FEBS Letters 580 (2006) 1621–1624

A possible role of exon-shuffling in the evolution of signal peptides of human proteins

Maria Dulcetti Vibranovski[a,b,1], Noboru Jo Sakabe[a,b,1], Sandro José de Souza[a,*]

[a] *Laboratory of Computational Biology, Ludwig Institute for Cancer Research, Sao Paulo Branch, Rua Prof. Antonio Prudente 109, 4° andar, CEP 01509-010, São Paulo, SP, Brazil*

[b] *Departamento 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 25 November 2005; revised 23 January 2006; accepted 31 January 2006

Available online 10 February 2006

Edited by Takashi Gojobori

Abstract It was recently shown that there is a predominance of phase 1 introns near the cleavage site of signal peptides encoded by human genes [Tordai, H. and Patthy, L. (2004) Insertion of spliceosomal introns in proto-splice sites: the case of secretory signal peptides. FEBS Lett. 575, 109–111]. It was suggested that this biased distribution was due to intron insertion at AG|G proto-splice sites. However, we found that there is no disproportional excess of AG|G that would support insertion at proto-splice sites. In fact, all nG|G sites are enriched in the vicinity of the cleavage site. Additional analyses support an alternative scenario in which exon-shuffling is largely responsible for such excess of phase 1 introns.

Keywords: Proto-splice site; Exon-shuffling; Signal peptide; Phase 1 introns

1. Introduction

Although there are evidences suggesting the existence of introns in the common ancestor of prokaryotes and eukaryotes, it is agreed that most introns were inserted during the evolution of eukaryotes [1]. What are the mechanisms involved in intron insertion? One possibility is that intron insertion is random and only those introns flanked by proper nucleotides (in a way that they constitute a suitable splicing site) remain [2,3]. One alternative possibility is that intron insertion is targeted to specific sites [4]. This second possibility is best represented by the proposal from Dibb and Newman that introns are inserted at C/AAG|R proto-splice sites ("|" represents the insertion site) [4]. The most plausible mechanism for intron insertion at proto-splice sites involves the attachment of a given excised intron to components of the spliceosome, a phenomenon already found in Nature [5,6]. By a reverse reaction, the intron is inserted back at a proto-splice site in a heterologous mRNA ensuring that the insertion happens in regions with the exonic signals needed for the splicing process. Reverse transcription and homologous recombination with the original gene puts the inserted intron back into the genome [5]. Although conserved exonic nucleotides in the splice site (AG|G) are observed for all introns, one of the evidences for the existence of proto-splice sites comes from the knowledge that novel introns present even stronger conservation of AG|G than other introns [7–9].

Recently, Tordai and Patthy [10] showed the existence of a significant excess of phase 1 introns (those that interrupt a codon after its first nucleotide) in the vicinity of the cleavage site of signal peptides encoded by human genes. These authors argued that the phase 1 bias was due to intron insertion at a putative proto-splice site AG|G enriched in that region. This enrichment would be due to a high frequency of glycines (encoded by GGn) at positions −1, −3, −4 and −5 in relation to the cleavage site of signal peptides, most probably because positions −1 and −3 can only bear small and neutral amino acids for proper cleavage [11].

However, an alternative possibility, not mentioned by Tordai and Patthy, is that exon-shuffling of the signal peptide itself generated the excess of phase 1 introns in the vicinity of the cleavage site. Support for this alternative possibility comes from the following: (i) most modern events of exon-shuffling involve exons flanked by phase 1 introns (1-1 exons) [12–15]; (ii) human proteins are enriched in modern domains that are encoded by 1-1 exons [16]; (iii) the distribution of modern and ancient domains are correlated to the signal of modern and ancient exon-shuffling (involving 1-1 and 0-0 exons, respectively) [17]; and (iv) exon-shuffling of target sequences has been observed previously [18,19]. Based on the above arguments we wondered whether the pattern observed by Tordai and Patthy [10] would be better explained by a model where signal peptides were predominantly acquired by exon-shuffling. Here we present our findings.

2. Results and discussion

Human protein sequences were downloaded from Swiss-Prot 47.0 [20]. Duplicates were removed resulting in 11 849 sequences, of which 2313 presented an N-terminal signal peptide. The position of the signal peptide was determined on the basis of the annotation provided by Swiss-Prot.

Intron positions and phases for 1823 and 6748 sequences with and without signal peptide, respectively, were obtained by cross referencing Swiss-Prot proteins to genes annotated by Ensembl 26.35 release [21]. Whenever an Ensembl gene

[*] Corresponding author. Fax: +55 11 3207 7001.
E-mail address: sandro@compbio.ludwig.org.br (S.J. de Souza).

[1] These authors equally contributed to this work.

0014-5793/$32.00 © 2006 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2006.01.094

presented more than one product, we selected the protein that corresponded to the Swiss-Prot sequence by comparison of length and identity (Blast alignment [22]). Our dataset is larger than Tordai and Patthy's because we used Ensembl 26.35 instead of EID version 132 [21,23].

In accordance to the results in Tordai and Patthy's [10, Fig. 1], the intron phase distribution along the first 100 amino acids of the proteins containing a signal peptide was biased for phase 1 introns in the vicinity of the cleavage site (Table 1 and Supplemental Figs. S1 and S2). When we selected the nearest introns to the signal peptide (±5 amino acids from the C-terminus of the signal peptide), the biased distribution of phase 1 introns was even more dramatic (Table 1; $\chi^2 = 158.5$; d.f. = 2; $P = 4.9 \times 10^{-35}$).

The explanation given by Tordai and Patthy for the existence of such phase 1 peak is based on the preference of intron insertion at specific sites, namely the AG|G proto-splice site [10]. We found that introns near the signal peptide cleavage site (positions −1, −3, −4 and −5) have a significant higher frequency of AG|G flanking splice sites than all other introns (Table 2, 34% versus 21%; $\chi^2 = 24.5$, d.f. = 1, $P = 9.63 \times 10^{-7}$). In principle, this is in accordance with the proposition from Tordai and Patthy. However, this excess is not restricted to AG|G. In fact we found that the proportion of AG|G to {CTG}G|G is statistically the same in both sets of intron positions – the ones located within the signal peptides compared to all other positions for the same genes (Table 2, 61% versus 58%; $\chi^2 = 0.7$, d.f. = 1, $P = 0.42$). In other words, there is an enrichment for all nG|G sites in the set of introns located close to the cleavage site. The prediction from the idea put forward by Tordai and Patthy is that only AG|G would be enriched in that region.

In fact, the higher proportion of AG|G to {CTG}|G sites is not restricted to genes encoding proteins with a signal peptide, but to introns in general [7]. AG|G corresponds to 61% of the nG|G triplets around the signal peptide and 58% in all other intron positions. For introns from the control set (genes without a signal peptide and no excess of phase 1 introns) AG|G also corresponds to 58% of all nG|G sites.

It should be noted that for phase 1 introns, all nG|G triplets encode glycine. Thus, the highest frequency of all nG|G sites is probably due to an enrichment of glycines in the region near the cleavage site of signal peptides. Together, these results do not support the hypothesis that the preferential intron insertion at AG|G proto-splice sites is the cause of the biased phase 1 intron distribution near signal peptides.

Furthermore, Tordai and Patthy did not fully explore their data since in their Table 1 it is shown that proteins containing a signal peptide have their *entire* genes enriched with phase 1 introns (Table 1 of Ref. [10]). We observed the same pattern in our datasets (Table 1 in this report). Phase 1 introns corresponded to 46% of all introns in proteins containing a signal peptide compared to 29% in proteins lacking a signal peptide ($\chi^2 = 1{,}557$; d.f. = 2; $P < 1.0 \times 10^{-86}$). When considering only introns located beyond the first 100 amino acids, there was still an elevated phase 1 frequency (Table 1; $\chi^2 = 1{,}227$; d.f. = 2; $P < 1.0 \times 10^{-50}$).

Based on the overall abundance of phase 1 introns in genes encoding proteins with signal peptide, the presence of such a steep peak of phase 1 introns only in the first 15–25 amino acids is intriguing. The data in Tordai and Patthy's Fig. 1 [10] may be viewed in a different way. If signal peptides are encoded by one or more exons, intron positions will be concentrated at the carboxy end of the signal peptide. As signal peptides have approximately the same length, the intron positions near the cleavage sites will be found around amino acids 15–25 (see signal peptide lengths in Fig. 1 of Tordai and Patthy [10]). On the other hand, the remaining exons will have different lengths and therefore their intron positions will differ, leading to a dilution of the density of phase 1 introns. We plotted the frequency of intron phases as a function of intron number in order to normalize exon lengths (Fig. 1). One can note that the peak related to signal peptides is considerably smaller (1.5-fold difference, instead of 3-fold), due to a higher frequency of phase 1 in all intron positions along the entire genes, as noted above.

The fact that signal peptides are encoded by one or more exons and genes encoding proteins with signal peptides are enriched with phase 1 introns led us to consider the possibility that these proteins were predominantly constructed by modern exon-shuffling.

We tested this scenario by comparing two experimental sets in regard to five parameters: (1) the presence of modern (present in eukaryotes only) and ancient (present in eukary-

Table 1
Intron phase frequencies for: (i) 1823 human genes encoding proteins with signal peptide, (ii) 6748 human genes encoding proteins without signal peptide, and (iii) 689 human genes with an intron near the cleavage site of the encoded signal peptide

Introns	Without signal peptide			With signal peptide		
	Phase 0	Phase 1	Phase 2	Phase 0	Phase 1	Phase 2
All	27438 (49%)	16331 (29%)	12684 (22%)	5593 (36%)	7076 (46%)	2838 (18%)
First 100 residues	4584 (45%)	3335 (33%)	2189 (22%)	844 (30%)	1458 (51%)	536 (19%)
After the first 100 residues	22854 (49%)	12996 (28%)	10495 (23%)	4749 (38%)	5618 (44%)	2302 (18%)
Near the signal peptide cleavage site (within ± 5 amino acids)	–	–	–	108 (16%)	535 (77%)	46 (7%)

Table 2
Number of phase 1 introns with nG|G sites and other triplets

	AG|G	{CTG}G|G	All other triplets
−1, −3, −4, −5 positions	81 (34%, 61% of nG|G sites)	51 (21%)	106 (45%)
Other positions	1363 (21%, 58% of nG|G sites)	996 (15%)	4227 (64%)

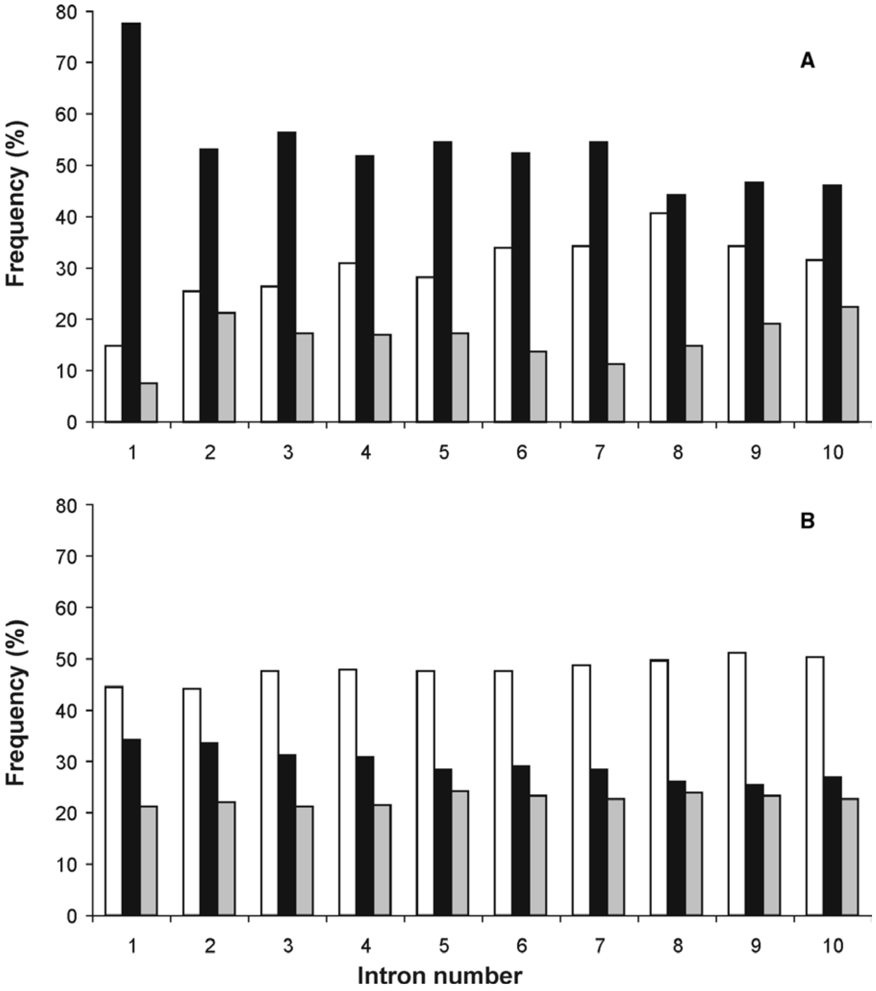

Fig. 1. Distribution of the frequency of intron phases along the first 10 intron positions of human genes, which normalizes exon lengths. (A) Proteins with signal peptides and introns near the cleavage site (689) and (B) without signal peptides (6748). Hollow bars: phase 0, black bars: phase 1, gray bars: phase 2.

otes and prokaryotes) domains according to Pfam 12.0 taxonomical information [24], (2) the similarity to prokaryote proteins using Blast alignment with E-value $<10^{-4}$ against the Swiss-Prot bacterial database [20,22], which would indicate the "antiquity" of the protein set, (3) the frequency of 1-1 domains known to have been shuffled, according to Bányai and Patthy [25], (4) the frequency of putative 1-1 exon-shuffling as evidenced by alignment of all exons against all exons (see Supplemental Material), (5) the excess of symmetrical exons. Data for these analyses are summarized in Table 3. First, analyses 1 and 2 clearly show that proteins containing a signal peptide are enriched with modern domains and depleted of ancient conserved regions, respectively. Second, proteins containing a signal peptide show a higher rate of exon-shuffling, as evaluated by the number of shuffled domains (analysis 3) and the excess of symmetric exons (analysis 5). In accordance with the above features, proteins with signal peptides show significantly more cases of putative exon-shuffling of 1-1 exons (analysis 4). Thus, in all comparisons, the results supported the notion that proteins with a signal peptide evolved predominantly through modern exon-shuffling involving 1-1 exons (Table 3).

Table 3
Comparisons between proteins with and without signal peptides in relation to parameters associated to "modern" exon-shuffling

Analysis	Comparison	Proteins without signal peptides	Proteins with signal peptides	Statistics/observations
1	Proteins with "modern" domains	49% (2337 of 4744)[a]	60% (871 of 1444)[a]	$\chi^2 = 54.2$; d.f. = 1; $P = 1.8 \times 10^{-13}$
2	Proteins similar to prokaryote	27% (1837 of 6748)	18% (330 of 1823)	$\chi^2 = 63.0$; d.f. = 1; $P = 1.9 \times 10^{-15}$
3	Proteins with >1 domain known to have been shuffled	0.014% (87 of 5961)	49% (821 of 1669)	See Supplemental Table S1
4	Frequency of putative 1-1 exon-shuffling	2.7% of exons are shuffled	8.8% of exons are shuffled	$\chi^2 = 725.6$; d.f. = 1; $P < 1.0 \times 10^{-50}$
5	Excess of symmetric exons	7%	23%	$\chi^2 = 43.4$; d.f. = 1; $P = 4.5 \times 10^{-11}$

[a] Not all proteins presented Pfam domains.

In spite of these evidences suggesting that proteins with a signal peptide were constructed through modern exon-shuffling, we failed to obtain direct evidence showing both a donor and acceptor gene for a shuffling event involving signal peptides. Several factors make the inference regarding homology almost impossible. Although signal peptides have constraints in their constitution (positively charged amino acids in the N-terminus and hydrophobic residues in the middle), they may vary substantially, presenting low sequence similarity. Furthermore, they evolve at higher rates than proteins in general [26]. The short-length of signal peptides is also an impediment; the statistical significance of alignments of very short sequences is always low. Therefore, true homologs may be difficult to identify; one can even mistaken them with random matches or convergent evolution.

Nevertheless, acquisition of targeting peptides through exon-shuffling has already been observed. Many studies have shown cases in which genes transferred from organelles (chloroplast, mitochondria and apicoplast) to the nucleus of plants and Apicomplexa acquired an N-terminal transit peptide so to allow transfer of the expressed protein from the cytoplasm back to the organelle ([18] and references therein). As these transit peptides present a downstream intron, as in the case of human signal peptides, they were most probably acquired by exon-shuffling in the nuclear version of the gene. In the case of a mitochondrial targeting peptide reported by Long et al. [19], both the donor and acceptor genes were identified.

Opposite to the view that signal peptides were acquired independently each time during evolution is the scenario where they were acquired few times and subsequently spread to other proteins. This could be achieved through gene duplication and subsequent divergence mainly through independent exon-shuffling events. One such example is plasma proteases, where regulatory modules seem to have been inserted in the phase 1 intron between the signal peptide and the zymogen activation domain of an ancestral protease. All the proteins derived from this ancestor as urokinase, tissue plasminogen activator, neurotrypsin and others bear a signal peptide and many shuffled 1-1 modules [12,13].

Regardless whether signal peptides were acquired independently through exon-shuffling or spread through gene duplication and exon-shuffling, we show here that the biased distribution of phase 1 introns in proteins with signal peptides is unlikely due solely to intron insertion at proto-splice sites. Rather, our data reinforce the importance of modern exon-shuffling in the construction of these mosaic proteins.

Acknowledgements: M.D.V. and N.J.S. are supported by fellowships from FAPESP.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2006.01.094.

References

[1] de Souza, S.J. et al. (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.

[2] Sadusky, T., Newman, A.J. and Dibb, N.J. (2004) Exon junction sequences as cryptic splice sites: implications for intron origin. Curr. Biol. 14, 505–509.

[3] Sverdlov, A.V., Rogozin, I.B., Babenko, V.N. and Koonin, E.V. (2004) Reconstruction of ancestral protosplice sites. Curr. Biol. 14, 1505–1508.

[4] Dibb, N.J. and Newman, A.J. (1989) Evidence that introns arose at proto-splice sites. EMBO J. 8, 2015–2021.

[5] Sharp, P.A. (1985) On the origin of RNA splicing and introns. Cell 42, 397–400.

[6] Tani, T. and Ohshima, Y. (1991) mRNA-type introns in U6 small nuclear RNA genes: implications for the catalysis in pre-mRNA splicing. Genes Dev. 5, 1022–1031.

[7] Stephens, R.M. and Schneider, T.D. (1992) Features of spliceosome evolution and function from an analysis of the information at human splice sites. J. Mol. Biol. 228, 1124–1136.

[8] Coghlan, A. and Wolfe, K.H. (2004) Origins of recently gained introns in *Caenorhabditis*. Proc. Natl. Acad. Sci. USA 101, 11362–11367.

[9] Qiu, W.G., Schisler, N. and Stoltzfus, A. (2004) The evolutionary gain of spliceosomal introns: sequence and phase preferences. Mol. Biol. Evol. 21, 1252–1263.

[10] Tordai, H. and Patthy, L. (2004) Insertion of spliceosomal introns in proto-splice sites: the case of secretory signal peptides. FEBS Lett. 575, 109–111.

[11] Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. 10, 1–6.

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

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

[14] Patthy, L. (1999) Genome evolution and the evolution of exon-shuffling–a review. Gene 238, 103–114.

[15] Tordai, H. et al. (2005) Modules, multidomain proteins and organismic complexity. FEBS J. 272, 5064–5078.

[16] Kaessmann, H., Zollner, S., Nekrutenko, A. and Li, W.H. (2002) Signatures of domain shuffling in the human genome. Genome Res. 12, 1642–1650.

[17] Vibranovski, M.D., Sakabe, N.J., de Oliveira, R.S. and de Souza, S.J. (2005) Signs of ancient and modern exon-shuffling are correlated to the distribution of ancient and modern domains along proteins. J. Mol. Evol. 61, 341–350.

[18] Kilian, O. and Kroth, P.G. (2004) Presequence acquisition during secondary endocytobiosis and the possible role of introns. J. Mol. Evol. 58, 712–721.

[19] Long, M., de Souza, S.J., Rosenberg, C. and Gilbert, W. (1996) Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. Proc. Natl. Acad. Sci. USA 93, 7727–7731.

[20] Boeckmann, B. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365–370.

[21] Hubbard, T. et al. (2005) Ensembl 2005. Nucleic Acids Res 33, D447–D453.

[22] Altschul, S. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

[23] Saxonov, S., Daizadeh, I., Fedorov, A. and Gilbert, W. (2000) EID: the exon–intron database – an exhaustive database of protein-coding intron-containing genes. Nucleic Acids Res. 28, 185–190.

[24] Bateman, A. et al. (2002) The Pfam protein families database. Nucleic Acids Res. 30, 276–280.

[25] Banyai, L. and Patthy, L. (2004) Evidence that human genes of modular proteins have retained significantly more ancestral introns than their fly or worm orthologues. FEBS Lett. 565, 127–132.

[26] Williams, E.J., Pal, C. and Hurst, L.D. (2000) The molecular evolution of signal peptides. Gene 253, 313–322.