the pattern observed in Figure 3A that shows a significant excess of 1-1 exons in the central part of proteins. To make the scenario even more complex, it is likely that ancient domains were also involved in modern exon-shuffling events. These events could explain the differences between the distribution of ancient 0-0 and 1-1 domains, showing a higher frequency of 1-1 domains at the end of proteins. The differences between these two distributions might also be due to an apparent selective advantage of 1-1 domains in shuffling events (Patthy 2003). Another possibility is that a fraction of modern domains actually corresponds to ancient domains but was not classified as such because of the incompleteness of the current protein databases.

Ancient and Modern Domains are Enriched with 0-0 and 1-1 Domains, Respectively

The data discussed in the last sections (Figs. 1–4) raised the possibility that modern domains in proteins are acquired at the extremities of proteins, mainly through 1-1 exon-shuffling events, while ancient 0-0 domains are centrally located. We performed an analysis to directly verify the representa-

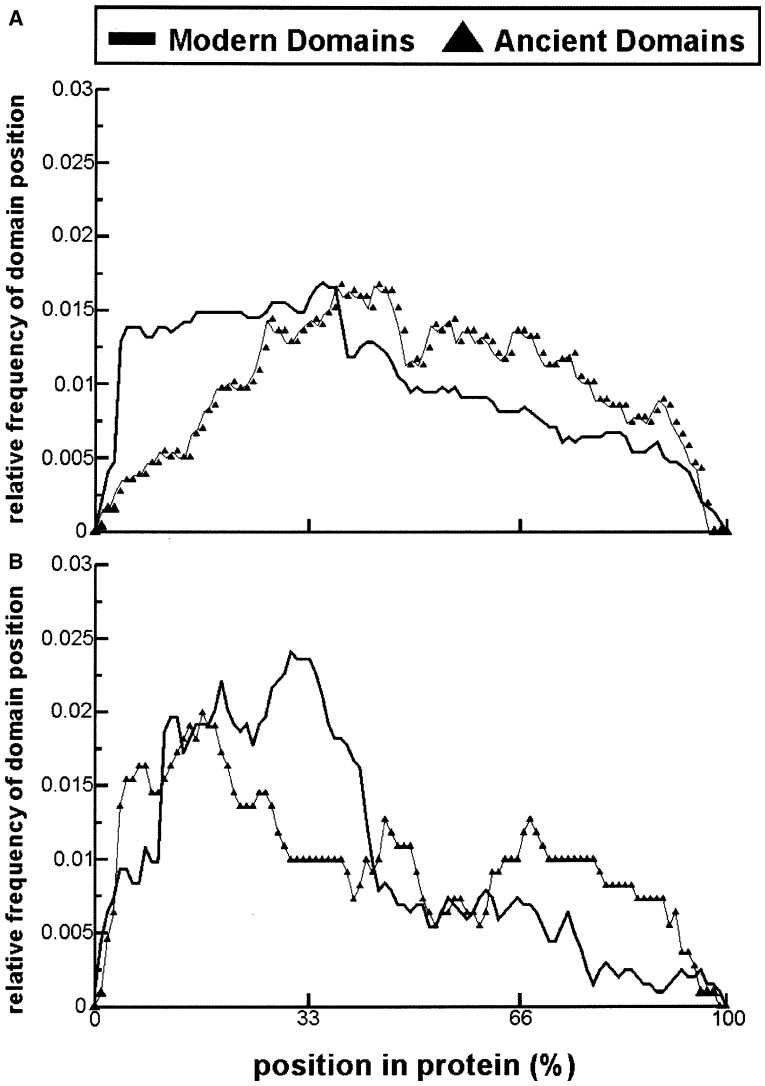

Fig. 4. Distribution of symmetric modern and ancient domains along "hybrid" proteins. (**A**) Domains 0-0 (183 proteins). (**B**) Domains 1-1 (143 proteins).

tion of both 0-0 and 1-1 shuffling units in ancient and modern domains.

Kaessmann et al. (2002) had already shown, using human sequences, that 1-1 domains are overrepresented in modern domains. Furthermore, they also found a marginal overrepresentation of 0-0 domains in the set of ancient domains. In order to directly test this association in our dataset, we checked the frequency of domains flanked by introns of the same phase. Ancient domains presented an excess of 0-0 domains and depletion of 1-1 domains, whereas modern domains showed the opposite pattern; an excess of 1-1 and a depletion of 0-0 domains (Table 2). Simulations involving random sets of the same size as the ancient and modern domains showed that the observed pattern was significant ($p \leq 0.0116$) (Table 2).

The Emergence of a Model of Protein Evolution Involving Ancient and Modern Exon-Shuffling

The analysis presented here is a natural extension of other reports (de Souza et al. 1996, 1998; Fedorov et al. 2001; Kaessmann et al. 2002; Long et al. 1995). In the present analysis, however, we were able to establish a correlation between the age of protein domains, their relative position along proteins and the nature of the introns flanking their corresponding genomic regions. Taken together, the data reported here and the data published previously, as quoted above, strongly suggest that these biased distributions of ancient and modern domains are due to exon-shuffling events occurring at different times during protein evolution. More interestingly, the data allow the proposal of a model that explains how ancient and modern domains were acquired by proteins (Fig. 5).

The model encompasses two major stages. In the first stage, before the divergence of prokaryotes and eukaryotes, the central portion of proteins was constructed mainly by the shuffling of domains flanked by phase 0-0 introns. This is supported by the following observations: (1) 0-0 ancient domains are preferentially located in the central portions of ancient proteins (Fig. 4), (2) ancient domains are en-

Table 2. Observed and expected numbers of ancient and modern symmetric domains (0-0 and 1-1) in the ExInt purged databases

		Sequences of ExInt 99%				Sequences of ExInt 20%			
Exon	Domain	Obs	Exp[a]	Excess (%)	p*	Obs	Exp[a]	Excess (%)	p*
0-0	Ancient	1,335	1,239	7.7	0	427	399	7	0.0115
0-0	Modern	473	569	−16.8	0	204	232	−12	0.0116
1-1	Ancient	356	502	−29.1	0	125	152	−17.7	0.0002
1-1	Modern	376	230	63.3	0	115	88	30.4	0.0002

[a]Expected values were calculated based on exon frequencies among all domains datasets (ancient + modern).
*p values were calculated by simulating 10,000 iterations.

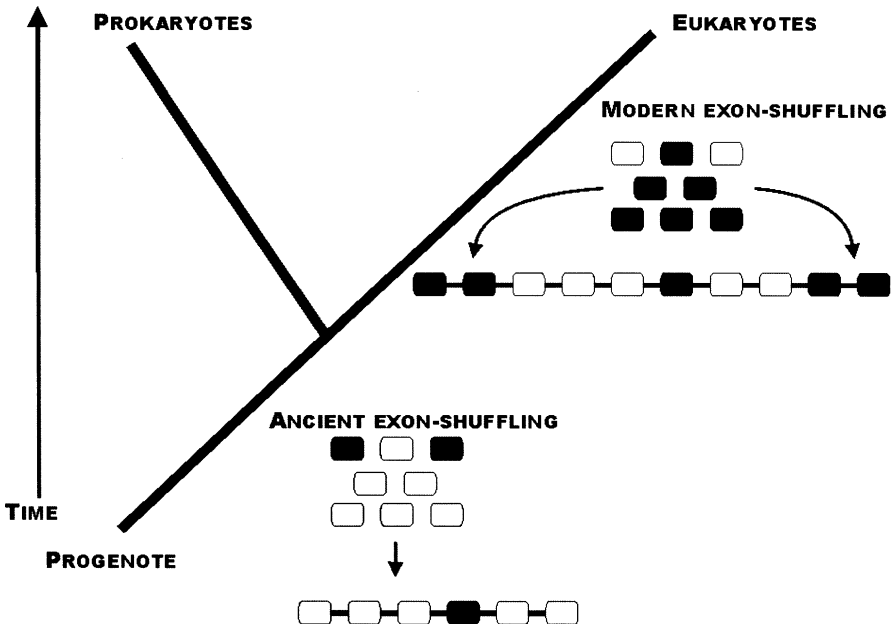

Fig. 5. Scheme representing the model proposed for the formation of proteins by ancient and modern events of exon-shuffling. Introns are represented by black lines, 0-0 exons by open boxes, and 1-1 exons by solid boxes.

riched with 0-0 exons and depleted of 1-1 exons (Table 2 in this report and see Kaessmann et al. 2002), (3) the position of phase 0 introns is correlated to the boundaries of protein modules as shown by de Souza et al. (1998), Roy et al. (1999), and Fedorov et al. (2001). The second stage of the model depicted in Fig. 5 represents modern exon-shuffling involving preferentially 1-1 exons occurring at the extremities of ancient proteins. This is supported by the following observations: (1) 1-1 modern domains are preferentially located at the extremities of proteins (Fig. 4), (2) modern domains are enriched with 1-1 exons and depleted of 0-0 exons (Table 2 in this report and see Kaessmann et al. 2002), and (3) exon-shuffling events in genes originated in multicellular animals are strongly associated with 1-1 exons (Patthy 1987, 1996, 1999, 2003).

It is important to emphasize that the model does not argue for the exclusive involvement of 0-0 and 1-1 domains in exon-shuffling events before and after the divergence between eukaryotes and prokaryotes, respectively. Long et al. (1995) and de Souza et al. (1998) observed a significant excess of 1-1 exons in ancient conserved regions. Exons flanked by these phase 1 introns could be involved in domain shuffling in the progenote. However, domains flanked by phase 0 introns would be predominant in exon-shuffling events in the progenote due to the excess of phase 0 introns. For eukaryotes, the excess of 1-1 domains probably reflects some selective advantage of this type of event.

Final Comments

Roy et al. (2002) elegantly showed that the signal of ancient introns tends to weaken in genes having a higher density of introns. Since sequence databases continue to grow at an exponential rate, it is expected that any signal of putative ancient introns could eventually be washed out. Interestingly, Roy et al. (2002) noticed that their observation matches the expectation of the synthetic theory of intron evolution.

From a theoretical viewpoint, the model proposed in the former report is important in the sense that it defined new parameters and predictions to be tested. The definition of subsets of introns believed to represent the primordial ones is expected to enrich any signal derived from them, if they are really ancient.

Recently, several reports have provided evidence for a putative biased mechanism of intron insertion that would lead to the observed higher frequency of phase 0 introns (Coghlan and Wolfe 2004; Qiu et al. 2004; Sadusky et al. 2004; Wolf et al. 2000). Qiu et al. (2004) argued that these observations, *per se*, weaken the synthetic theory of intron evolution, since they suggest that peculiarities on the distribution of phase 0 introns could be explained by a biased insertional model. Their interpretation does not take into account an alternative model in which introns of different phases were inserted but are under differential selection. Even if we assume conservatively that intron insertion, by an unknown mechanism, is biased towards phase 0 introns, we are still left with evidence that cannot be explained by this model. One would have to assume different types of biases in the mechanism of intron insertion to explain all the reported evidence such as the correlation between phase zero introns and boundaries of proteins modules and excess of symmetric exons in ancient conserved regions.

In summary, the data presented here not only give support to the synthetic theory of intron evolution but allow the emergence of a model that is, *per se*, an extension of such a theory. In the last decade, studies from different groups have generated convincing evidence for the ancient nature of several introns and for the role of exon-shuffling in the creation of proteins in the progenote. From a philosophical perspective, the "introns-early" theory has matured to a more complete, detailed theory that incorporates, in a constructive way, new data and evidence without losing its main conceptual component: the ancient nature of introns.

Acknowledgments. The authors thank Walter Gilbert, Ricardo Brentani, and Natanja Kirschbaum-Slager for the critical reading of the manuscript. M.D.V. and N.T.S. are supported by fellowships from FAPESP.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) The Pfam protein families database. Nucleic Acids Res 30:276–280

Blake CCF (1978) Do genes-in-pieces imply proteins-in-pieces? Nature 273:267

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31:365–370

Cavalier-Smith T (1991) Intron phylogeny: a new hypothesis. Trends Genet 7:145–148

Coghlan A, Wolfe KH (2004) Origins of recently gained introns in *Caenorhabditis*. Proc Natl Acad Sci USA 101:11362–11367

de Souza SJ (2003) The emergence of a synthetic theory of intron evolution. Genetica 118:117–121

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

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

Doolittle WF (1978) Genes in pieces: were they ever together? Nature 272:581–582

Fedorov A, Suboch G, Bujakov M, Fedorova L (1992) Analysis of nonuniformity in intron phase distribution. Nucleic Acids Res 20:2553–2557

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

Gilbert W (1978) Why genes in pieces? Nature 271:501

Gilbert W (1987) The exon theory of genes. Cold Spring Harb Symp Quant Biol 52:901–905

Kaessmann H, Zollner S, Nekrutenko A, Li WH (2002) Signatures of domain shuffling in the human genome. Genome Res 12:1642–1650

Liu M, Grigoriev A (2004) Protein domains correlate strongly with exons in multiple eukaryotic genomes: evidence of exon shuffling? Trends Genet 20:399–403

Long M, Rosenberg C, Gilbert W (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. Proc Natl Acad Sci USA 92:12495–12499

Palmer JD, Logsdon JM Jr (1991) The recent origins of introns. Curr Opin Genet Dev 1:470–477

Patthy L (1987) Intron-dependent evolution: preferred types of exons and introns. FEBS Lett 214:1–7

Patthy L (1991) Molecular exchange principles in proteins. Curr Opin Struct Biol 1:351–361

Patthy L (1996) Exon shuffling and other ways of module exchange. Matrix Biol 15:301–310

Patthy L (1999) Genome evolution and the evolution of exon-shuffling: review. Gene 238:103–114

Patthy L (2003) Modular assembly of genes and the evolution of new functions. Genetica 118:217–231

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448

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 (2003) Recent evidence for the exon theory of genes. Genetica 118:251–266

Roy SW, Nosaka M, de Souza SJ, Gilbert W (1999) Centripetal modules and ancient introns. Gene 238:85–91

Roy SW, Fedorov A, Gilbert W (2002) The signal of ancient introns is obscured by intron density and homolog number. Proc Natl Acad Sci USA 99:15513–15517

Sadusky T, Newman AJ, Dibb NJ (2004) Exon junction sequences as cryptic splice sites: implications for intron origin. Curr Biol 14:505–509

Sakharkar M, Passetti F, de Souza JE, Long M, de Souza SJ (2002) ExInt: an exon intron database. Nucleic Acids Res 30:191–194

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

Stoltzfus A, Logsdon JM, Palmer JD, Jr., Doolittle WF (1997) Intron ''sliding'' and the diversity of intron positions. Proc Natl Acad Sci USA 94:10739–10744

Wolf YI, Kondrashov FA, Koonin EV (2000) No footprints of primordial introns in a eukaryotic genome. Trends Genet 16:333–334