Exon elongation added intrinsically disordered regions to proteins and facilitated the emergence of the last eukaryotic common ancestor

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

[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

* To whom correspondence should be addressed. Tel: +81-27-265-7334; Fax: +81-27-226-5168; Email: khomma@maebashi-it.ac.jp

ABSTRACT

Most prokaryotic proteins consist of a single structural domain (SD) with little intrinsically disordered regions (IDRs) that by themselves do not adopt stable structures, while typical eukaryotic proteins are comprised of multiple SDs and IDRs. 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 eight eukaryotes including vertebrates, invertebrates, a fungus and plants. Based on this observation, we propose the 'small bang' model: many exons unconstrained by SDs elongated to encode IDRs, which facilitated the acquisition of additional SDs to make the last common ancestor of eukaryotes. One prediction of the model is that long internal exons are mostly unconstrained exons. Analytical results of the eight eukaryotes were consistent with the prediction. In support of the model, we identified cases of internal exons that elongated after the rat-mouse divergence and discovered that the expanded sections are mostly in unconstrained exons and preferentially encode IDRs. Another prediction that genes with longer internal exons tend to have downstream exons that encode more SDs was also verified. The model that accounts for the dichotomy between prokaryotic and eukaryotic proteins is thus plausible.

INTRODUCTION

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 evolutionary process leading to the LECA probably 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 exonisation of introns (12-14). While prokaryotic proteins mostly consist of a single SD, eukaryotic proteins tend to have multiple SDs (15-17).

We endeavoured to propose a plausible evolutionary process that explains 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 SRSF3 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 tumours (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.

MATERIALS AND METHODS

All the sequence, exon and orthologue data used in this study were downloaded from the Ensembl database (22). The exons were made non-redundant by in-house C programs. IDRs were predicted by DISOPRED3 (23), DICHOT (24) and POODLE-L (25), while SCOP domains (26) were assigned in an intermediate step in DICHOT prediction.

Using an in-house C script, we analyzed the proteins to calculate the fraction of internal exons encoding SCOP domains to get the expected occurrence of constrained exons. If L is the total length encoded by internal exons rounded to the nearest integer, the probability of an exon encoding y amino acids to have both boundaries located inside of a SCOP domain of length x_i is given by $(x_i - y - 1)/(L-y+1)$ if $x_i > y+1$ and is zero otherwise. The total probability is the sum of the probability over all SCOP domains encoded by the gene. Only the portions of SCOP domains encoded by internal exons are considered.

Mouse or rat specific insertion segments were identified from the MAFFT (27) alignments of orthologous protein sequences with default settings. We wrote a C program to assess enrichment of GO terms according to SwissProt/UniProt (28) annotations (release 2021_04). An in-house C script computed the frequencies of nucleotides and amino acids.

To count the number of SCOP domains encoded by specified exons, we wrote a C program. The same program also calculates the expected occurrence of SCOP domains, based on the following: if D is the overall density of SCOP domains per residue in all proteins of the species, the expected number of SCOP domains in a coding sequence of z residues is given by zD.

RESULTS

The longer internal exons are, the more frequently they encode IDRs

We examined the internal exons of eight model eukaryotes including three vertebrates (*Homo sapiens*, *Mus musculus* and *Rattus norvegicus*), two invertebrates (*Caenorhabditis elegans* and *Drosophila melanogaster*), one fungus (*Schizosaccharomyces pombe*) and two plants (*Arabidopsis thaliana* and *Oryza sativa* subsp. *japonica*). Internal exons shorter than 241 nt constitute a majority in all the genomes (Figure 1). Interestingly, we found that the longer internal exons are, the more frequently they encode IDRs (predicted by DISOPRED3 (23)) in all species (Figure 1). The correlation coefficients are all positive and significantly different from 0 (Supplementary Table S1, $p<0.001$, *t*-test).

To assess if the observation is robust to IDR prediction methods used, we used two other programs: DICHOT (24) and POODLE-L (25). We found the positive correlation without exception irrespective of IDR prediction methods (Supplementary Figures S1, S2), although three correlations are not statistically significant from 0 at $p<0.001$ (Supplementary Figure S1D, Supplementary Figure S2 panels D and F and Supplementary Table S1). 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 a tendency to have higher cytidine contents, those of *A. thaliana* did not (Supplementary Figure S3). The inclusion of long internal exons is thus unlikely to be universally mediated by the SRSF3 splicing factor.

Genes with long internal exons are enriched with those with nucleus-specific functions

To identify functions of genes with long internal exons, we analysed 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 (28), 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 number of genes in each range is shown in Supplementary Table S2) in all the 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 nucleus-specific functions. The same analysis of the genes with extremely long exons (>960 nt) identified four GO terms that are significantly highly represented (chi-square test, $p<0.05$): nucleoplasm, nuclear body, mismatch repair complex (GO:0032300) and microtubule. Again, all but the final term are functions specific to the nucleus.

The small bang model

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 most of the elongated sections encoded IDRs. Splicing site alteration, exonisation of introns and repeat sequence expansion are possible mechanisms of internal exon elongation. Which primordial short exons were selected to expand in length? An internal exon that encodes a portion of an SD (stippled in Figure 2A), which we call a constrained exon, is presumably resistant to elongation because the inserted segment(s) inside the SD would be structurally destabilizing in most cases and thus would be rarely tolerated. We thus considered it probable that internal exons that expanded were mostly unconstrained exons. Moreover, we consider it likely that the added sections encoding IDRs facilitated recombination to give rise to multiple SDs: firstly, IDRs probably facilitate recombination as they enable recombination to take place at an inter-SD site, leaving SDs intact, while recombination in the absence of IDRs frequently disrupts SDs to destabilize protein structure and thus is evolutionarily selected against. Secondly, IDRs are likely to promote domain accretion as they serve to keep a proper spacing between SDs, which is crucial in many protein functions (29,30).

We therefore propose the small bang model (Figure 2A): in early eukaryotes a plethora 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 often in the nucleus. Some of the genes with extended exons recombined with each other to give rise to those encoding multiple SDs with the IDRs in the extended segment serving as a spacer between SDs. These 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.

Long internal exons are mostly unconstrained exons

If the model is correct, long internal exons in extant eukaryotes must be mostly unconstrained exons, that is, the longer internal exons are, the less the observed fraction of constrained exons will be. To test this prediction, we used SCOP domains (26) (Figure 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 SDs whose homologues have not been structurally determined remain unclassified by SCOP and thus may be latent in the segment. The results (Figure 2C-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 (Supplementary Table S3) since SCOP domains longer than the region encoded by long exons are rare (Materials and Methods). As latent SCOP domains, i.e., those that have not been identified, will increase both the observed and expected fractions to the same extent, they do not essentially affect the results.

Insertions preferentially occur in unconstrained exons and mostly encode IDRs

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 (Figure 3A and D) 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 cases of elongated internal exons, the identified cases are probably genuine ones. It turned out that a majority of insertions are in short (1- 340 nt) internal exons (Supplementary 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 (Figure 3B and D). The fractions of constrained exons are extremely low and are less than the expected values (Figure 3C and F, Supplementary Table S3).

Four examples of segments inserted in internal segments are presented (Figure 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 presumably 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 exonisation of the intervening intron (Figure 4D) and the inserted segment again encodes an IDR. The latter three are cases of alternative splicing that affected previously existing exons. The fact that the extended segments encode IDRs dovetails with the report that alternatively spliced fragments are mostly associated with IDRs (31). All the presented inserted segments are located 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.

Genes with longer internal exons tend to encode more SDs

A prediction of the latter half of the small bang model is that exons downstream of longer internal exons are apt to encode more SDs. To test it, we counted SCOP domains that are entirely encoded by the longest internal exon itself as well as downstream exons, because intron losses after acquisition of SDs may have eliminated exon boundaries (Figure 5A). With no exceptions, all the eight species exhibited the predicted tendency (Figure 5B-I). We also calculated the mean length of the proteins encoded by the longest internal exon and the downstream exons. Exons downstream of longer internal exons tend to encode longer protein regions (Figure 5B-I). As domain accretion results in elongation of encoded proteins according to our model, the latter observation is consistent with the hypothesis, too.

DISCUSSION

The small bang model accounts for the dichotomy between prokaryotic and eukaryotic proteins: eukaryotic proteins have the higher fraction of IDRs and the more SDs than prokaryotic counterparts. Firstly, the tendency of longer internal exons to encode more IDRs supports this model. Although the correlations are all statistically significant, the slopes in some species, especially in *S. pombe* (Figure 1F), are small. We consider it likely that massive exon losses that had occurred in some branches of eukaryotes after the LECA weakened the initial signal. The fact that the trend is detectable in all the eight species, however, makes it plausible that the tendency is traceable to the LECA. The same signal is predicted to be detectable in all other eukaryotes with a significant number of internal exons.

Genes with long internal exons were enriched with those with nuclear protein functions. This observation is congruent with the report that IDRs play crucial roles in transcription factors (9). Some of the newly acquired IDRs may have participated in condensates that regulate multitudes of cellular functions, conferring evolutionary advantage.

Secondly, the model is consistent with the observed low frequency of constrained exons in long internal exons in all eight eukaryotes. It is likely that many unconstrained exons elongated with the elongated sections encoding mostly IDRs. If the model is correct, this phenomenon must be observed in other eukaryotes with a large enough number of internal exons. The actual cases of internal exons that probably elongated in the mouse or rat genome are compatible with the model.

Lastly, the finding that genes with longer internal exons tend to encode more SDs in the downstream exons without exception also supports the model. According to the model, this tendency must be universally observed in eukaryotes that retain a large enough number of internal exons. The elongated sections of internal exons that mostly encoded IDRs probably facilitated recombination to add SDs. We acknowledge that this piece of evidence is also compatible with the alternative model with the ordering of events reversed, namely, domain accretion occurred prior to the expansion of unconstrained exons. However, we consider the proposed model more plausible: elongation of exons mostly produces segments encoding IDRs, which are likely to enable recombination that adds SDs while preserving the structures and in some cases promoting functions. As many multidomain proteins provide novel functions, this step plausibly conferred selective advantage. Since recombination events in the alternative model are predicted to produce dysfunctional genes with disrupted SDs at a high frequency, it is less plausible than the small bang model.

Having acquired IDRs and multidomain proteins necessary for efficient cellular functions, many branches of eukaryotes such as most protists experienced a massive loss of introns, as they no longer gave significant selective advantage. The putative evolutionary advantage provided by introns through attenuation of R-loop formation must be minimal as introns have been entirely lost in a number of extant eukaryotes. On the other hand, the existence of introns in primordial eukaryotes made it possible for many internal exons to expand, which conferred selective advantage, according to the small bang model. The small bang model thus proposes a plausible beneficial function of introns.

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.

AVAILABILITY

All data are included in the manuscript and the supplementary file.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We thank K. Hosoda for setting up a system to run POODLE-L in house.

FUNDING

This work was supported by a grant to K.H. from Maebashi Institute of Technology.

CONFLICT OF INTEREST

None declared.

REFERENCES

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

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

FIGURES

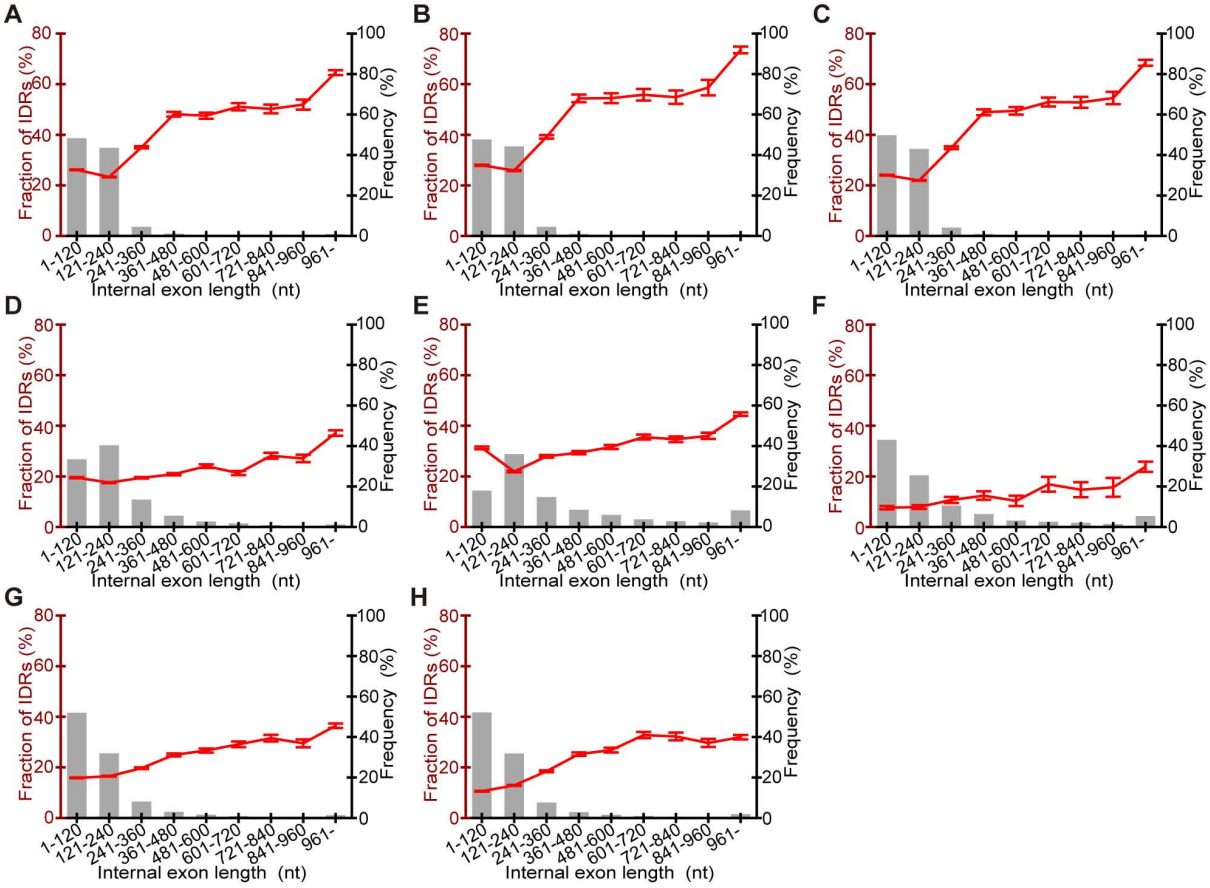

Figure 1. Long internal exons constitute a minority and tend to encode IDRs. The frequencies of occurrence of internal exons in the indicated length ranges are shown by bar graphs (right axis), while the mean fractions of IDRs encoded by the internal exons in each length range together with standard errors of the mean (SEMs) are displayed by line graphs (left axis). 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*.

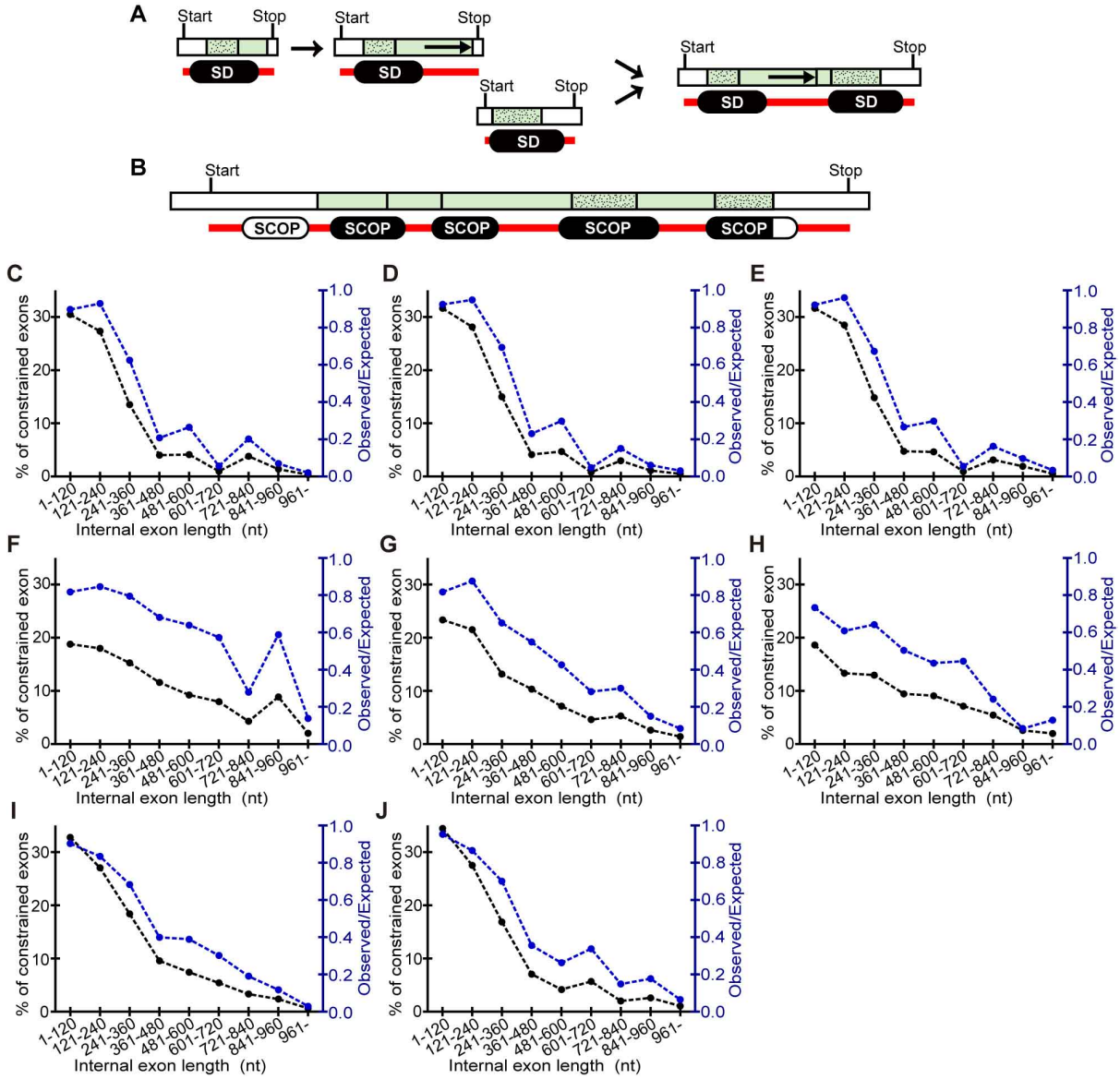

Figure 2. The small bang model and the tendency of longer internal exons to contain less constrained exons. (**A**) The small bang model. The terminal and internal exons are shown as white and pale green rectangles, respectively, with the constrained exons stippled. The IDRs and SDs in the encoded proteins are respectively presented as red bars and rounded black rectangles. The 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 (black line, left axis) and the ratio of the observed fraction to the expected fraction (blue line, right axis) 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*.

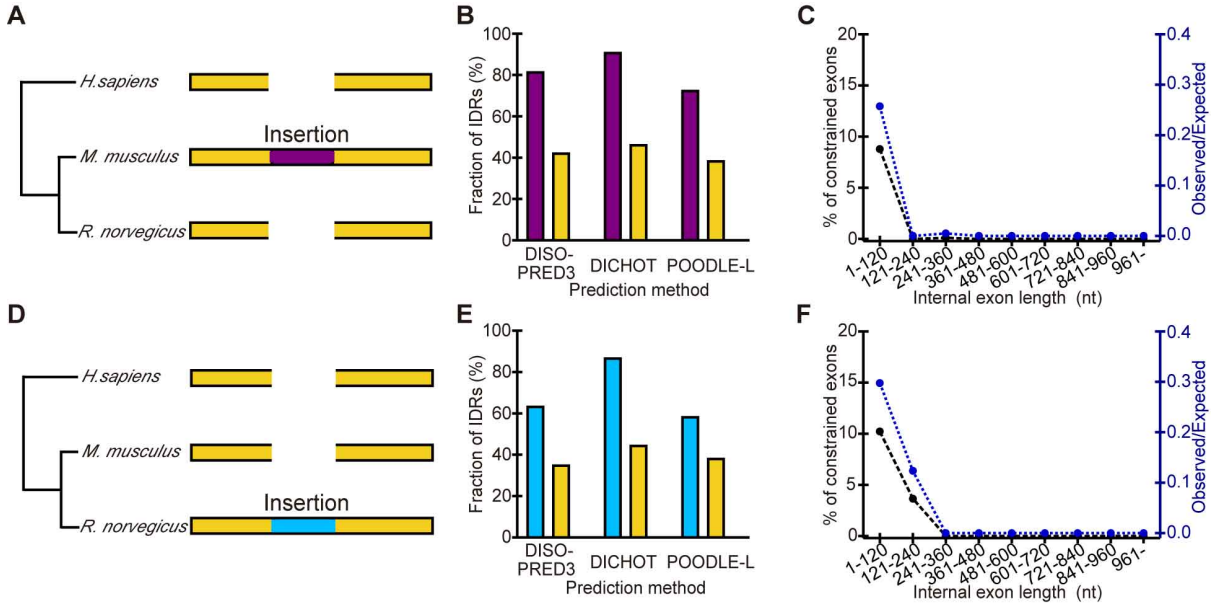

Figure 3. Segments inserted after the rat-mouse divergence encode mostly IDRs and preferentially occur in unconstrained exons. (**A**-**C**) Mouse-specific insertions. (**D**-**F**) Rat-specific insertions. (**A**, **D**) A scheme to identify probable inserted segments in internal exons. (**B**, **E**) The fraction of IDRs in the inserted segments (purple and sky blue, respectively) and that in the rest of internal exons of all genes (yellow) by diffrent prediction methods. (**C**, **F**) The observed fraction of constrained exons in the inserted segments (black line, left axis) and the ratio of the observed fraction to the expected fraction (blue line, right axis) in each length range of internal exons. The expected fractions are those of all internal exons.

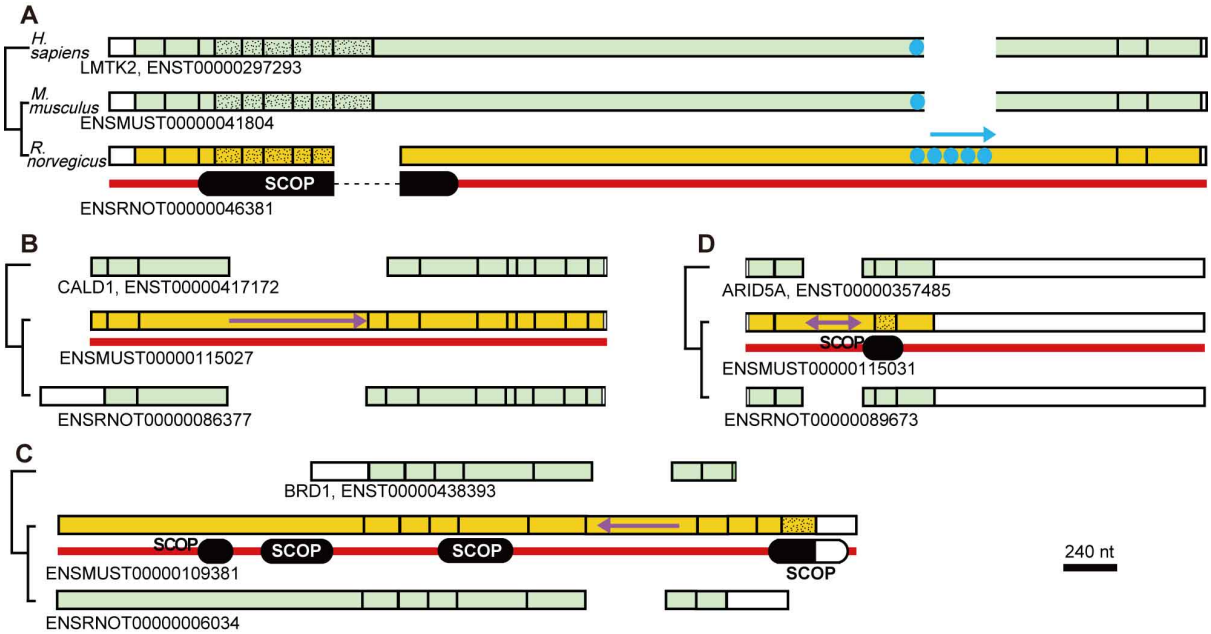

Figure 4. Actual examples of elongation in internal exons. Human, mouse and rat orthologues are drawn with corresponding exons aligned vertically on the indicated scale. The same phylogenetic tree was added to the left of each pane. The coding regions of the terminal exons (white), those with insertion in internal exons (yellow), and the rest of internal exons (pale green) of orthologues are shown with the constrained exons stippled. The single-headed arrows signify the probable directions of expansion, while the double-headed arrow indicates an exonised intron. The transcript identifications 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.

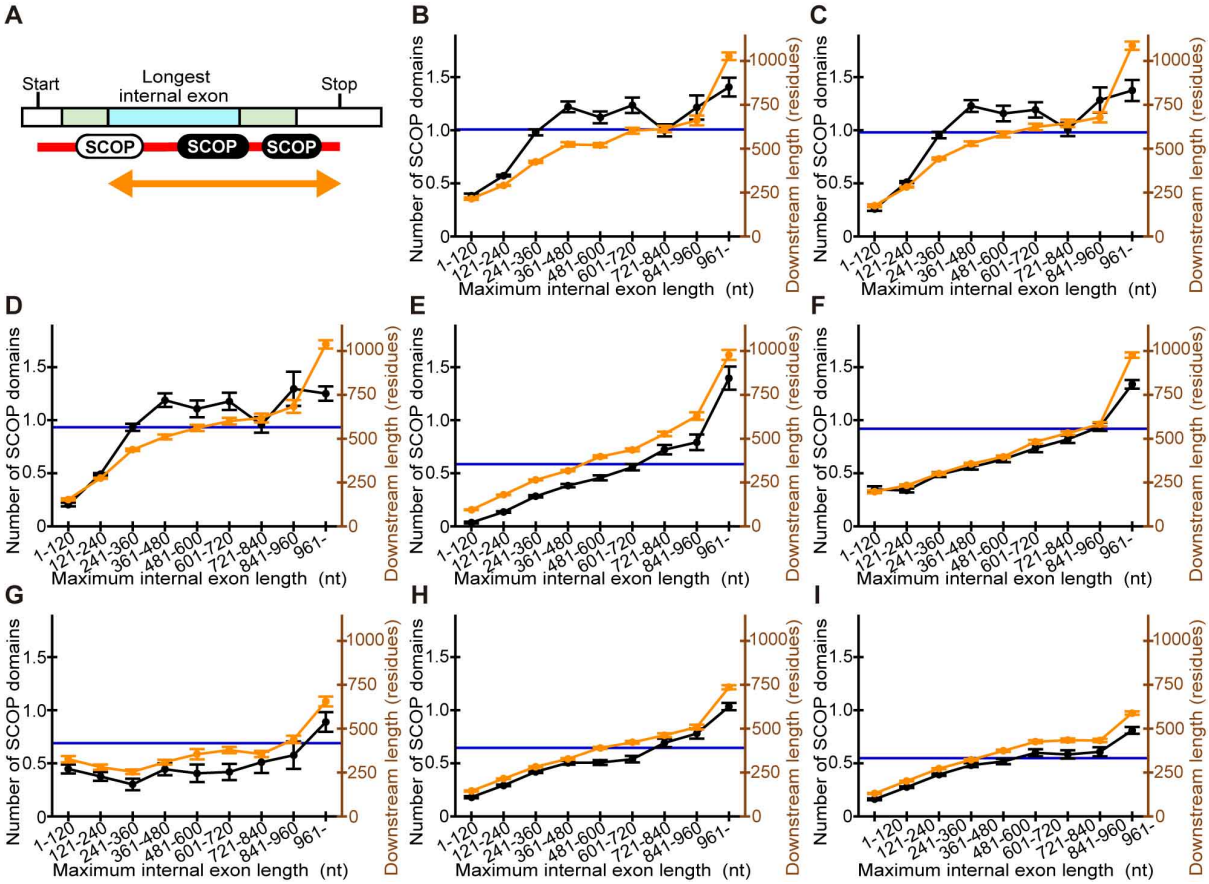

Figure 5. Genes with longer internal exons tend to encode proteins with more SDs. (**A**) A hypothetical case to explain the procedure. The number of SCOP domains in the region indicated by the orange arrow is counted, that is, only those SCOP domains (black) encoded entirely by the longest internal exon and the downstream exons are counted. The sky-blue rectangle signifies the longest internal exon, while the pale green ones represent the other internal exons. The mean number of SCOP domains with SEM in genes with each range of maximum internal exons (black, left axis), the expected number of SCOP domains (blue, left axis) and the mean length of encoded proteins with SEM (orange, right axis) in (**B**) *H. sapiens*, (**C**) *M. musculus*, (**D**) *R. norvegicus*, (**E**) *C. elegans*, (**F**) *D. melanogaster*, (**G**) *S. pombe*, (**H**) *A. thaliana* and (**I**) *O. sativa*.