Evolution of the intron–exon structure of eukaryotic genes

Manyuan Long, Sandro Jose de Souza and Walter Gilbert

Harvard University, Cambridge, USA

The origin and evolution of intron–exon structures continue to be controversial topics. Two alternative theories, the 'exon theory of genes' and the 'insertional theory of introns', debate the presence or absence of introns in primordial genes. Both sides of the argument have focused on the positions of introns with respect to protein and gene structures. A new approach has emerged in the study of the evolution of intron–exon structures: a population analysis of genes. One example is the statistical analysis of intron phases — the position of introns within or between codons. This analysis detected a significant signal of exon shuffling in the DNA sequence database containing both ancient and modern exon sequences: intron phase correlations, that is, the association together within genes of introns of the same phase. The results of this analysis suggest that exon shuffling played an important role in the origin of both ancient and modern genes.

Current Opinion in Genetics & Development 1995, **5**:774–778

Introduction

Soon after the discovery of introns in eukaryotic genes, many questions arose about the origin and evolution of these intervening sequences. In general, there are two competing theories for the origin of introns. The introns-early hypothesis [1–3] or exon theory of genes [4] suggests that introns were present in the progenote. In contrast, the insertional theory of introns, also called the introns-late theory, invokes recent events of intron insertion into eukaryotic genomes [5–9].

The presence or absence of introns in the progenote has profound consequences for the origin and evolution of the genes. The existence of introns opens the possibility that exon shuffling events are the major evolutionary force in creating new genes. In a scenario of intronless genomes, new genes are created by gene duplication with modification in one of the copies. We know today that different types of introns exist, some of them with self-splicing activity. This raises another question: what is the origin of the classical nuclear spliceosomal introns? On the basis of some structural and functional similarities, it is believed that they are derived from group II introns (reviewed recently in [10,11•]). The insertionists believe that the symbiotic event which gave rise to the mitochondria and chloroplast provided the eukaryotic cell with a source of group II introns. Cavallier-Smith [12] suggested that these group II introns would invade the nuclear genes of the eukaryotic cell by retroposition creating, after modification, most of the introns that we see today. On the other hand, the exon theory of genes states that these nuclear introns are descendants of self-splicing introns (ribozymes) that were present in the progenote. After the divergence of the three major kingdoms, most of these ancient introns were lost in the Eubacteria and Archaebacteria. In the eukaryotes, however, these introns evolved to a more efficient and complex type of intron, the spliceosomal intron.

Historically, three lines of evidence have been used to support the exon theory of genes. These are the correlation between protein modules and exons [13–15], the correspondence of intron positions between plant and animal genes [16–20], and, finally, the correlation between intron positions in genes that diverged in the progenote [21••,22–24]. On the other hand, the phylogenetic distribution of introns [8] and the mobile activity of some introns [10,11•,25•] has been taken as evidence for the insertional theory of genes. Recently, we have developed a novel approach to the question of the origin of introns: a population analysis of genes in the DNA sequence database. By studying the intron phase distribution and the correlations of intron phases, we have obtained results which, in our opinion, provide strong evidence for the exon theory of genes [26••]. This review discusses this work on intron phase distributions while reviewing recent progress in the field.

Searching for ancient introns

Phylogenetic analysis of eukaryotic genes that exist in both nuclear and organellar versions revealed that these genes are products of ancient duplications, which probably antedate the prokaryote/eukaryote divergence. If introns are in identical positions in such genes, they are good candidates for ancient introns. The glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene provides a convincing example. Kersanach *et al.*

 ISSN 0959-437X

[21••] pointed out that five introns are in identical positions between nuclear and chloroplast GAPDH, suggesting that these introns were likely to exist in the ancestor of eukaryotes and prokaryotes. Other examples of this phenomenon are the cytosolic and mitochondrial malate dehydrogenase and aspartate aminotransferase where two and five introns, respectively, are in identical positions [22,23]. The insertionists, however, see these coincidences as non-random insertions [27•]. In their view, there is a limited number of sites where introns can be inserted, and, therefore, the coincidences discussed above are not statistically significant, that is, they could have occurred by chance. The existence of such limited sites for insertions, however, remains questionable.

The major difficulty in tracing ancestral introns is that what one sees today is the final product of intron loss, intron drift, and even occasional cases of intron acquisition. Although the insertionists always claim that clear examples of intron drift are missing (see [8] and reply of [28••]), a clear example of intron sliding has been found in two paralogous genes in *Volvox* [29]. In the case of intron loss, two recent examples reinforce the notion that this is a common phenomenon [30•,31•].

Correlations between introns and protein modules

Is there a correlation between exons and units of protein structure, such as the relations of exons to modules as suggested by Go [15]? The historical perspectives of this assumption are well described in the literature [4,13,14,32•]. The most striking successes of this approach were the predictions of novel intron positions [13,15] that were later observed [33,34]. Stoltzfus *et al.* [35••] tried to approach the question of whether exons correspond to units of protein structure by examining four ancient genes. They concluded that the "exon theory of genes is untenable" on the basis of their failure to show any such correlations. Their argument, however, is weakened by several factors. Their first approach was to test a correlation between exons and secondary structural elements in proteins. However, it is not surprising that they did not find such a correlation because one has known since 1979 that introns often break α-helices [36]. They also tested whether or not introns are close to a center of mass of the protein structure. Although such notion of centrality might have been held in the early 1980s, in 1985 Straus and Gilbert [37] had already pointed out that centrality was not a useful feature for triosephosphate isomerase gene (TPI). Thus, in these two directions, they attacked straw men. Finally, Stoltzfus *et al.* [35••] tried to test the correlation between intron positions and modules as defined by Go [15]. With the possible exception of TPI, they again did not find any positive correlation, but this analysis is again not without problems. The first, and major, problem is the definition of ancestral introns, as today's intron–exon structures are the final products of intron loss, intron sliding, and possibly intron gain. Secondly, issues of intron position always involve issues of protein sequence alignment which may be problematic. In comparing genes that diverged a very long time ago, there can be regions that lack enough sequence similarity to make a reliable alignment. Thus what appears to be different intron positions in the central regions of the globins, for example, may only reflect alignment artifacts.

Addition of introns

One view of the addition of introns is that they behave as transposable elements entering some sequence, a protosplice set, in a gene [38]. When one examines the sequences at intron boundaries, one occasionally notices that the introns are flanked by repeating sequences, an AG/GT sequence before and after the intron. This has suggested the possibility the intron entered a pre-existing AG/GT sequence, the intron carrying within it a /GT...AG/ sequence. This question has been extensively examined by Stephens and Schneider [39] who surveyed 1 800 human introns to investigate the structural features of splice sites. Using an information measure of sequence conservation, they found that most of the information is confined within the introns: 82% of information in the donor (5′) sequences and 97% of information in the acceptor (3′) sequences are on the intron sides of the junctions. There is low conservation for an AG/G sequence on the exon sides of the junctions and essentially none for the AG/GT sequence. In *Caenorhabditis elegans* [40] only an AG/R sequence appears, and in *Drosophila melanogaster* [41], only AG/RT (where R is A or G). Although this bias in sequence at the exon boundary could possibly be the result of drift away from an original AG/GT sequence, hypothetically used as a site of insertion, it could also just be the result of a drift of sequence to enhance pairing with the small nuclear RNAs of the splicing machinery.

Recently, a demonstration of the addition of spliceosomal introns to nuclear genes and a mechanism for that to occur has been achieved through the work of Tokio Tani and Yasumi Ohshima and their collaborators [42–44]. In several yeasts the U6 small nuclear RNA has been found to have added introns, one in *Schizosaccharomyces pombe* and five in different species of *Rhodotorula*. The U6 small nuclear RNA plays a catalytic role in premessenger RNA splicing. The distribution suggests strongly that these introns were added to the U6 gene. The likely mechanism is reverse splicing: introns spliced out of a messenger RNA were spliced into the catalytic U6 RNA. That intron containing RNA molecule was then later, by accident, copied by reverse transcriptase into cDNA and inserted, by recombination, into the nuclear gene. This is a general mechanism for the movement or insertion of spliceosomal introns into RNA. However, if one looks at the sequences into which these introns

have been inserted in the small nuclear RNAs, there is no sign of a protosplice site. The six sequences in U6 RNA are: AA/AU, GA/GA, AG/AU, UU/AG, GG/AU and CU/GC. Even though inside the intron the standard GT...AG sequences are used, there is no sign of any sequence conservation on the exon side. This supports the concept that no protosplice sequence was used and that any conservation on the exon side is the result of drift toward more effective splicing.

Signal of exon shuffling: intron phase correlations

Intron phase, defined as the position of the intron within a codon [45], is a conserved evolutionary character, as a non-deleterious phase change needs a double mutation of insertion and deletion. To analyze the intron phase distribution, we first built a database with more than 13 000 exon sequences that were extracted from 1 925 independent or nearly independent eukaryotic genes. We then calculated the phases of all introns in these genes. We found a great excess of phase zero introns (i.e. those that lie between codons) over the other two phases and a highly significant correlation of intron phases: the three types of introns are not arranged randomly within genes but the phases of adjacent introns prefer to be associated together, that is, there is an excess of symmetric exons (exons flanked by two introns of the same phase) [26••]. Furthermore, similar correlations of the intron phases flanking sets of exons also appear. In all cases, the greatest excess of symmetric patterns is a roughly 30% excess of (1,1) patterns over the expectation (on the basis of the biased intron phase distribution). What is the explanation of these observations?

The excess of phase zero introns could itself be taken as evidence for exon shuffling, as exon shuffling works more efficiently if the introns are all in the same phase. The simplest form of an insertional model predicts no phase bias and equal numbers of introns in all three phases if the addition of introns is independent of sequence. However, models in which introns are added to protosplice sites, such as the AG/GT sequence, will produce a phase bias.

The correlation of phases and the excess of symmetric exons, however, is a far stronger argument for exon shuffling. Exon shuffling requires that the shuffled exon or sets of exons be symmetric [46,47] because the addition of a symmetric exon or symmetric exon sets into an intron of the same phase does not disturb the reading frame.

Could the correlation of intron phases be caused by alternative splicing? Alternative splicing that adds alternative sets of exons to the beginning or end of genes would not restrict the phases or the phase combinations of exons. Alternative splicing events that add an additional internal exon to a previously functioning structure do require that the added exon be symmetric, but, of course, also require that added element code for a peptide that can be added as a complete structure, a unit of structure or function. Furthermore, alternative splicing is solely dependent on the biological properties of individual genes [48,49] rather than the distances between the introns. Therefore, alternative splicing is not likely to provide an explanation for the correlations as we observed. The simplest explanation for the excess of symmetric exons is that it is a signal of exon shuffling.

Exon shuffling in modern genes

In the controversial area of evolution of intron–exon structure, one concept is reasonably well accepted —that of exon shuffling, at least in modern genes. Since the initial reports of examples of exon shuffling in the low density lipoprotein receptor [50] and the regulatory proteases of the blood coagulation [51], many further examples have been documented (comprehensive reviews can be found in [47,52–54]). One recent example of exon shuffling in the sterol regulatory element binding protein-2 gene (SREBP-2) in hamster shows an evolution of the intron–exon structure in the laboratory, where exon shuffling conferred a new function selected in a new environment [55•]. A recent report of exon shuffling in another specific gene in sunflower [56•] discredits the view that exon shuffling is only limited to one evolutionarily recent lineage of vertebrates [8,35••,57].

How general is exon shuffling? Our detection of a sign of exon shuffling in a database of eukaryotic exons makes it possible to estimate how many exons are involved in that signal. At least 19% of the exons had to be involved in order to create the observed excess of symmetric exons over expectation. (If all of the deviation from the 1/3 expectation of intron phases were due to exon shuffling, then at least 28% of the database has to be involved.) This provides only a minimum estimate on how much shuffling might have occurred, a lower boundary, as much of the shuffling might not appear as an excess and factors such as intron drift will weaken the signal of intron correlations. Thus, quantitatively, exon shuffling is very important in evolution.

Exon shuffling in ancient genes

To study the origins of exon shuffling, we identified in our database those regions of genes which were homologous to prokaryotic genes. These are ancient conserved regions (ACR) [58] which represent complete genes or portions of genes that descended in a relatively unchanged manner from a common ancestor. These regions have no introns in the prokaryotes, but they have introns in the eukaryotic versions. On any

intron-late model these introns must have been inserted. For such models they could not participate in exon shuffling because the ACR regions are colinear with the ancestral molecules, which according to the introns-late hypothesis came into existence before there were introns. We examined these ACR introns for intron phase bias and for intron phase correlation. We discovered that the introns within these ACRs show a similar phase bias to the overall database, about 55% in phase zero, and show intron phase correlations. There is an excess of symmetric exons and symmetric sets of exons, significant at the 1% level, for these introns. We interpret this excess of symmetric exons again as evidence of exon shuffling. However, this exon shuffling would have had to have occurred in the progenote, and hence these introns would have existed in the progenote. This is a novel argument for the ancient nature of introns, based on principles different from those adduced before, which provides independent support for the concept that the introns are very old.

Conclusions

Although the debate as to early versus late introns continues, new experimental findings show more examples of exon shuffling in organisms. Our data on the correlation of intron phases suggest that there is an excess of symmetric exons and hence that a large portion of the exon database has been involved in exon shuffling, which further suggests a very important role of exon shuffling in the evolution of genes. The same signal detects exon shuffling in ancient conserved regions, which in turn strongly supports the idea that introns were present in the progenote.

Acknowledgements

We thank the NIH for their support, grant GM37997 to W Gilbert. We thank Nancie Munroe for secretarial assistance. SJ de Souza is supported by a post-doctoral fellowship from Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (Brasil).

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Gilbert W: **Why gene in pieces?** *Nature* 1978, **271**:501.

2. Blake CCF: **Do genes-in-pieces imply protein in pieces?** *Nature* 1978, **273**:267.

3. Doolittle WF: **Gene-in-pieces: were they ever together?** *Nature* 1978, **272**:581–582.

4. Gilbert W: **The exon theory of genes.** *Cold Spring Harbor Symp Quant Biol* 1987, **52**:901–905.

5. Cavallier-Smith T: **Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox.** *J Cell Sci* 1978, **34**:247–278.

6. Rogers J: **Exon shuffling and intron insertion in serine proteases genes.** *Nature* 1985, **315**:458–459.

7. Rogers J: **How were introns inserted into nuclear genes?** *Trends Genet* 1989, **5**:213–216.

8. Palmer JD, Logsdon JM: **The recent origins of introns.** *Curr Opin Genet Dev* 1991, **1**:470–477.

9. Cavallier-Smith T: **Selfish DNA and the origin of introns.** *Nature* 1985, **315**:283–284.

10. Nilsen TW: **RNA–RNA interaction in the spliceosome: unraveling the ties that bind.** *Cell* 1994, **78**:1–4.

11. • Michel F, Ferat J-L: **Structure and activities of group II introns.** *Annu Rev Biochem* 1995, **64**:435–461.
A part of this review is devoted to a discussion on the evolution of group II introns including the possible evolutionary relationship between this group and the classical spliceosomal introns.

12. Cavallier-Smith T: **Intron phylogeny: a new hypothesis.** *Trends Genet* 1991, **7**:145–148.

13. Gilbert W, Marchionni M, McKnight G: **On the antiquity of introns.** *Cell* 1986, **46**:151–153.

14. Gilbert W, Glynias M: **On the ancient nature of introns.** *Gene* 1993, **135**:137–144.

15. Go M: **Correlations of DNA exonic regions with protein structural units in haemoglobin.** *Nature* 1981, **291**:90–93.

16. Marchionni M, Gilbert W: **The triosephosphate isomerase gene from maize: introns antedate the plant/animal divergence.** *Cell* 1986, **46**:133–141.

17. Imajuku Y, Hirayama T, Endoh H, Oka A: **Exon-intron organization of the *Arabidopsis thaliana* protein kinase genes CDC2a and CDC2b.** *FEBS Lett* 1992, **304**:73–77.

18. Nawrath C, Schell J, Koncz C: **Homologous domains of the largest subunit of eucaryotic RNA polymerase II are conserved in plants.** *Mol Gen Genet* 1990, **223**:65–75.

19. Shah DM, Hightower RC, Meagher RB: **Genes encoding actin in higher plants: intron positions are highly conserved but the coding sequences are not.** *J Mol Appl Genet* 1983, **2**:111–126.

20. Pardo JM, Serrano R: **Structure of plasma membrane H^+-ATPase gene from the plant *Arabidopsis thaliana*.** *J Biol Chem* 1989, **264**:8557–8562.

21. •• Kersanach R, Brinkmann H, Liaud M-F, Zhang D-X, Martin W, Cerff R: **Five identical intron positions in ancient duplicated genes of eubacterial origin.** *Nature* 1994, **367**:387–389.
The authors found three new intron positions (two were previously known) precisely conserved between chloroplast and nuclear-encoded glyceraldehyde-3-phosphate dehydrogenase (GAPDH) what gives a strong support to the exon theory of genes.

22. Obaru K, Tsuzuki T, Setoyama C, Shimada K: **Structural organization of the mouse aspartate aminotransferase isozyme genes: introns antedate the divergence of cytosolic and mitochondrial isozymes genes.** *J Mol Biol* 1988, **200**:13–22.

23. Setoyama C, Joh T, Tsuzuk T, Shimada K: **Structural organization of the mouse cytosolic malate dehydrogenase gene: comparison with that of the mouse mitochondrial malate dehydrogenase gene.** *J Mol Biol* 1988, **202**:355–364.

24. Iwabe N, Kuma K, Kishino H, Hasegawa M, Miyata T: **Compartmentalized isoenzymes and the origin of intron.** *J Mol Evol* 1990, **31**:205–210.

25. • Giroux MJ, Clancy M, Baier J, Ingham L, McCarty D, Hannah LC: ***De novo* synthesis of an intron by the maize transposon element dissociation.** *Proc Natl Acad Sci USA* 1994, **91**:12150–12154.

A clear observation that mobile elements can be converted to a classical intron.

26. •• Long M, Rosenberg C, Gilbert W: **Intron phase correlations and the evolution of the intron–exon structure of genes.** *Proc Natl Acad Sci USA* 1995, in press.
An extensive statistical analysis of intron phase distribution in a large exon database created recently by the authors was carried out, revealing a significant correlation of intron phases in both modern and ancient genes. The power of population analysis of genes is obvious in this study, in that it detected clear signals for the existence of early introns against evolutionary noise of loss, gain, and sliding of introns.

27. • Logsdon JM, Palmer JD, Stoltzfus A: **Origin of introns - early or late [letter]?** *Nature* 1994, 369:526–527.
A critique of [21••], where the authors discuss the phylogenetic distribution of introns in GAPDH (Logsdon and Palmer) as well as the probability of site-specific intron insertion events (Stoltzfus). A reply from the authors of [21••] follows these comments.

28. •• Senapathy P; and Bertolaet BL, Seidel HM, Knowles JR: **Introns and the origin of protein-coding genes [letter].** *Science* 1995, **268**:1366–1367.
Two comments about [35••] where the authors indicate that the lack of correlation between exons and units of protein structure (as suggested in [35••]) does not disprove their respective theories, i.e., the split-gene hypothesis (Senapathy) and the exon microgene theory (Bertolaet *et al.*). Both letters are followed by a reply of Stoltzfuz *et al.*

29. Muller K, Schmitt R: **Histone genes of *Volvox carteri*: DNA sequence and organization of two H3-H4 gene loci.** *Nucleic Acids Res* 1988, **16**:4121–4136.

30. • Luque T, Atrian S, Danielsson O, Jornvall H, Gonzalez-Duarte R: **Structure of the *Drosophila melanogaster* glutathione-dependent formaldehyde dehydrogenase/octanol dehydrogenase gene (class III alcohol dehydrogenase).** *Eur J Biochem* 1994, **225**:985–993.
Analysis of the *Drosophila* and other class III ADH suggests intron loss due to retroposition events.

31. • Geiss KT, Abbas GM, Makaroff CA: **Intron loss from the NADH dehydrogenase subunit 4 gene of lettuce mitochondrial DNA. Evidence for homologous recombination of a cDNA intermediate.** *Mol Gen Genet* 1994, **243**:97–105.
A clear example of intron loss due to a different kind of retroposition involving a recombination between an intronless copy of cDNA and the original gene.

32. • Mattick JS: **Introns: evolution and function.** *Curr Opin Genet Dev* 1994, **4**:823–831.
Based on the introns-late model, the author proposed that introns evolved as functional elements within genomes.

33. Jensen EO, Paludan K, Hyldig-Nielsen JJ, Jorgensen P, Markere KA: **The structure of a chromosomal leghaemoglobin gene from soybean.** *Nature* 1981, **291**:677–679.

34. Tittiger C, Whyard S, Walker VK: **A novel intron site in the triosephosphate isomerase gene from the mosquito *Culex tarsalis*.** *Nature* 1993, **361**:470–472.

35. •• Stoltzfus A, Spencer DF, Zucker M, Logsdon JM, Doolittle WF: **Testing the exon theory of genes: the evidence from protein structure.** *Science* 1994, **265**:202–207.
The author critically examined the exon theory of genes using several tests they designed to investigate the correlation between exons and units of proteins structures in four ancient genes.

36. Gilbert, W: **Introns and exons: playgrounds of evolution.** *ICN–UCLA Symp Mol Cell Biol* 1979, **14**:1–12.

37. Straus D, Gilbert W: **Genetic engineering in the pre-cambrian: structure of the chicken triosephosphate isomerase gene.** *Mol Cell Biol* 1985, 5:3497–3506.

38. Dibb NJ, Newman AJ: **Evidence that introns arose at proto-splice sites.** *EMBO J* 1989, **8**:2015–2021.

39. Stephens RM, Schneider TD: **Features of spliceosome evolution and function.** *J Mol Biol* 1992, **228**:1124–1136.

40. Fields C: **Information content of *Caenorhabditis elegans* splice site sequences varies with intron length.** *Nucleic Acids Res* 1990, **18**:1509–1512.

41. Mount SM, Burks C, Hertz G, Stormo GD, White O, Fields C: **Splicing signals in *Drosophila*: intron size, information content, and consensus sequences.** *Nucleic Acids Res* 1992, **20**:4255–4262.

42. Takahashi Y, Urushiyama S, Tani T, and Ohshima Y: **An mRNA-type intron is present in the *Rhodotorula hasegawae* U2 small nuclear RNA gene.** *Mol Cell Biol* 1993, **13**:5613–5619.

43. Tani T, Ohshima Y: **mRNA-type introns in U6 small nuclear RNA genes: implications for the catalysis in pre-mRNA splicing.** *Genes & Devel* 1991, 5:1022–1031.

44. Tani T, Ohshima Y: **The gene for the U6 small nuclear RNA in fission yeast has an intron.** *Nature* 1989, **337**:87–90.

45. Sharp PA: **Speculations on RNA splicing.** *Cell* 1981, **23**:643–646.

46. Patthy L: **Intron-dependent evolution: preferred types of exons and introns.** *FEBS Lett* 1987, **214**:1–7.

47. Patthy L: **Modular exchange principles in proteins.** *Curr Opin Struct Biol* 1991, **1**:351–361.

48. Hodges D, Bernstein SI: **Genetic and biochemical analysis of alternative RNA splicing.** *Adv Genet* 1994, **31**:207–228.

49. McKeown M, **Alternative mRNA splicing.** *Annu Rev Cell Biol* 1992, **8**:133–155.

50. Sudhof TC, Russell DW, Goldstein JL, Brown MS, Sanchez-Pescador R, Bell GI: **Cassette of eight exons shared by genes for LDL receptor and EGF precursor.** *Science* 1985, **228**:893–895.

51. Patthy L: **Evolution of the proteases of blood coagulation and fibrinosis by assembly from modules.** *Cell* 1985, **41**:657–663.

52. Doolittle RF: **The multiplicity of domains in proteins.** *Annu Rev Biochem* 1995, **64**:287–314.

53. Dorit RL, Gilbert W: **The limited universe of exons.** *Curr Opin Struct Biol* 1991, **1**:973–977.

54. Dorit RL, Schoenbach L, Gilbert W: **How big is the universe of exons?** *Science* 1990, **250**:1377–1382.

55. • Yang J, Brown MS, Ho YK, Goldstein JL: **Three different rearrangements in a single intron truncate sterol regulatory element binding protein-2 and produce sterol-resistant phenotype in three cell lines.** *J Biol Chem* 1995, **270**:12152–12161.
In three different cell lines selected by sterol-resistance, the phenotype was due to independent exon shuffling events occurring in one intron of sterol regulatory element binding genes.

56. • Domon C, Steinmetz A: **Exon shuffling in anther-specific genes from sunflower.** *Mol Gen Genet* 1994, **244**:312–317.
The first report of exon shuffling in plant genes.

57. • Patthy L: **Exons — original building blocks of proteins** *Bioessays* 1991, **13**:187–192.

58. Green P, Lipman D, Hillier L, Waterson R, States D, Claverie J-M: **Ancient conserved regions in new gene sequences and the protein databases.** *Science* 1993, **259**:1711–1717.

M Long, SJ de Souza and W Gilbert, Biological Laboratories, Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA.