PNAS

Main Manuscript for

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

Satoshi Fukuchi[a], Tamotsu Noguchi[b], Hiroto Anbo[a], and Keiichi Homma[a, 1]

[a] Program for Information Systems, Division of Informatics, Bioengineering and Bioscience, Maebashi Institute of Technology; 460-1 Kamisadori-machi, Maebashi-shi, 371-0816 Japan, [b] Pharmaceutical Education Research Center, Meiji Pharmaceutical University; 2-522-1 Noshio, Kiyose, Tokyo, 204-8588 Japan

[1]To whom correspondence should be addressed. Tel: +81-27-265-7334; Fax: +81-27-226-5168.

Email: khomma@maebashi-it.ac.jp

Author Contributions: S.F. and K.H. designed research; H.A. and K.H. performed research; T.N. provided analytical tools; K.H. analyzed data; and T.N. and K.H. wrote the paper.

Competing Interest Statement: The authors declare no competing interest.

This PDF file includes:

Main Text
Figures 1 to 5

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 from the proteomic viewpoint: the protoeukaryotic genes had no introns and mostly encoded one SD each, but a majority of them subsequently divided into multiple exons (step 1). Many exons unconstrained by SDs elongated to encode IDRs (step 2). The elongated exons encoding IDRs frequently facilitated the acquisition of multiple SDs to make the last common ancestor of eukaryotes (step 3). One prediction of the model is that long internal exons are mostly

unconstrained exons. Analytical results of the eight eukaryotes are consistent with this 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. The model also predicts that SDs followed by long internal exons tend to have other SDs downstream. This prediction was also verified in all the eukaryotic species analyzed. Our model accounts for the dichotomy between prokaryotic and eukaryotic proteins and proposes a selective advantage conferred by IDRs.

Significance Statement

Eukaryotic proteins generally consist of structural domains and intrinsically disordered regions that do not form unique three-dimensional structures by themselves. We propose a model for the evolution of the last common eukaryotic ancestor from the primordial eukaryote. According to our model, many internal exons unconstrained by structural domains elongated to encode intrinsically disordered regions, and then the added intrinsically disordered regions facilitated recombination to produce genes encoding multiple structural domains. Our model explains why typical eukaryotic proteins consist of multiple structural domains and many intrinsically disordered regions, while most prokaryotic proteins contain only one structural domain with little intrinsically disordered regions. Here, we present data from eight eukaryotes of diverse branches that support our model.

Main Text

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 exonization of introns (12-14). While prokaryotic proteins mostly consist of a single SD, eukaryotic proteins tend to have multiple SDs (15-17).

We endeavored 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. These exons 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.

Results

The longer internal exons are, the more frequently they encode IDRs. We examined the internal exons of the 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 (Fig. 1). Interestingly, we found that the longer internal exons are, the more frequently they encode IDRs (predicted by DISOPRED3 (22)) in all species (Fig. 1). The correlation coefficients are all positive and significantly different from 0 (*SI Appendix*, 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 (23) and POODLE-L (24). We found a positive correlation without exception irrespective of IDR prediction methods (*SI Appendix*, Figs. S1 and S2), although three correlations are not statistically significant from 0 at $P<0.001$ (*t*-test, *SI Appendix*, Fig. S1*D*, Fig. S2 *D* and *F*, and 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 (*SI Appen*dix, Fig. S3). The inclusion of long internal exons is thus unlikely to be universally mediated by the SRSF3 splicing factor.

Genes with long internal exons preferentially encode 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 (25), a high-quality protein database. The genes with internal exons longer than 720 nt are annotated with the following GO terms significantly more often (chi-square test, $P<0.05$, the number of genes in each range is shown in *SI Appendix*, Table S2) in all the species examined: 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, exonization of introns, and repeat sequence expansion are possible mechanisms of internal exon elongation. What kind of primordial short exons expanded in length? An internal exon that encodes a portion of an SD (stippled in Fig. 2*A*), 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 thought it likely that the added sections encoding IDRs facilitated recombination to give rise to multiple SDs: firstly, IDRs probably facilitate recombination 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 (26-28).

We therefore propose the small bang model from the proteomic viewpoint (Fig. 2*A*): the genes of the protoeukaryote had no introns and most of them encoded a single SD, just like prokaryotic genes. Most genes subsequently divided into exons (step 1). Then a plethora of unconstrained exons expanded with the addition of segments that encode IDRs (step 2). 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 (step 3). As they conferred a selective advantage, these events occurred multiple times to produce the LECA. According to this model, the LECA proteome was comprised mostly of proteins having multiple SDs and a significant fraction of IDRs, accounting for the dichotomy between prokaryotic and eukaryotic proteins.

Long internal exons are mostly unconstrained exons. Our model would predict that long internal exons in extant eukaryotes are mostly unconstrained exons, that is, the longer internal exons are, the smaller the observed fraction of constrained exons will be. To test this prediction,

we used SCOP domains (29) in our analysis (Fig. 2*B*), 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 (Fig. 2*C*-*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 (*SI Appendix*, Table S3) since SCOP domains longer than the region encoded by long exons are rare (Materials and Methods). Presumably "latent" SCOP domains, i.e., those that have not been identified, would increase both the observed and expected fractions to the same extent, and would essentially not affect our 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 before the emergence of the LECA is difficult, we searched for cases of internal exon elongation after the rat-mouse divergence by comparison of human, mouse, and rat orthologs. (The generally high reliability of human, rat, and mouse data prompted us to analyze the three species.) We identified the segments in internal exons that exist only in the rat or mouse genome 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 obtain segments inserted within internal exons (purple in Fig. 3*B* and sky blue in Fig. 3*F*), we removed inserted segments that coincide with entire exons (gray in Fig. 3*A* and *E*). Although this method cannot capture all the cases of internal exons elongated after the rat-mouse divergence, the selected cases are probably genuine ones.

A majority of insertions turned out to be in short (1- 360 nt) internal exons (*SI Appendix*, 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 (Fig. 3*C* and *G*). The fractions of constrained exons are extremely low and are in all cases less than or equal to the expected values (Fig. 3*D* and *H*, *SI Appendix*, Table S3).

Four examples of segments inserted in internal segments are presented (Fig. 4). Fig. 4*A* 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). Fig. 4*B* 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 Fig. 4*C* 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. 4*D*) 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 (30). 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.

An SD followed by a long internal exon frequently has another SD downstream. In step 3 of the small bang model, many genes with long internal exons acquired SDs downstream (Fig. 2*A*). This would predict that genes encoding an SD tend to encode another SD downstream if a long internal exon exists downstream of the exon(s) encoding the first SD. To test this prediction, we identified segments between SCOP domains (group A, indicated by the orange arrow in Fig. 5*A*) with the expectation to frequently find long internal exons in the segments that facilitated acquisition of the downstream SCOP domains. For comparison, we chose the segment after the last SCOP domain in each gene (group B, the blue arrow in Fig. 5*A*), as it is predicted not to contain a long internal exon at high frequency as it has not facilitated the obtainment of SCOP domain(s) downstream. If the model were correct, the length of the longest exon in group A would tend to be longer than that of group B.

We determined the longest exon in each group A segment in genes with SCOP domain(s) and calculated the mean length in each range of segment length (orange in Fig. 5 *B-I*). We also identified the longest exon in the group B segment in all genes with SCOP domain(s) and computed the mean length in each bin of distance from the last SCOP domain (blue in Fig. 5 *B-I*). The means of a distance bin are plotted only if the numbers of both groups are more than 20 (*SI Appendix*, Table S5). (Mainly due to the rarity of internal exons in *S. pombe* genes, the number of analyzable cases in this species is small.) Without exception, in all the eight species examined, the longest internal exons are longer if a SCOP domain is followed by aother SCOP domain. The differences of the weighted averages of the two groups are statistically significant at $P<0.001$ (*t*-test) except for *R. norvegicus* and *S. pombe* which are significantly different at $P<0.05$ (*t*-test). (Since sections of group A tend to be shorter than those in group B (*SI Appendix*, Table S5), the differences in the weighted averages the groups are generally smaller than those of individual distance ranges.) This result comports with our model. One concern is that identification of novel

SCOP domains may change the result. To assess the effect of latent SCOP domains, we carried out the same analyses with half of SCOP domains randomly neglected. Although the number of samples is reduced (*SI Appendix*, Table S6), the result is essentially the same as before (*SI Appendix*, Fig. S4). Thus, the identification of latent SCOP domains would probably not affect the conclusion.

Discussion

The small bang model accounts for the dichotomy between prokaryotic and eukaryotic proteins: eukaryotic proteins have a higher fraction of IDRs and 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* (Fig. 1*F*), 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 are enriched with those with nuclear protein functions. This observation comports 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 an evolutionary advantage.

Secondly, the model is consistent with the observed low frequency of constrained exons in long internal exons in all eight eukaryotes analyzed. 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 our model.

Lastly, we found in all the eight eukaryotes examined that genes encoding an SD tend to have another SD downstream if a long internal exon exists in the downstream section. This result is also in agreement with 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 (step 3 in Fig. 2*A*) occurred prior to the expansion of unconstrained exons (step 2 in Fig. 2*A*). However, we consider our 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 a 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 our 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 after the LECA, as they no longer gave a significant selective advantage. The putative evolutionary advantage provided by introns through attenuation of R-loop formation (7) is probably minimal as introns have been entirely lost in a number of extant eukaryotes. On the other hand, the existence of introns in early eukaryotes made it possible for many internal exons to expand, which conferred a selective advantage, according to the small bang model. Our 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 novel combinations of SDs that do not exist in prokaryotes. Future research will hopefully test and further refine the model.

Materials and Methods

All the sequence, exon and ortholog data used in this study were downloaded from the Ensembl database (30). Though we analyzed all variants except for the analyses of the segments inserted after the rat/mouse split, we made the exons non-redundant by in-house C programs to avoid double counting. IDRs were predicted by DISOPRED3 (22), DICHOT (23), and POODLE-L (24), while SCOP domains (29) were assigned in an intermediate step in DICHOT prediction.

Using another 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 (32) alignments of orthologous protein sequences with default settings. We wrote a C program to assess enrichment of GO terms according to SwissProt/UniProt (25) annotations (release 2021_04).

A different in-house C script identified the longest internal exon among candidate exons and calculated the mean length and SEM in each range of distance from the last SCOP domain. In this program, an internal exon is considered as a candidate of the longest internal exon of group

A if the exon encodes the C-terminal end of the last SCOP domain. In group A, an internal exon is also considered as a candidate if the beginning of the exon encodes a residue between contiguous SCOP domains. An internal exon is included as a candidate of the longest internal exon of group B if the last residue the exon encodes is downstream of the last SCOP domain. Genes encoding no SCOP domains are neglected in both groups.

Acknowledgments

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

References

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. *Nucleic 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. UniProt Consortium, UniProt: the universal knowlegebase in 2021. *Nucleic Acids Res.* **49**, D480-D489 (2021).
26. R. A. Sturm, W. Herr, The POU domain is a bipartite DNA-binding structure. *Nature* **336**, 601-604 (1988).
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. U.S.A.* **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).

Figures

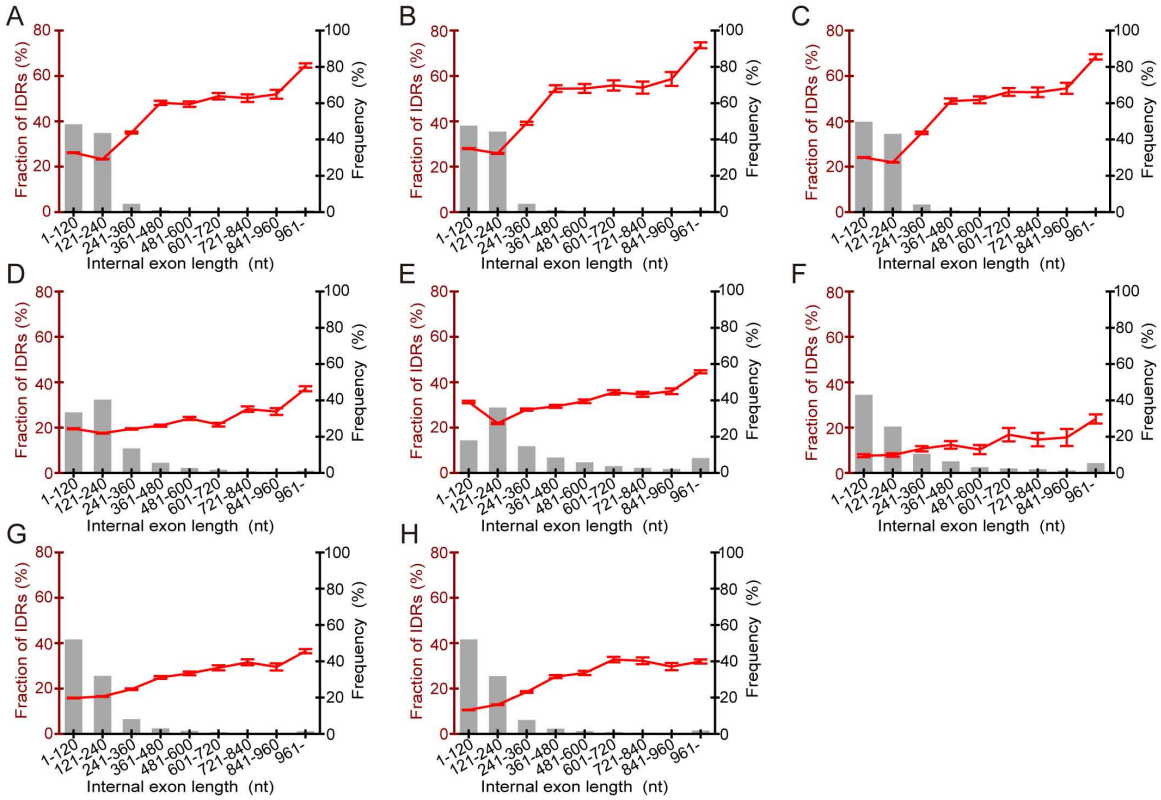

Fig. 1. Long internal exons constitute a minority and tend to encode IDRs. The frequencies of occurrence of internal exons in the indicated length ranges are shown by bar graphs (right axis), while the mean fractions of IDRs encoded by the internal exons in each length range together with standard errors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

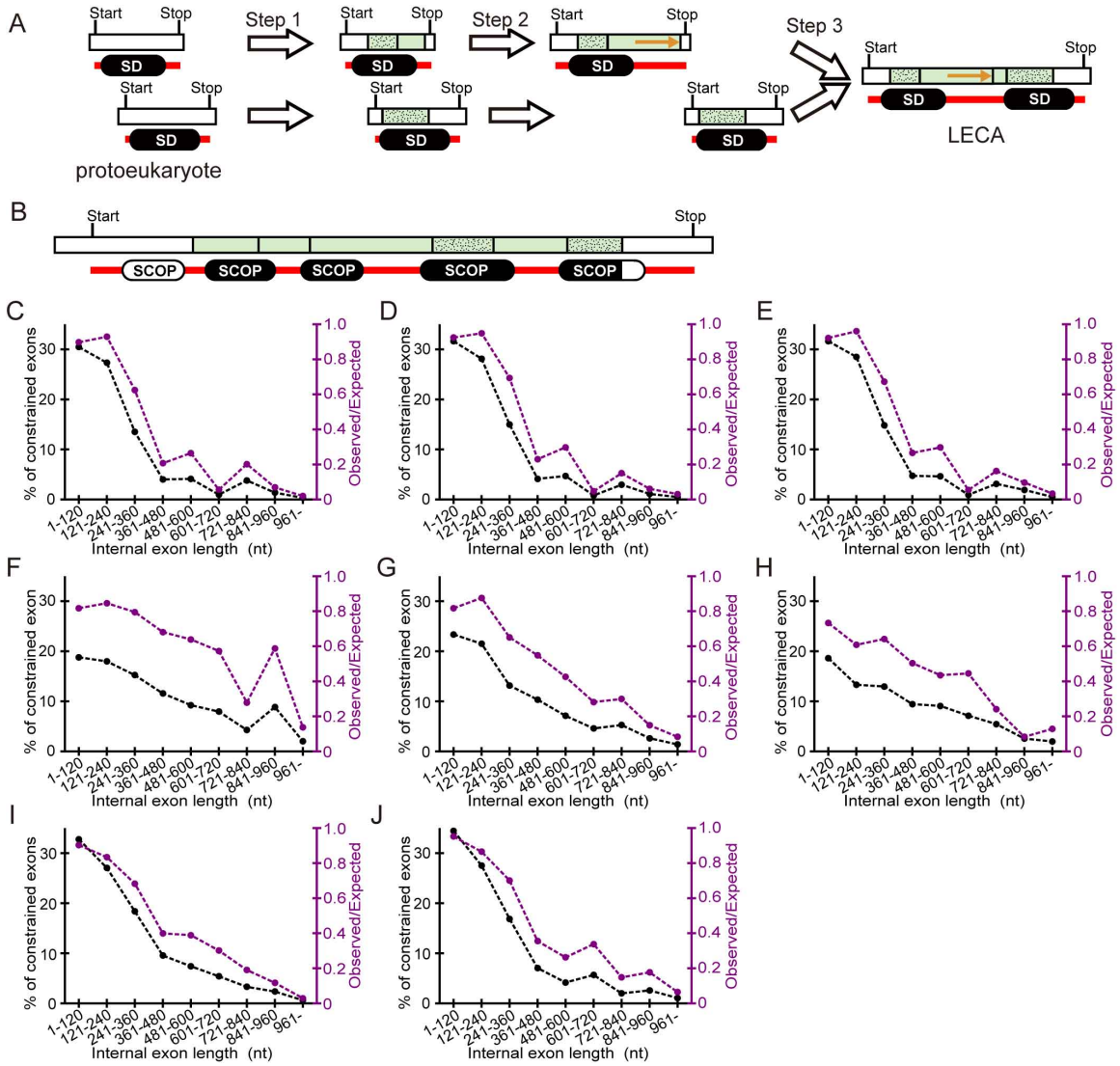

Fig. 2. The small bang model and the tendency of longer internal exons to contain less constrained exons. *(A)* The small bang model. The terminal and internal exons are shown as white and pale green rectangles, respectively, with the constrained exons stippled. The IDRs and SDs in the encoded proteins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

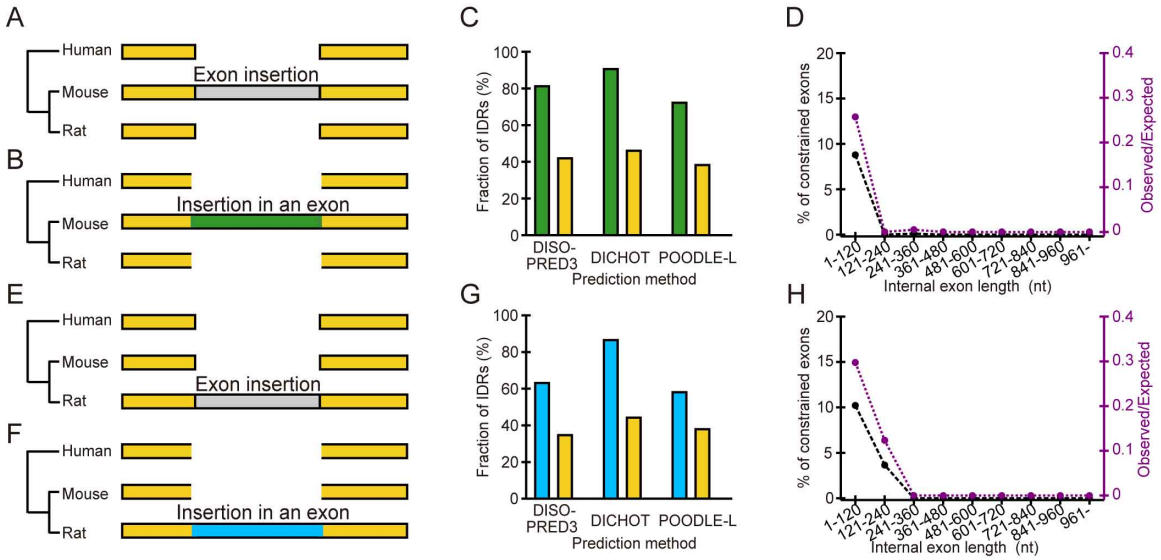

Fig. 3. Segments inserted in internal exons after the rat-mouse divergence encode mostly IDRs and preferentially occur in unconstrained exons. (*A*-*D*) Mouse-specific insertions. (*E*-*H*) Rat-specific insertions. (*A, B*, *E*, and *F*) The rectangles represent internal exons. The fraction of IDRs in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

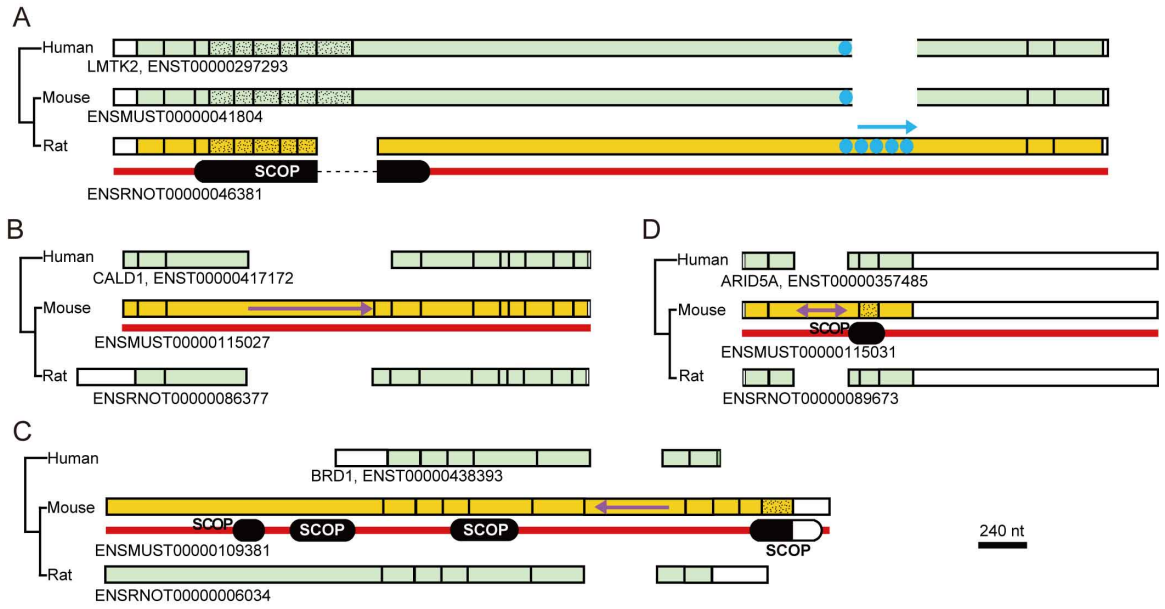

Fig. 4. Examples of elongated internal exons. Human, mouse, and rat orthologs are drawn with corresponding exons aligned vertically on the indicated scale. The same phylogenetic tree was added to the left of each pane. The coding regions of the terminal exons (white), those with insertion in internal exons (yellow), and the rest of internal exons (pale green) of orthologs 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 *B*.

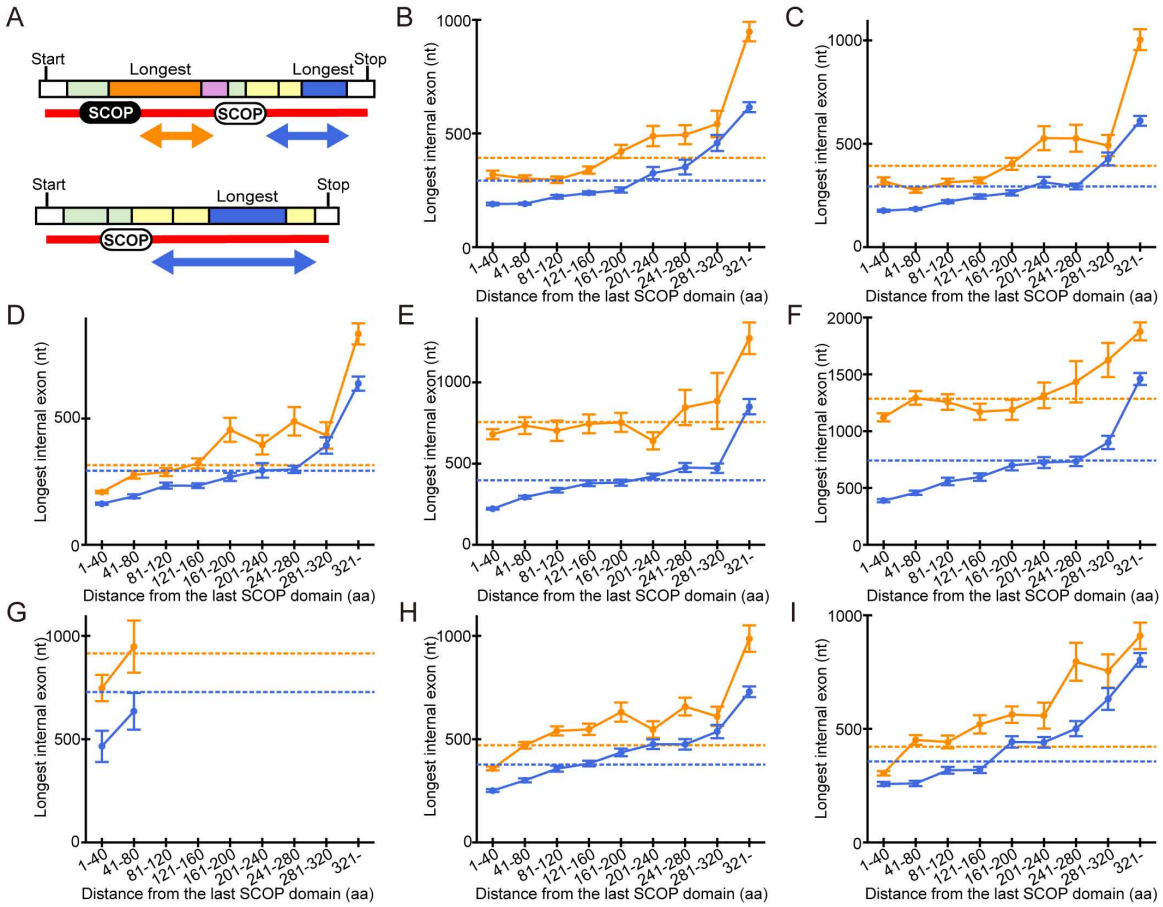

Fig. 5. A SCOP domain followed by a long internal exon frequently has another SCOP domain downstream. (*A*) Hypothetical cases to explain the procedure. Internal exons are colored. If multiple SCOP domains exist, the longest internal exon among the candidates (orange and pink rectangles) in the inter-SCOP region (orange arrow) is identified (group A). In the regions downstream of the last SCOP domain (blue arrows), the longest internal exons among the candidates (yellow and blue rectangles) are determined (group B). (*B*-*I*) The mean lengths and SEMs of the longest internal exons in group A (orange) and B (blue) segments are shown as solid lines if the numbers of both groups are in excess of 20. The orange and blue dotted lines represent the weighted averages of group A and B, respectively. The species plotted are (*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*.