Title: Exon elongation added disordered regions to proteins to facilitate evolution of early eukaryotes

Authors: Satoshi Fukuchi[1], Tamotsu Noguchi[2], Hiroto Anbo[1], Keiichi Homma[1*]

Affiliations:

[1]Program for Information Systems, Division of Informatics, Bioengineering and Bioscience, Maebashi Institute of Technology; 460-1 Kamisadori-machi, Maebashi-shi, 371-0816 Japan.

[2]Pharmaceutical Education Research Center, Meiji Pharmaceutical University; 2-522-1 Noshio, Kiyose, Tokyo, 204-8588 Japan.

*Corresponding author. Email: khomma@maebashi-it.ac.jp

Abstract: Most prokaryotic proteins consist of a single structural domain (SD), while eukaryotic proteins are typically comprised of multiple SDs and intrinsically disordered regions (IDRs) that by themselves do not adopt stable structures. How eukaryotic proteins evolved to differ from prokaryotic proteins has not been fully elucidated. Here, we found that the longer internal exons are, the more frequently they encode IDRs in diverse branches of eukaryotes. Based on this observation, we propose the "small bang" model: many exons unconstrained by SDs elongated to encode IDRs, which in turn facilitated the acquisition of additional structural domains to make the last common ancestor of eukaryotes. As two predictions of the model were verified, the model that accounts for the dichotomy between eukaryotic and prokaryotic proteins is plausible.

One-Sentence Summary: We propose the small bang model that explains how eukaryotic proteins evolved to differ from prokaryotic proteins and present evidence for the model.

Main Text:

The last eukaryotic common ancestor (LECA) probably had an intron-rich genome (*1,2*), but massive intron losses subsequently occurred in a few branches of eukaryotes (*3*). Alternative splicing produces multiple variants from one gene and can explain the variation in complexity among eukaryotes (*4*) and approximately 95% of multiexon human genes undergo alternative splicing (*5*). Introns, regarded as selectively disadvantageous, proliferated in the LECA most plausibly by meiotic processes (*6*). Recent research, however, uncovered a potential advantage of having introns: they attenuate the formation of DNA-RNA hybrids (R-loops) that are deleterious in transcription (*7*). Prokaryotic proteins are comprised mostly of structural domains (SDs) and little intrinsically disordered regions (IDRs) that do not adopt stable three-dimensional structures by themselves, while on average ~33% of residues in eukaryotic proteins are in IDRs (*8*). IDRs in eukaryotic proteins are especially prevalent in transcription factors (*9*) and participate in condensates that regulate diverse cellular functions (*10*). IDRs can arise de novo by expansion of repetitive DNA sequences (*11*) and exonization of introns (*12-14*). While prokaryotic proteins mostly consist of a single SD, eukaryotic proteins tend to have multiple SDs (*15-17*).

We endeavor to propose a plausible evolutionary process to explain domain accretion and addition of IDRs in eukaryotic proteins. Although most human internal exons are between 50 to 200 nucleotides (nt) in length (*18*), 5% are exons longer than 1000 nt, are evolutionarily conserved across mammals, and are mostly expressed and translated (*19*). The SRS3 splicing factor promotes the inclusion of large exons enriched with C-nucleotides in vertebrates (*20*) and plays important roles in the occurrence and development of tumors (*21*). Intriguingly, long exons in vertebrates frequently encode IDRs (*20*). We investigated if this tendency is observed in other eukaryotes and thus is likely to have been present in the LECA.

We examined the internal exons of eight model eukaryotes (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana* and *Oryza sativa* subsp. *japonica*) and found that those shorter than 241 nt constitute a majority in all the genomes (Fig. 1). Interestingly, the longer internal exons are, the more frequently they encode IDRs (predicted by DISOPRED3 (*22*)) in all species including non-vertebrates (Fig. 1). Though the correlation coefficients are all positive and significantly different from 0 (table S1, $p<0.001$, *t*-test), they are small mostly because of the preponderance of short internal exons. To assess if the observation is independent of IDR prediction methods used, we used two other programs: DICHOT (*23*) and POODLE-L (*24*). We found the positive correlation without exception irrespective of IDR prediction methods (figs. S1, S2), although three correlations are not statistically significant at $p<0.001$ (fig. S1D, fig. S2 panels D and F, and table S1).

To examine if disorder promoting amino acids are preferentially encoded by long internal exons, we analyzed the frequency distribution of each amino acid. The amino acids were ordered by their propensity to promote disorder (*25*). In long internal exons, the seven least disorder-promoting amino acids are encoded less (fig. S3), the next nine amino acids are encoded at about the same frequency as in short internal exons (fig. S4), while the four most disorder-promoting amino acids are encoded more (fig. S5). The tendencies of individual amino acids are consistent with the general propensity of long internal exons to encode IDRs.

We also examined if long internal exons have high cytidine contents as reported on human exons (*20*). Though longer internal exons of most species displayed the tendency to have

higher cytidine contents, those of *A. thaliana* did not (fig. S6). Thus, SRS3 splicing factor orthologs probably do not operate in all eukaryotes to include long internal exons.

To identify functions of genes with long internal exons, we analyzed the gene ontology (GO) terms that are enriched in such genes in *H. sapiens*, *M. musculus*, *S. pombe* and *A. thaliana* as these are the only ones in the eight species for which a majority of proteins have been annotated in SwissProt/UniProt (*26*), a high-quality protein database. The genes with internal exons longer than 720 nt are annotated with the following GO terms significantly more (chi-square test, $p<0.05$, the gene numbers are shown in table S2) in all four species: nucleus (GO:0005634), nucleoplasm (GO:0005654), nuclear body (GO:0016604), mRNA processing (GO:0006397), double-strand break repair (GO:0006302), and microtubule (GO:0005874). All except for the last one are related to characteristics of nuclear proteins. The same analysis of the genes with extremely long exons (>960 nt) identified four GO terms that are significantly highly represented (chi-square test, $p<0.05$, table S2): nucleoplasm, nuclear body, mismatch repair complex (GO:0032300) and microtubule. Again, all but the last are traits of nuclear proteins.

One explanation for the tendency of long internal exons to encode IDRs is that some short internal exons which existed in early eukaryotes elongated and the elongated sections encoded IDRs. Splicing site alteration, exonization of introns and repeat sequence expansion are possible mechanisms of internal exon elongation. What kind of primordial short exons expanded in length, however? An internal exon that encodes a portion of an SD (stippled in Fig. 2A), which we call a constrained exon, is unlikely to elongate because the inserted segment(s) inside the SD would be structurally destabilizing in most cases and thus would be rarely tolerated. We thus consider it plausible that internal exons that expanded were mostly unconstrained exons.

We propose the small bang model (Fig. 2A) to account for the observation: in early eukaryotes a number of unconstrained exons expanded with the added segments encoding IDRs. The acquired IDRs provided new beneficial functions to the proteins such as efficient binding of transcribed DNA sequences and facilitation of condensation, most frequently in the nucleus. Some of the genes with extended exons recombined with others to give rise to those encoding multiple SDs with the IDRs in the extended segment serving as a spacer between SDs. Proper spacing between SDs is often crucial in protein functions (*27,28*). Similar events occurred multiple times to produce the LECA as they conferred selective advantage. According to this model, the LECA proteome was comprised mostly of multiple SDs and a significant fraction of IDRs, accounting for the dichotomy between prokaryotic and eukaryotic proteins. Having acquired IDRs and multidomain proteins necessary for efficient cellular functions, introns were subsequently lost in a number of branches of eukaryotes as they no longer conferred selective advantage.

If the model is correct, long internal exons in extant eukaryotes are mostly unconstrained exons, that is, the longer internal exons are, the less the observed fraction of constrained exons compared to the expected fraction are. To test this prediction, we used SCOP domains (*29*) (Fig. 2B), because SD boundaries are not easily identifiable from the outputs of IDR prediction programs as a single SD is often artifactually subdivided into multiple segments. Notice that the absence of a SCOP domain alignment does not automatically mean that the segment is an IDR, because an SD whose homologs have not been structurally determined remains unclassified by SCOP and thus may be latent in the segment. The results (Fig. 2C to J) verify the prediction: in all species examined, both the fraction of constrained exons and the

observed to expected ratio decrease as internal exons become longer. In general expected fractions of constrained exons in long internal exons are low (table S3) since SDs longer than the residues encoded by long exons are rare (materials and methods). Since latent SCOPs, i.e., those that have not been identified, will increase both the observed and expected fractions to the same extent, they do not appreciably affect the results.

Is it possible to identify elongated segments and determine if they preferentially occur in unconstrained exons and encode mostly IDRs? Since unequivocal identification of such cases that had happened in early eukaryotes is difficult, we searched for cases of internal exon elongation after the rat-mouse divergence. We identified the segments in internal exons that exist only in the rat or mouse genome (Fig. 3A and B) and regarded them as those inserted after the divergence: the mouse/rat-specific segment is unlikely to have been deleted twice in the human lineage and in the rat/mouse lineage. To exclude cases of alternative splicing involving entire exons, we removed insertions both of whose boundaries coincide (within one amino acid) with exon boundaries. Although this method cannot capture all the inserted segments, the identified cases are likely to be those of elongated internal exons. A majority of insertions were found in short (1- 340 nt) internal exons (table S4). The segments inserted in internal exons encode mostly IDRs with the fractions much higher than those of the rest of internal exons, irrespective of the prediction methods used (Table 1). The fractions of constrained exons are extremely low and are less than the expected values (Fig. 3B and D, table s3).

Four examples of segments inserted in internal segments are presented (Fig. 4). Figure 4A is a case in which an internal exon likely expanded by sequence repetition in the rat lineage, adding an IDR-encoding segment as previously suggested (*11*). Figure 4B represents an example in which the 3' splicing site of an internal exon was probably altered in the mouse lineage to incorporate a previously intronic segment in the exon. An alteration of the 5' splicing site of an internal exon accounts for the case shown as Figure 4C and the inserted segment also encodes an IDR. Two previously contiguous internal exons apparently formed one long exon by exonization of the intervening intron (Fig. 4D) and the inserted segment again encodes an IDR. The latter three are cases of alternative splicing that affect previously existing exons. The fact that the extended segments encode IDRs dovetails with the report that alternatively spliced fragments are mostly associated with IDRs (*30*). All the presented inserted segments except for that of Figure 4D are in unconstrained exons. These findings are congruent with the notion that most long internal exons were produced by insertion of IDR-encoding segments to unconstrained exons. We note that the length distribution of internal exons is likely to be in equilibrium in extant eukaryotes: insertions are offset by deletions, since otherwise, some internal exons would lengthen indefinitely.

A prediction of the latter half of the small bang model is that genes with longer internal exons have more SDs. We analyzed if genes with longer maximum internal exons tend to encode more SCOP domains. With no exceptions, the eight species exhibited the predicted tendency (Fig. 5). We also calculated the average length of the proteins encoded by the genes in each range. The general tendency is that genes with longer maximum internal exons encode longer proteins (Fig. 5). As domain accretion results in elongation of encoded proteins by our model, this observation is consistent with the hypothesis.

The small bang model accounts for the dichotomy between prokaryotic and eukaryotic proteins: eukaryotic proteins have more IDRs and more SDs than prokaryotic counterparts. The model is consistent with the observed tendency of long internal exons to encode IDRs and

the low frequency of constrained exons in long internal exons. The actual cases of internal exons that likely elongated in the mouse or rat genome are compatible with the model. The fact that genes with longer internal exons exhibit a tendency to encode more SDs also supports the model. The model predicts that the lengths of constrained exons are more evolutionarily conserved than those of unconstrained exons. Another prediction is that genes with longer internal exons have more combinations of SDs that do not exist in prokaryotes. Future research will hopefully test and further refine the model.

References and Notes

1. E. V. Koonin, The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol. Direct* **1**, 22 (2006).
2. M. Irimia, S. W. Roy, Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb. Perspect. Biol.* **6**, a016071 (2014).
3. M. Csuros, I. B. Rogozin, E. V. Koonin, A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput. Biol.* **7**, e1002150 (2011).
4. L. Chen, S. J. Bush, J. M. Tovar-Corona, A. Castillo-Morales, A. O. Urrutia, Correcting for differential transcript coverage reveals a stong relationship between alternative splicing and organism compexity. *Mol. Biol. Evol.* **31**, 1402-1413 (2014).
5. Q. Pan, O. Shai, L. J. Lee, B. J. Frey, B. J. Blencowe, Deep surveying of alternative splicing compexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413-1415 (2008).
6. A. M. Poole, Did group II intron proliferation in an endosymbiont-bearing archaeon create eukaryotes? *Biol. Direct* **1**, 36 (2006).
7. D. K. Niu, Protecting exons from deleterious R-loops: a potential advantage of having introns. *Biol. Direct* **2**, 11 (2007).
8. J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, D. T. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635-645 (2004).
9. T. Minezaki, K. Homma, A. R. Kinjo, K. Nishikawa, Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. *J. Mol. Biol.* **16**, 1137-1149 (2006).
10. M. Fuxreiter, M. Vendruscolo, Generic nature of the condensed states of proteins. *Nat. Cell Biol.* **23**, 587-594 (2021).
11. P. Tompa, Intrinsically unstructured proteins evolve by repeat expansion. *BioEssays* **25**, 847-855 (2003).
12. F. A. Kondrashov, E. V. Koonin, Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends Genet.* **19**, 115-119 (2003).
13. R. Sorek, The birth of new exons: mechanisms and evoutionary consequences. *RNA* **13**, 1603-1608 (2007).
14. Y. Marquez, M. Höpfler, Z. Ayatollah, A. Barta, M. Kalyna, Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity. *Genome Res.* **25**, 995-1007 (2015).
15. E. V. Koonin, L. Aravind, A. S. Kondrashov, The impact of comprehensive genomics on our understanding of evolution. *Cell* **101**, 573-576 (2000).
16. E. V. Koonin, Y. I. Wolf, G. P. Karev, The structure of the protein universe and genome evolution. *Nature* **420**, 218-223 (2002).
17. H. Tordai, A. Nagy, K. Farkas, L. Bányai, L. Patthy, Modules, multidomain proteins and organismic compexity. *FEBS J.* **272**, 5067-5078 (2005).
18. E. S. Lander *et al*., Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
19. M. T. Bolisetty, K. L. Beemon, Splicing of internal large exons is defined by novel *cis*-acting sequence elements. *Nucl. Acids Res.* **40**, 9244-9254 (2012).

20. T. Kawachi *et al*., Regulated splicing of large exons is linked to phase-separation of vertebrate transcription factors. *EMBO J.* **40**, e107485 (2021).
21. J. Xiong, Y. Chen, W. Wang, J. Sun, Biological function and molecular mechanism of SRSF3 in cancer and beyond. *Oncol. Lett.* **23**, 21 (2022).
22. D. T. Jones, D. Cozzetto, DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857-863 (2015).
23. S. Fukuchi, K. Hosoda, K. Homma, T. Gojobori, K. Nishikawa, Binary classification of protein molecules into intrinsically disordered and ordered segments. *BMC Struct. Biol.* **11**, 29 (2011).
24. S. Hirose, K. Shimizu, S. Kanai, Y. Kuroda, T. Noguchi, POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* **23**, 2046-2053 (2007).
25. A. K. Dunker, *et al*., The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* **9**, S1 (2007).
26. UniProt Consortium, UniProt: the universal knowlegebase in 2021. *Nucl. Acids Res.* **49**, D480-D489 (2021).
27. H. C. van Leeuwen, M. J. Strating, M. Rensen, W. de Laat, P. C. van der Vilet, Linker length and composition influence the flixibility of Oct-1 DNA binding. *EMBO J.* **16**, 2043-2053 (1997).
28. S. Brodsky *et al*., Intrinsically disordered regions direct transcription factor *in vivo* binding specificity. *Mol. Cell* **79**, 459-471 (2020).
29. A. Andreeva, E. Kulesha, J. Gough, A. G. Murzin, The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucl. Acids Res.* **48**, D376-D382 (2019).
30. P. R. Romero *et al*., Alternative splicing in concert with protein instinsic disorder enables increased functional diversity in multicellular organisms. *Proc. Natl Acad. Sci. USA* **103**, 8390-8395 (2006).
31. F. Cunningham *et al*., Ensembl 2022. *Nucl. Acids Res.* **50**, D988-D995 (2022).
32. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).

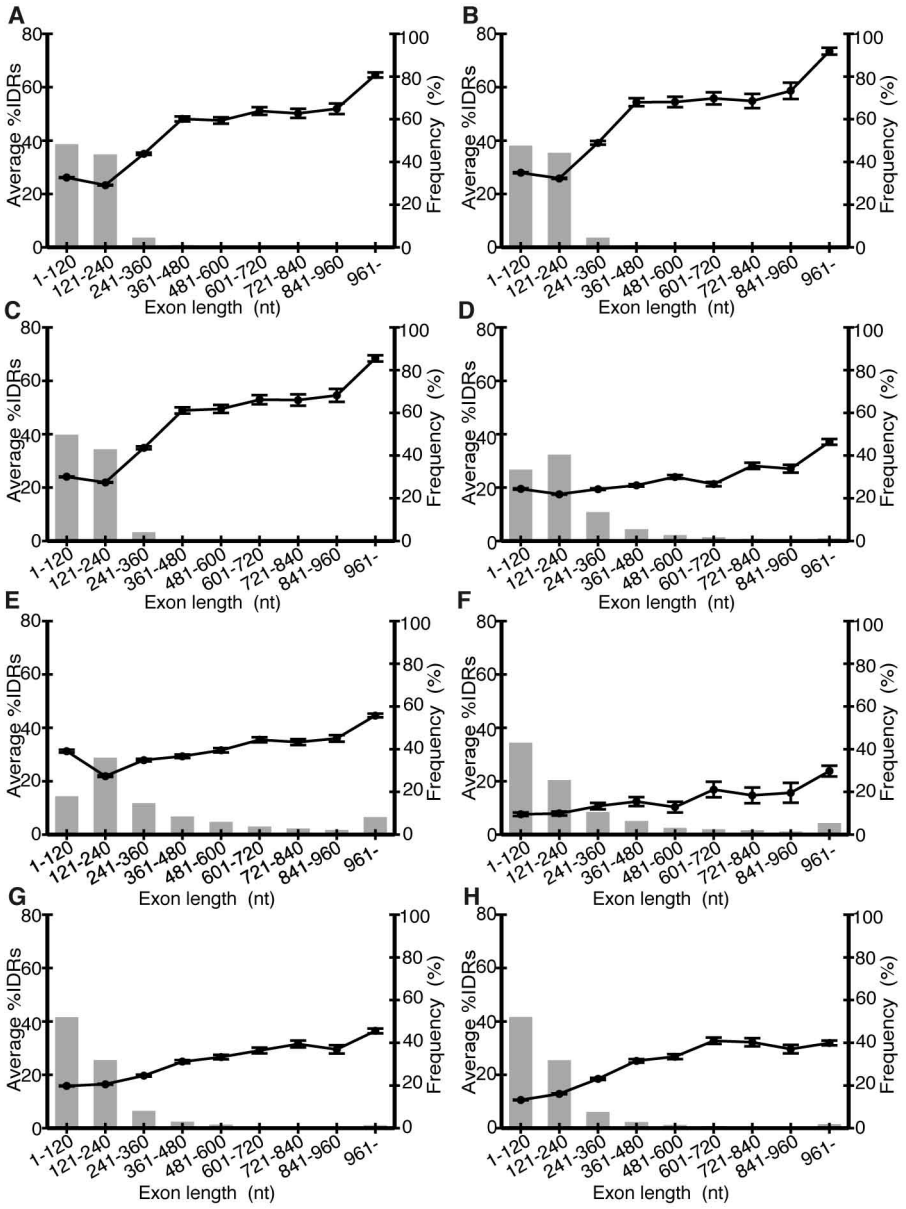

Fig. 1. Long internal exons constitute a minority and tend to encode IDRs. The frequencies of occurrence of internal exons in the indicated length ranges are shown by bar graphs (right scale), while the average fractions of IDRs encoded by the internal exons in the length ranges with standard errors of the mean (SEMs) are displayed by line graphs (left scale). The species presented are (**A**) *H. sapiens*, (**B**) *M. musculus*, (**C**) *R. norvegicus*, (**D**) *C. elegans*, (**E**) *D. melanogaster*, (**F**) *S. pombe*, (**G**) *A. thaliana*, and (**H**) *O. sativa*.

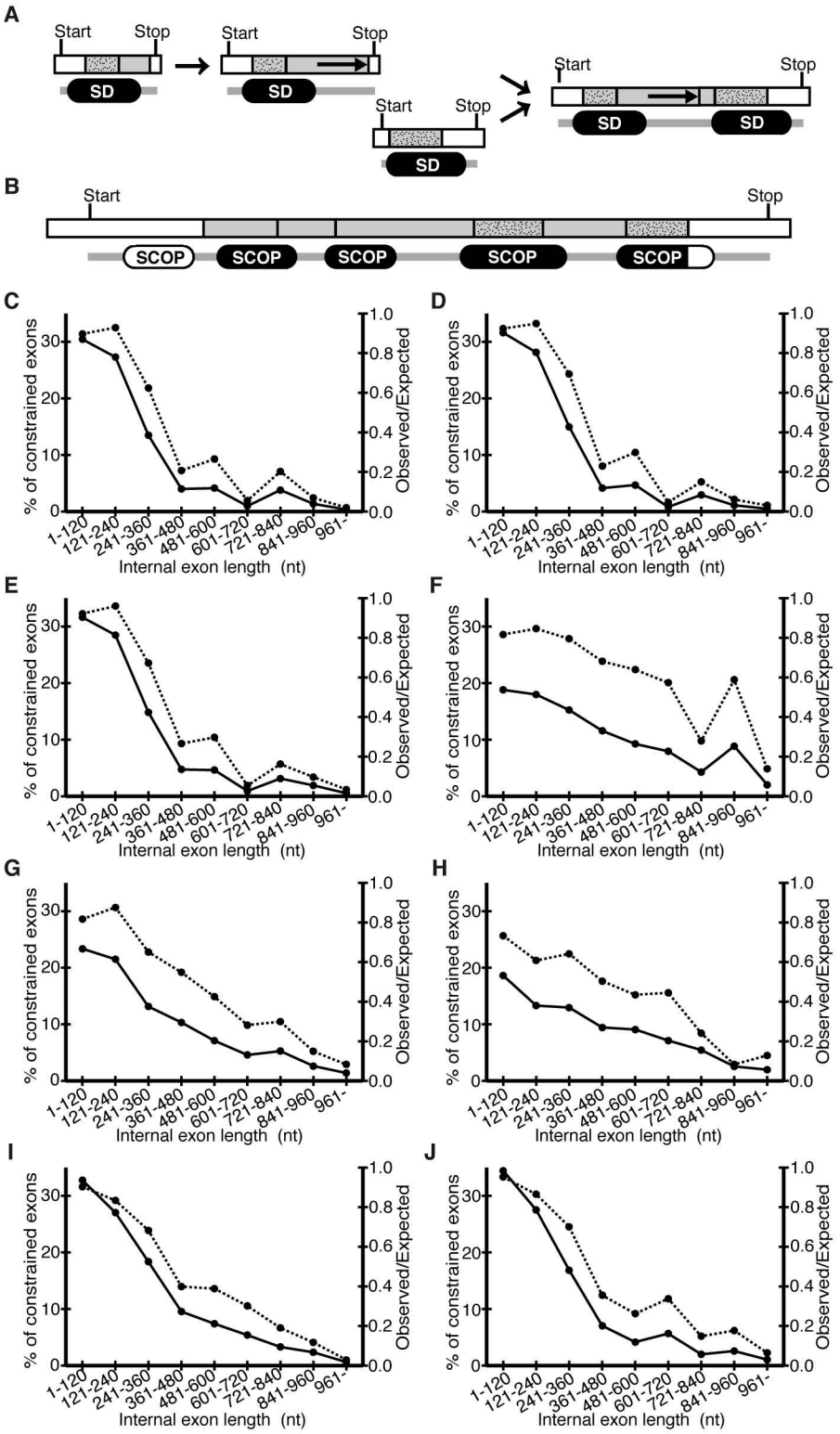

Fig. 2. The small bang model and the tendency that longer internal exons contain less constrained exons. (**A**) The small bang model. Terminal and internal exons are shown as white and grey rectangles, respectively, with constrained exons stippled. The IDRs and SDs in the encoded proteins are respectively presented as grey bars and rounded rectangles. Arrows inside exons indicate an elongated section. (**B**) A hypothetical case presented as in (A) except

that SDs are replaced by SCOP domains. Internal exons both of whose boundaries fall inside a SCOP domain are classified as constrained exons. Only sections of SCOP domains encoded by internal exons (black) are taken into account. (**C** to **J**) The observed fraction of constrained exons (solid line, left scale) and the ratio of the observed fraction to the expected fraction (dashed line, right scale) are shown. The species are (**C**) *H. sapiens*, (**D**) *M. musculus*, (**E**) *R. norvegicus*, (**F**) *C. elegans*, (**G**) *D. melanogaster*, (**H**) *S. pombe*, (**I**) *A. thaliana*, and (**J**) *O. sativa*.

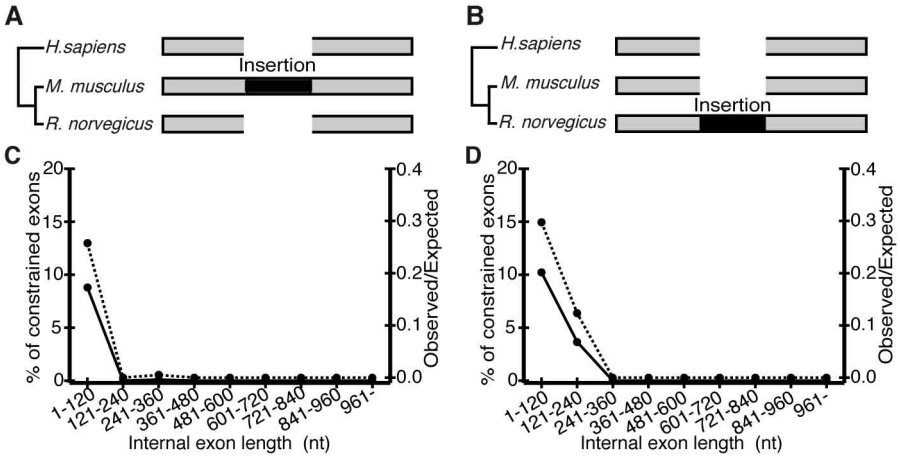

Fig. 3. Segments inserted after the rat-mouse divergence encode mostly IDRs and preferentially occur in unconstrained exons. (**A** and **C**) Mouse-specific insertions. (**B** and **D**) Rat-specific insertions. (**A** and **B**) A scheme to identify probable inserted segments in internal exons. (**C** and **D**) The observed fraction of constrained exons in the inserted segments (solid line, left scale) and the ratio of the observed fraction to the expected fraction (dashed line, right scale). The expected fractions are those of all internal exons.

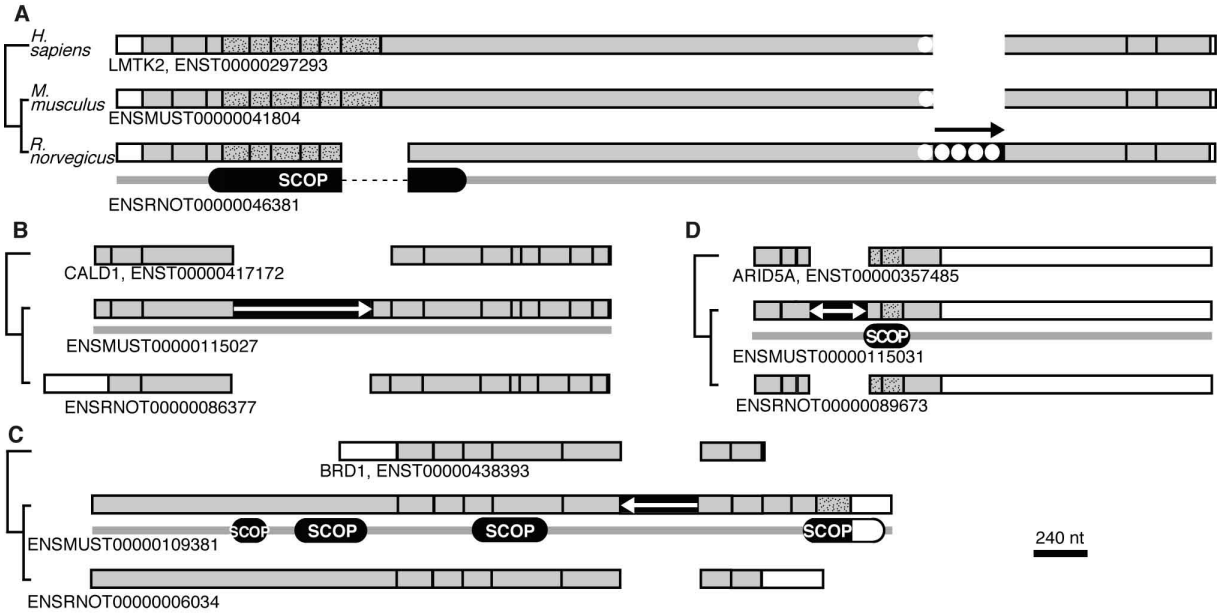

Fig. 4. Actual examples of elongation in internal exons. Human, mouse, and rat orthologs are drawn with corresponding exons mostly aligned vertically on the indicated scale. The coding regions of the terminal exons (white), probable inserted segments in internal exons (black), and the rest of internal exons (grey) of orthologs are shown with the constrained exons stippled. A single-headed arrow signifies the probable directions of expansion, while a double-headed arrow indicates an exonized intron. The transcript IDs and human gene names are added to the right of the phylogenetic trees and the SCOP domains in encoded proteins are depicted beneath exons. No SCOP domains were assigned in panel B.

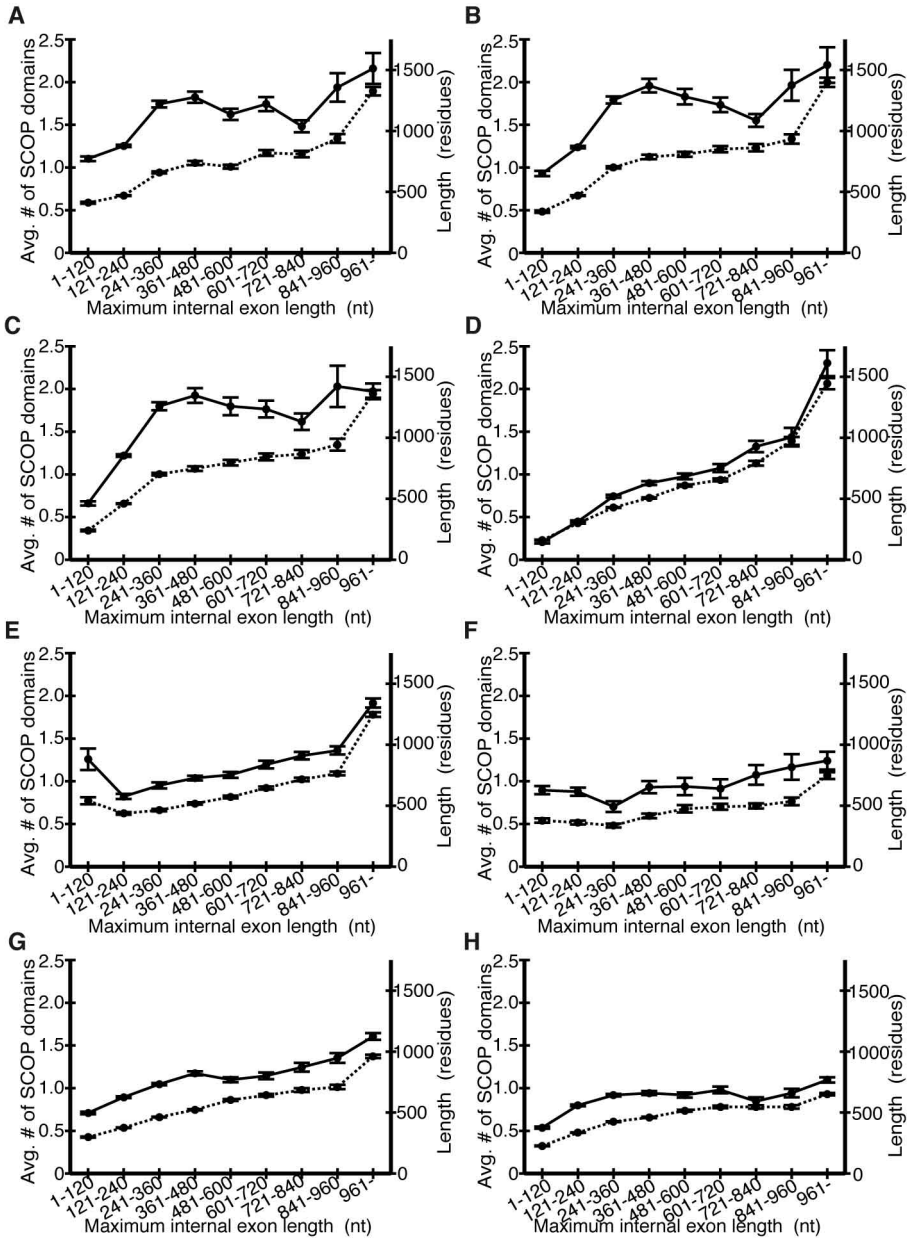

Fig. 5. Genes with longer internal exons tend to encode proteins with more SDs. The average number of SCOP domains with SEM in genes with each range of maximum internal exons (solid line, left scale) and the average length of encoded proteins with SEM (dashed line, right scale) in (**A**) *H. sapiens*, (**B**) *M. musculus*, (**C**) *R. norvegicus*, (**D**) *C. elegans*, (**E**) *D. melanogaster*, (**F**) *S. pombe*, (**G**) *A. thaliana*, and (**H**) *O. sativa*.

Table 1. Per cent IDRs encode by inserted segments and by the rest of internal exons.

	M. musculus			*R. norvegicus*		
Prediction method	DISO-PRED3	DICHOT	POODLE-L	DISO-PRED3	DICHOT	POODLE-L
Inserted segments	81.2	90.6	72.2	63.1	86.5	58.1
Others	41.9	46.0	38.2	34.7	44.2	37.9