**Article**

# Elongated exons encoding disordered regions expedited eukaryotic evolution

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

[1]Department of Life Science and Informatics, 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: Keiichi Homma, Department of Life Science and Informatics, Maebashi Institute of Technology, 460-1 Kamisadori-machi, Maebashi-shi, 371-0816 Japan, Tel.: +81 27 265 7360, Fax: +81 27 226 5168, E-mail: khomma@maebashi-it.ac.jp.

**Abstract**

**Most prokaryotic proteins consist of one structural domain (SD), while a typical eukaryotic protein is comprised of multiple SDs and intrinsically disordered regions (IDRs) that by themselves do not adopt stable structures[1-4]. In eukaryotes IDRs are especially prevalent in transcription factors[5] and are involved in condensates that regulate diverse cellular functions[6]. How eukaryotic proteins evolved to differ from prokaryotic proteins has not been fully elucidated. Here we find that the longer internal exons are, the more frequently they encode IDRs in general. Moreover, the boundaries of long internal exons tend to fall outside of structural domains. These findings are consistent with the idea that some internal exons whose boundaries were aligned outside of structural domains acquired sections encoding IDRs. Most of the internal exon segments that probably lengthened after the rat-mouse divergence encode IDRs, supporting the view. We surmise that many exons whose boundaries were not aligned to the inside of structural domains elongated to add IDRs to the encoded proteins, often conferring new functions. Some of the IDRs possibly served as spacers in acquiring structural domains through recombination. We propose that the 'small bang' that lengthened exons facilitated the evolution that led to the last common ancestor of eukaryotes.**

## Main text

The last eukaryotic common ancestor (LECA) probably had an intron-rich genome[7,8], but massive intron losses subsequently occurred in a few branches of eukaryotes[9]. Alternative splicing produces multiple variants from one gene and can explain the variation in complexity among eukaryotes[10] and approximately 95% of multiexon human genes undergo alternative splicing[11]. Introns, regarded as selectively disadvantageous, proliferated in the LECA most plausibly by meiotic processes[12]. 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[13]. 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 eukaryotic proteins consist of IDRs[1]. IDRs in eukaryotic proteins are especially prevalent in transcription factors[5] and are involved in condensates that regulate diverse cellular functions[6]. IDRs arise de novo by expansion of repetitive DNA sequences[14] and exonisation of introns[15-17]. While prokaryotic proteins mostly consist of one SD, eukaryotic proteins tend to have multiple SDs[2-4]. We endeavour to present 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 large exons enriched with C-nucleotides in vertebrates[20] and plays important roles in the occurrence and development of tumors[21]. Intriguingly, large exons in vertebrates frequently encode IDRs[20]. We here show that this tendency is generally observed in eukaryotes and thus is likely to have been present in the LECA. We argue that exons whose boundaries do not align with structural domains elongated with the expanded section encoding IDRs, which made it possible to acquire more structural domains, and thus produced the LECA.

### Long internal exons frequently encode IDRs

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 more than half of them shorter

than 241 nt in all the genomes (Fig. 1a-h). To test if same conclusions are reached irrespective of IDR prediction methods, we used three programs: DISOPRED3[22], DICHOT[23] and POODLE-L[24]. Interestingly, the longer internal exons are, the more frequently they encode IDRs in all species including non-vertebrates with all the correlations statistically significant if DISOPRED3 results are used (Fig. 1a-h, $p<0.001$, *t*-test). The phenomenon is generally observed irrespective of the three IDR prediction methods used (Fig. 1i, Extended Data Figs. 1, 2), although some correlations are not significant at $p<0.001$ (Extended Data Figs. 1d, f and 2d).

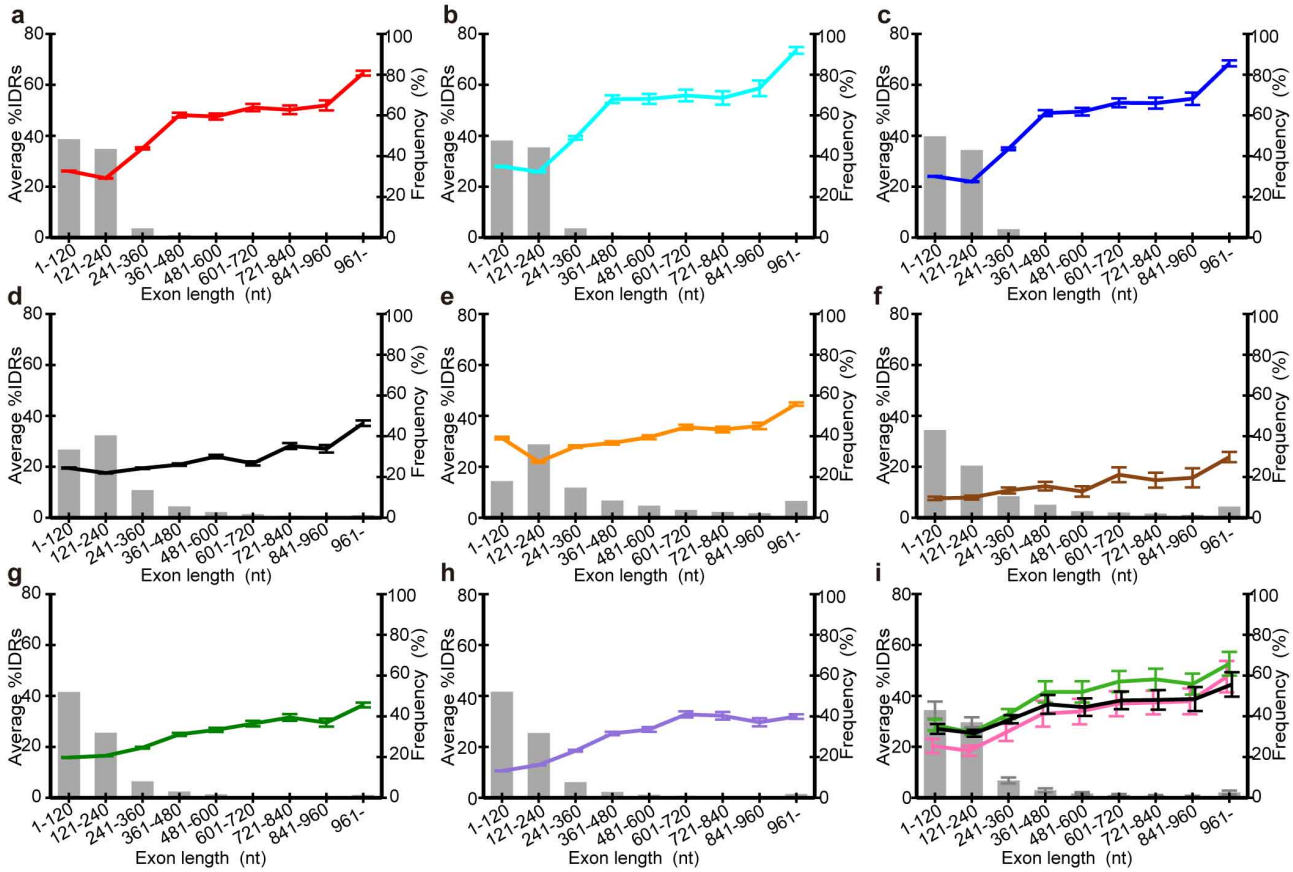
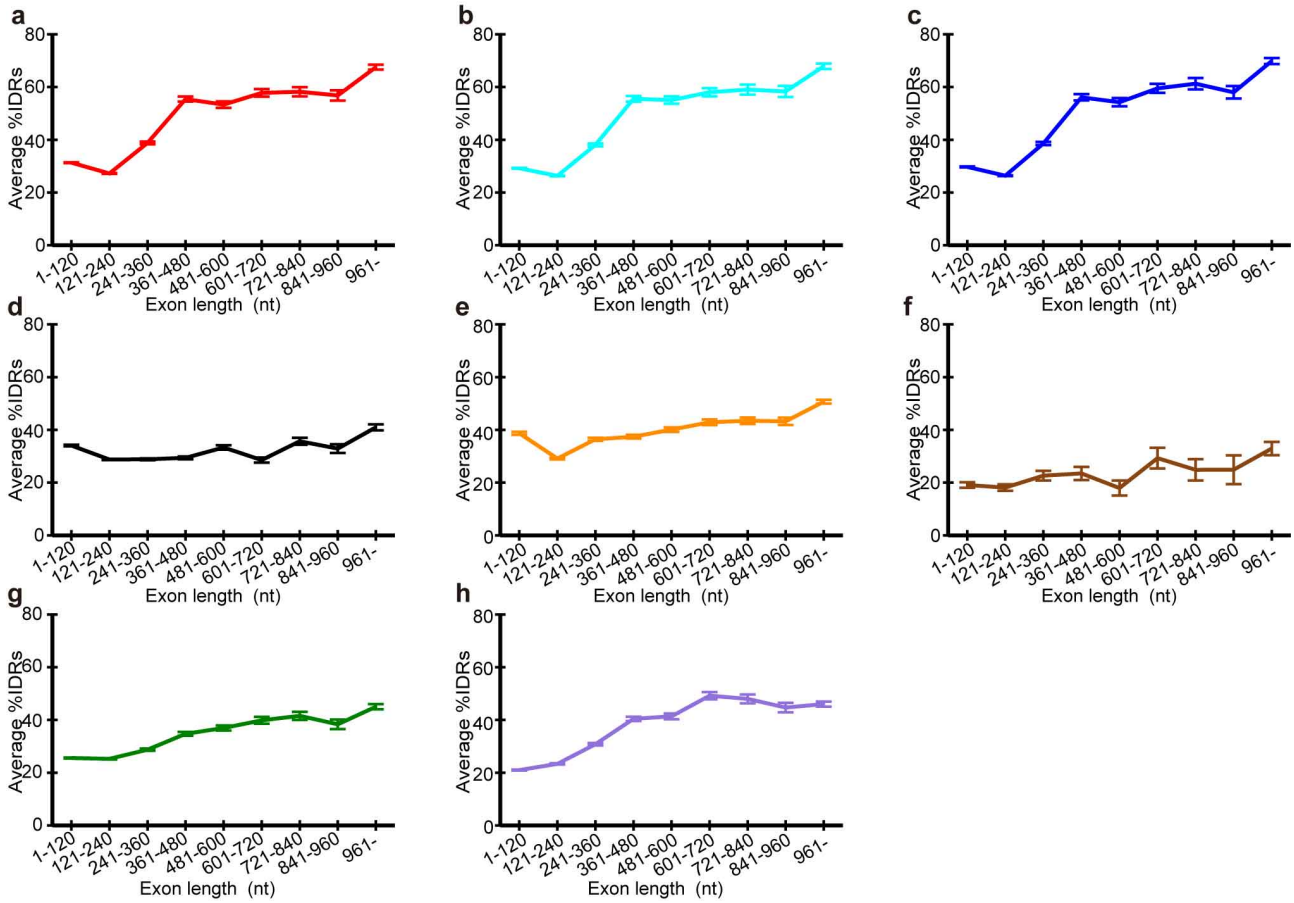
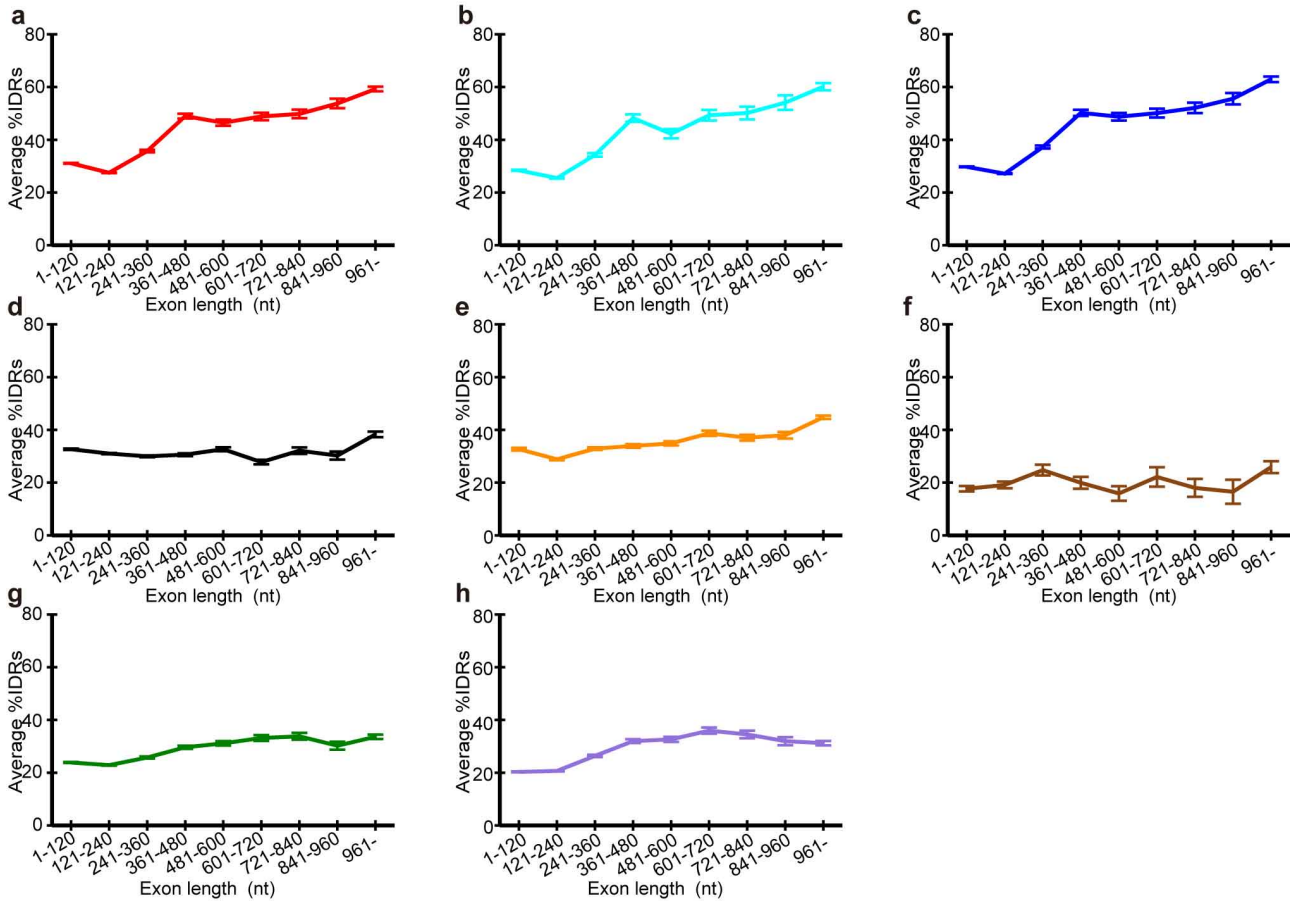



**Fig. 1| Long internal exons constitute a minority and tend to encode IDRs.** The frequencies of occurrence of internal exons in the indicated length ranges are shown by bar graphs, while the average fractions of IDRs encoded by the internal exons in the length rages with standard errors of the mean (SEMs) are displayed by line graphs. **a-h,** IDRs by DISOPRED3. **a,** *H. sapiens*. **b,** *M. musculus*. **c,** *R. norvegicus*. **d,** *C. elegans*. **e,** *D. melanogaster*. **f,** *S. pombe*. **g,** *A. thaliana*. **h,** *O. sativa*. **i,** Averages and SEMs of the eight eukaryotes. IDRs were predicted by DISOPRED3 (pink), DICHOT (green) and POODLE-L (black).

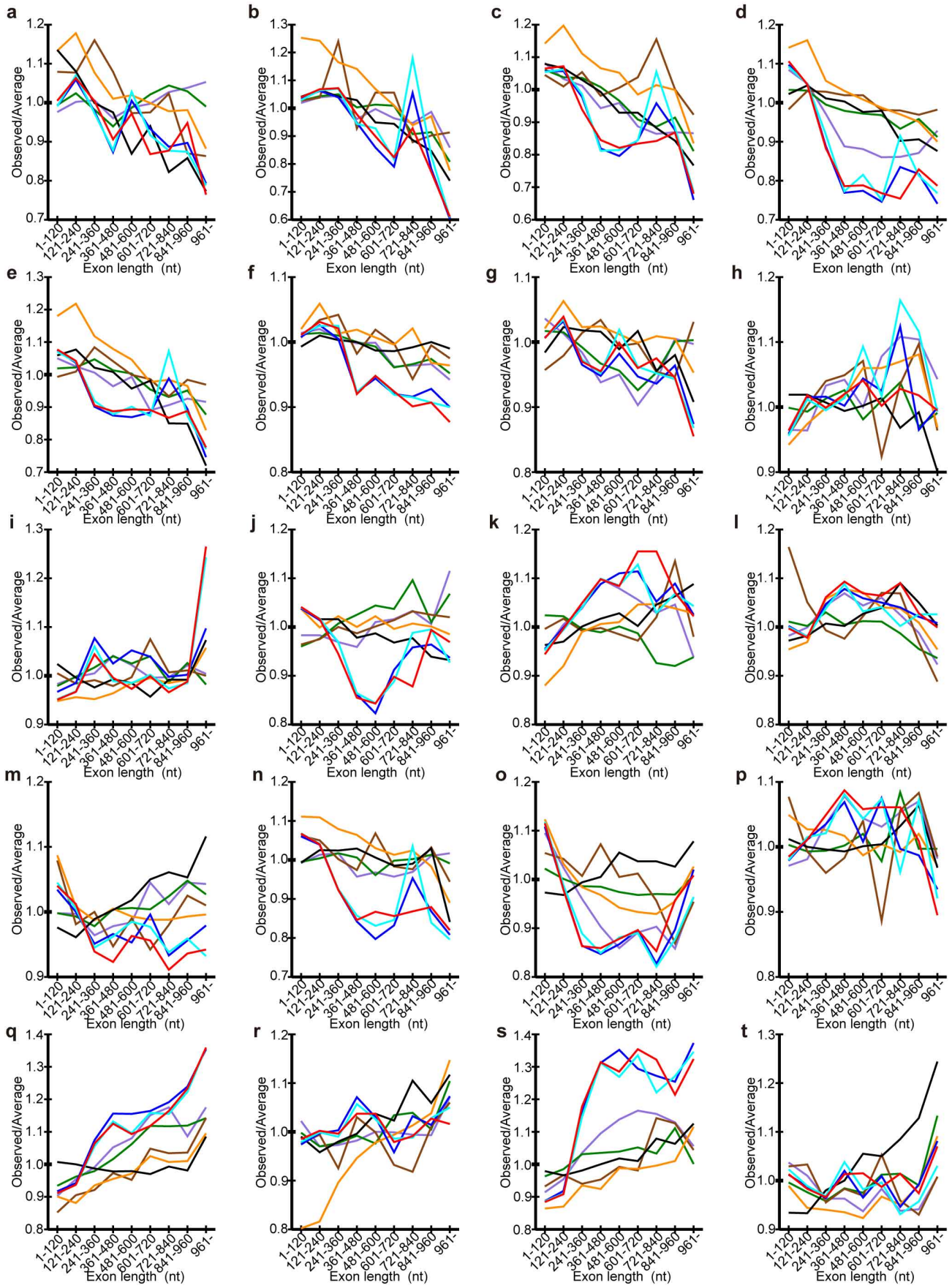



**Extended Data Figure 1| DICHOT results show that long internal exons tend to encode IDRs.** The average fractions of IDRs encoded by the internal exons in the length rages with SEMs are displayed by line graphs. IDRs were predicted by DICHOT. The data of the same species as in Fig. 1 are presented.



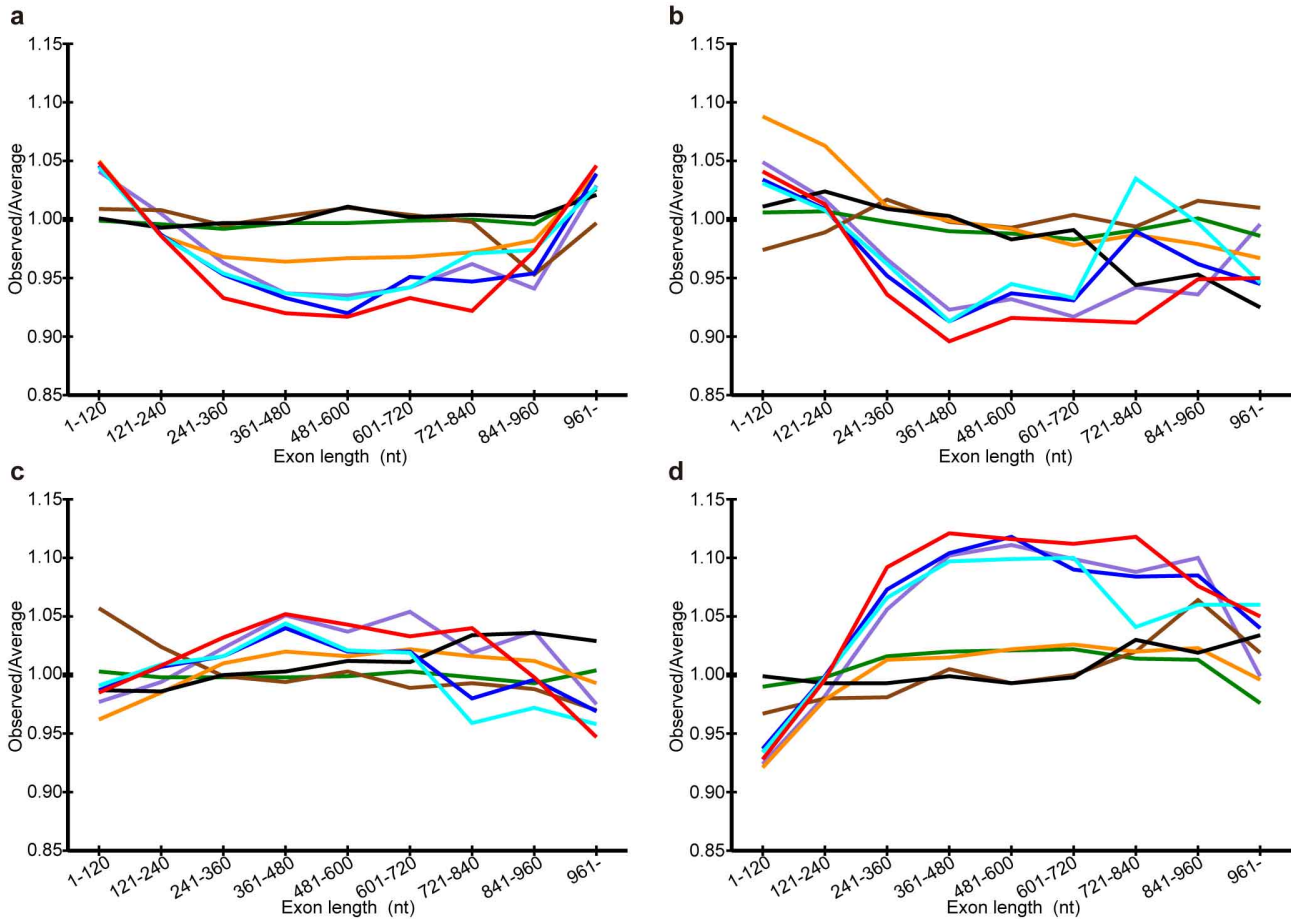

**Extended Data Figure 2| POODLE-L results show that long internal exons tend to encode IDRs.** The average fractions of IDRs encoded by the internal exons in the length rages with SEMs are displayed by line graphs. IDRs were predicted by POODLE-L. The data of the same species as in Fig. 1 are shown.

Consistently, order-promoting amino acids[23] encoded less often than expected, while disorder-promoting amino acids[23] are encoded more by longer internal exons (Extended Data Fig. 3): the seven order-promoting amino acids cysteine, tryptophan, tyrosine, isoleucine, phenylalanine, valine and leucine showed negative correlations of relative occurrence and exon length, while the four disorder-promoting amino acids, serine, glutamine, proline and glutamate exhibited positive correlations (Extended Data Fig. 3).


a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t
Observed/Average
Exon length (nt)
1-120
121-240
241-360
361-480
481-600
601-720
721-840
841-960
961-

**Extended Data Figure 3| Long internal exons encode order-promoting amino acids less often than expected, while preferentially encode disorder-promoting amino acids.** The amino acids are ordered from the most to the least order-promoting amino acids. The observed frequency of an amino acid encoded by the internal exons in each length bin is divided by the average frequency of the amino acid encoded by all the internal exons in the species. The ratios in each species (*H. sapiens* (red), *M. musculus* (cyan), *R. norvegicus* (blue), *C. elegans* (black), *D. melanogaster* (orange), *S. pombe* (brown), *A. thaliana* (green) and *O. sativa*. (purple)) are plotted. **a,** cysteine. **b,** tryptophan. **c,** tyrosine. **d,** isoleucine. **e,** phenylalanine. **f,** valine. **g,** leucine. **h,** histidine. **i,** threonine. **j,** asparagine. **k,** alanine. **l,** glycine. **m,** aspartate. **n,** methionine. **o,** lysine. **p,** arginine. **q,** serine. **r,** glutamine. **s,** proline. **t,** glutamate.

Though longer internal exons of most species displayed the tendency to have higher cytidine contents, confirming the observation on human exons[20], those of *A. thaliana* did not (Extended Data Fig. 4).




**Extended Data Figure 4| Long internal exons in most species have high cytidine contents.** The observed frequency of a nucleotide in the internal exons in each length bin is divided by the average frequency of the nucleotide in all the internal exons in the species. The ratios in each species (coloured as in Extended Data Figure 3) are shown in line graphs. **a,** adenine. **b,** thymidine. **c,** guanosine. **d,** cytidine.

**Genes with long internal exons preferentially encode nuclear proteins**

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, $p<0.05$) 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 are related to characteristics of nuclear proteins. The same analysis of the genes with extremely long exons (>960 nt) identified the 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 one can be considered as traits of nuclear proteins.

**The boundaries of long internal exons tend to fall outside of structural domains**

We thought it plausible that elongation of short exons produced most of long internal exons. Exons whose boundaries are aligned to the inside of SDs are not likely to be extended at the boundaries as the segment added to the exon boundaries would insert residues to SDs, oftentimes destabilising the structure. By contrast exons whose boundaries fall outside of SDs can elongate freely, with most of the extended segments encoding IDRs. We thus tested if the boundaries long internal exons tend to fall outside of SDs. For this analysis, we used SCOP domains[27], because SD boundaries are not easily identifiable from the outputs of prediction programs as a single SD is often artifactually subdivided into multiple segments. Note that the absence of an SCOP domain alignment does not always mean that the segment is an IDR, because an SD whose homologues have not been structurally determined is unclassified by SCOP and thus may be present in the segment. We examined if the two boundaries of internal exons are aligned to the inside of SCOP domains and determined the dependence on the lengths of the internal exons. Indeed, the longer internal exons are, the less often their boundaries fall inside of SCOP domains in all the species examined (Fig. 2).

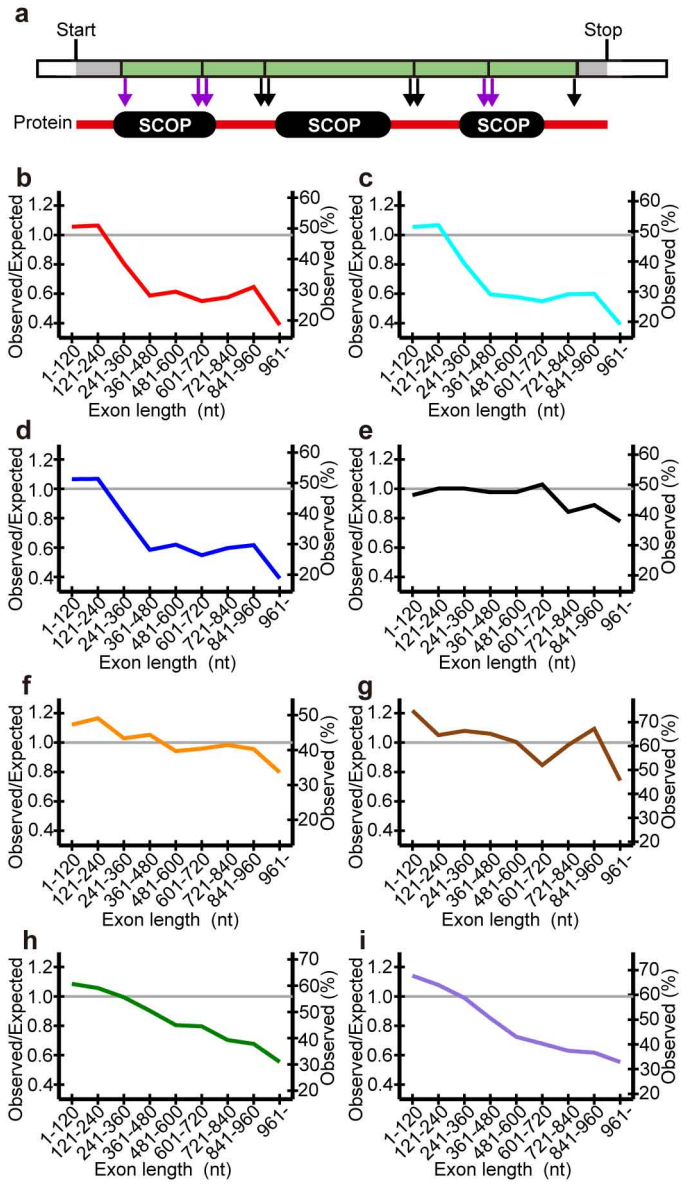
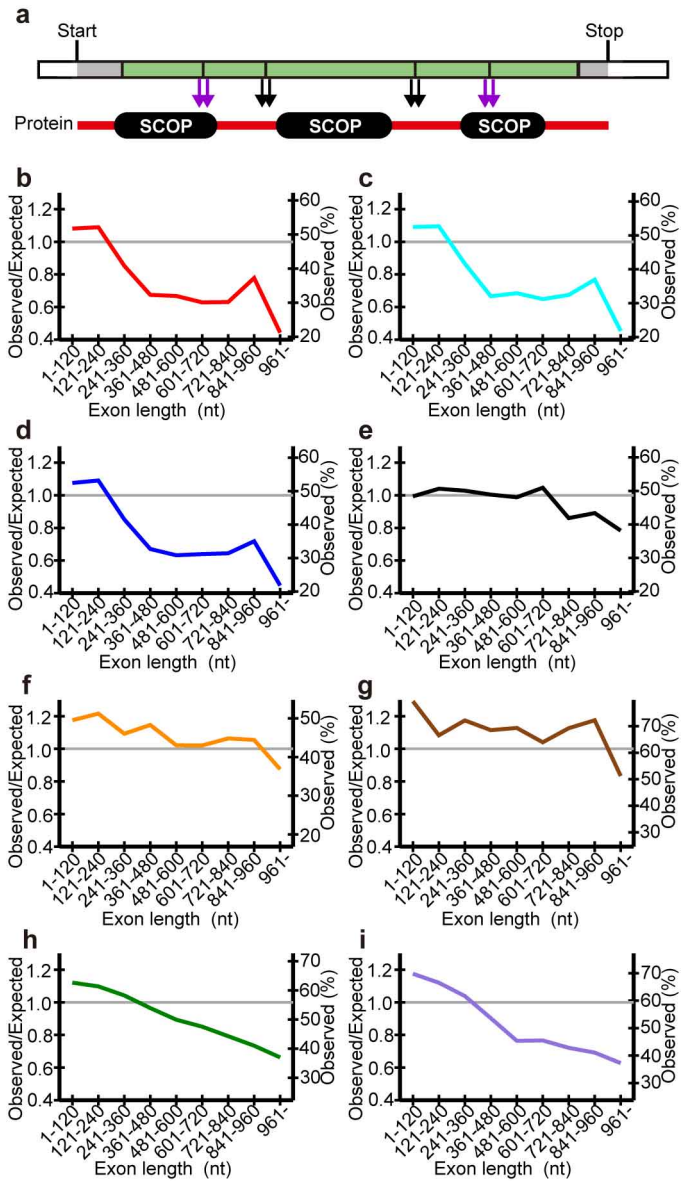



**Fig. 2| The longer internal exons are, the less their boundaries fall inside of SCOP domains. a**. A hypothetical case for explanation. Only the boundaries of internal exons are assessed (arrows). The boundaries are judged if they fall inside of SCOP domains (purple arrows) or not (black arrows). Two arrows are drawn for each boundary to signify the fact that the dependence on the lengths of both the upstream and downstream exons is assessed. **b-i,** The observed fractions of cases in which the exon boundaries are aligned inside of SCOP domains (right scale) and the ratios of the observed fraction to that expected if SCOP domains are randomly placed (left scale). **b,** *H. sapiens*. **c,** *M. musculus*. **d,** *R. norvegicus*. **e,** *C. elegans*. **f,** *D. melanogaster*. **g,** *S. pombe*. **h,** *A. thaliana*. **i,** *O. sativa*.

Moreover, the observed cases of the boundaries aligning within SCOP domains become less than the expected values in extremely long internal exons with no exceptions. Admittedly, the assumption that the 5' boundary of the second exon and the 3' boundary of the penultimate exon as internal exon borders is not completely justified, as they are also the 3' boundary of the 5'-terminal exon and the 5' boundary of the last exon, respectively. Since the exclusion of the boundaries with the 5'- and 3'-terminal exons does not appreciably alter the results (Extended Data Fig. 5), however, we consider this ambiguity immaterial. These observations lend credence to the hypothesis that most long internal exons resulted from elongation of exons whose boundaries do not fall inside of SDs with the added segments mostly encoding IDRs.




**Extended Data Figure 5| The longer internal exons are, the less their boundaries fall inside of SCOP domains even if terminal boundaries are included in analysis. a,** A hypothetical case for explanation. Only the boundaries between internal exons are assessed (arrows). The boundaries are judged if they fall inside of SCOP domains (purple arrows) or not (black arrows). **b-i,** The observed fractions of cases that the exon boundaries are aligned inside SCOP domains (right scale) and the ratios of the observed fraction to that expected if SCOP domains are randomly placed (left scale). **b,** *H. sapiens*. **c,** *M. musculus*. **d,** *R. norvegicus*. **e,** *C. elegans*. **f,** *D. melanogaster*. **g,** *S. pombe*. **h,** *A. thaliana*. **i,** *O. sativa*.

**Segments inserted in internal exons mostly encode IDRs**

Is it possible to identify elongated segments in internal exons and determine if they encode IDRs? Since uncovering such cases that took place in early eukaryotes is hard, we searched for cases of elongated internal exons after the rat-mouse divergence. The segments in internal exons that exist only in rat or mouse genome were identified (Fig. 3a) and regarded 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 that resulted from alternative splicing of entire exons, we removed those whose both boundaries coincided with exon boundaries. Although this method cannot capture all the inserted segments in rat and mouse, the identified cases are likely to be those of elongated internal exons. Insertions occurred more frequently in long internal exons (Fig. 3b). The segments inserted in internal exons encode mostly IDRs, irrespective of prediction methods used (Fig. 3c, d).

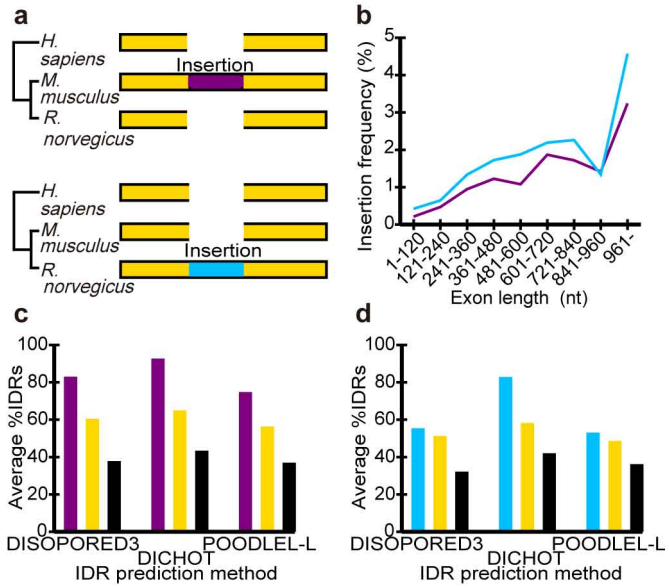



**Fig. 3| The segments are frequently inserted in long internal exons and mostly encode IDRs. a,** A scheme to identify probable insertion in internal exons. **b,** Frequency of insertion in mouse (purple) and rat (sky blue) in each internal exon length range. **c,** The average %IDRs encoded by the inserted segments (purple), internal exons in genes with inserted segments (gold) and internal exons in other genes (black) of mouse. **d,** The average %IDRs encoded by the inserted segments (sky blue), internal exons in genes with inserted segments (gold) and internal exons in other genes (black) of rat.

These findings are congruous to the notion that long internal exons generally result from insertion of segments that encode IDRs. Intriguingly, the rest of the internal exons in genes that have at least one insertion segments also encode IDRs frequently, but with the frequency less than that of inserted segments. This observation indicates that internal exons in genes that already have high IDR contents exhibit a high tendency to elongate. We note that the length distribution of internal exons is likely to be in equilibrium in extant eukaryotes: insertions are largely offset by deletions. Otherwise, some internal exons would lengthen without bound, which is not what is observed. In

the period that led to the LECA, however, elongation of internal exons probably predominated, resulting in long internal exons that are now observed in diverse branches of eukaryotes.

Four examples of insertions in internal segments are presented (Fig. 4). Fig. 4a is a case in which an internal exon likely expanded by sequence repetition in the rat lineage, adding an IDR-encoding segment just as previously suggested[14]. Fig. 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. Alteration of the 5' splicing site of an internal exon accounts for the case shown as Fig. 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 (Fig. 4d) and the inserted segment again encodes an IDR. Apparently the latter three are cases of alternative splicing and the fact that the added segments encode IDRs dovetails with the report that alternatively spliced fragments are mostly associated with IDRs.[28]

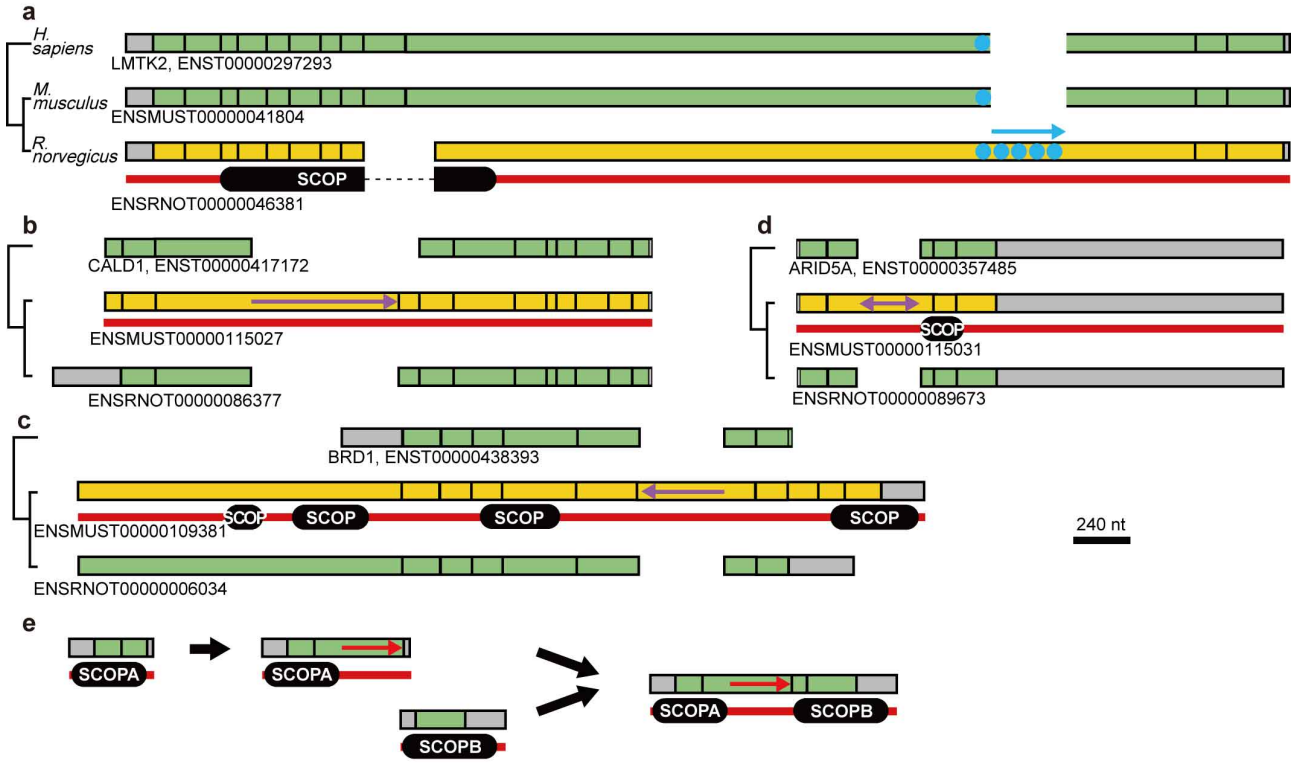



**Fig. 4| Actual examples of elongated internal exons and a model. a-d,** Human, mouse and rat orthologues. The terminal exons (grey) and internal exons (green or gold) of four groups of orthologs were scaled identically are shown with corresponding exons mostly aligned vertically. The transcript IDs and human gene names are added to the right the phylogenetic trees and the SCOP domains in encoded proteins are depicted beneath exons. Note that **b** has no SCOP domains. Arrows represent probable insertion segments. **e,** A proposed model of exon evolution that led to the LECA.

## The small bang model

We propose that in early eukaryotes some exon boundaries that are not aligned to the inside of structural domains gradually expanded with the added segments encoding IDRs (Fig. 4e). Splicing site alteration, exongenisation of introns and repeat sequence expansion are possible mechanisms of internal exon elongation. 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 generate genes 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[29,30]. Similar events occurred multiple times to produce the LECA as they conferred selective advantage.

The small bang model is consistent with the observed tendency of long internal exons to encode IDRs and the low frequency of boundaries of long internal exons that fall inside of structural domains. The actual cases of internal exons that probably elongated in the mouse or rat genome make the elongation mechanism plausible. Moreover the finding that genes with long internal exons are preferentially involved in nuclear functions supports this model. Furthermore, accretion of functional IDRs together with attenuation of R-loop formation may have conferred selective advantage to early eukaryotes that had introns. The model predicts that exon boundaries tend to separate SDs in encoded proteins. 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, many eukaryotes subsequently lost introns as they no longer conferred selective advantage. Intron losses obscured the characteristics of the LECA, but not completely: long internal exons tend to encode IDRs and their boundaries preferentially fall outside of SCOP domains in diverse branches of eukaryotes.

**Main references**


1. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635-645 (2004).
2. Koonin, E. V., Aravind, L. & Kondrashov, A. S. The impact of comprehensive genomics on our understanding of evolution. *Cell* **101**, 573-576 (2000).
3. Koonin, E. V. Wolf, Y. I. & Karev G. P. The structure of the protein universe and genome evolution. *Nature* **420**, 218-223 (2002).
4. Tordai, H., Nagy, A., Farkas, K., Bányai, L. & Patthy L. Modules, multidomain proteins and organismic compexity. *FEBS J.* **272**, 5067-5078 (2005).
5. Minezaki, T., Homma, K., Kinjo, A. R. & Nishikawa, K. Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. *J. Mol. Biol.* **16**, 1137-1149 (2006).
6. Fuxreiter, M. & Vendruscolo, M. Generic nature of the condensed states of proteins. *Nat. Cell Biol.* **23**, 587-594 (2021).
7. Koonin, E. V. 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).
8. Irimia, M. & Roy S. W. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb. Perspect. Biol.* **6**, a016071 (2014).
9. Csuros, M., Rogozin, I. B. & Koonin, E. V. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput. Biol.* **7**, e1002150 (2011).
10. Chen, L., Bush, S. J., Tovar-Corona, J. M., Castillo-Morales, A. & Urrutia A. O. Correcting for differential transcript coverage reveals a stong relationship between alternative splicing and organism compexity. *Mol. Biol. Evol.* **31**, 1402-1413 (2014).
11. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing compexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413-1415 (2008).

12. Poole, A. M. Did group II intron proliferation in an endosymbiont-bearing archaeon create eukaryotes? *Biol. Direct* **1**, 36 (2006).

13. Niu, D. K. Protecting exons from deleterious R-loops: a potential advantage of having introns. *Biol. Direct* **2**, 11 (2007).

14. Tompa, P. Intrinsically unstructured proteins evolve by repeat expansion. *BioEssays* **25**, 847-855 (2003).

15. Kondrashov, F. A. & Koonin, E. V. Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends Genet.* **19**, 115-119 (2003).

16. Sorek, R. The birth of new exons: mechanisms and evoutionary consequences. *RNA* **13**, 1603-1608 (2007).

17. Marquez, Y., Höpfler, M., Ayatollah, Z., Barta, A. & Kalyna, M. Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity. *Genome Res.* **25**, 995-1007 (2015).

18. Lander, E. S et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

19. Bolisetty, M. T. & Beemon, K. L. Splicing of internal large exons is defined by novel *cis*-acting sequence elements. *Nucl. Acids Res.* **40**, 9244-9254 (2012).

20. Kawachi, T. et al. Regulated splicing of large exons is linked to phase-separation of vertebrate transcription factors. *EMBO J.* **40**, e107485 (2021).

21. Xiong, J., Chen Y., Wang, W. & Sun, J. Biological function and molecular mechanism of SRSF3 in cancer and beyond. *Oncol. Lett.* **23**, 21 (2022).

22. Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857-863 (2015).

23. Fukuchi, S., Hosoda, K., Homma, K., Gojobori, T. & Nishikawa, K. Binary classification of protein molecules into intrinsically disordered and ordered segments. *BMC Struct. Biol.* **11**, 29 (2011).

24. Hirose, S., Shimizu, K., Kanai, S., Kuroda, Y. & Noguchi, T. POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* **23**, 2046-2053 (2007).

25. Dunker, A. K. 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. Andreeva, A., Kulesha, E., Gough, J. & Murzin, A. G. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucl. Acids Res.* **48**, D376-D382 (2019).

28. Romero P. R. et al. Alternative splicing in concert with protein instinsic disorder enebles increased functional diversity in multicellular organisms. *Proc. Natl Acad. Sci. USA* **103**, 8390-8395 (2006).

29. van Leeuwen, H. C., Strating, M. J., Rensen, M., de Laat, W. & van der Vilet, P. C. Linker length and composition influence the flixibility of Oct-1 DNA binding. *EMBO J.* **16**, 2043-2053 (1997).

30. Brodsky, S. et al. Intrinsically disordered regions direct transcription factor *in vivo* binding specificity. *Mol. Cell* **79**, 459-471 (2020).

## Methods

All the sequence, exon and orthologue data used in this study were downloaded from the Ensembl database[31]. IDRs were predicted by DISOPRED3[22], DICHOT[23] and POODLE-L[24], while SCOP domains[27] were assigned in a step in DICHOT. Using an in-house C script, we analyzed the proteins with SCOP alignments to calculate the fraction of the internal exons encoding the SCOP domains to get the expected occurrence of internal exon boundaries inside SCOP domains. Mouse or rat specific inserted 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[26] annotations (release 2021_04). The frequencies of nucleotides and amino acids were computed by an in-house C script.

## Methods references

31. Cunningham F. et al. Ensembl 2022. *Nucl. Acids Res.* **50**, D988-D995 (2022).
32. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).

## Acknowledgement

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

## Author contributions

S. F. and K. H. planned the research, T. N. provided the POODLE-L program, H. A. carried out DISOPRED and DICHOT predictions and K. H. analyzed data and wrote the paper.

## Competing interest declaration

None declared.