BioMedNet information for advertisers > click here <
DoubleTwist.com
BioMedNet Home Library Databases Collaborations Job Exchange Shopping Mall Your Room

FULL TEXT (TEXT + FULL FIGURES)

Current Opinion in Genetics & Development
Vol. 9, No. 6, December 1999

Orthologs, paralogs and genome comparisons

[Review article]

J Peter Gogarten, Lorraine Olendzenski

Current Opinion in Genetics & Development 1999, **9**:630-636.

Viewing options [Help]

text only | + thumbnails | **+ full figures** | PDF

Associated links

- Publications by Lorraine Olendzenski | J Peter Gogarten
- Related fulltext articles on BioMedNet

Outline

- Abstract
- Abbreviations
- Introduction – homology and the evolution of protein families
- Delineation of protein families
- Ancient gene duplications and the root of the tree of life
- Gene and genome duplications, speciation, and innovation in eukaryotes
- Convergence versus concerted evolution
- Conclusions
- Acknowledgements
- References and recommended reading
- Copyright

Abstract

During the past decade, ancient gene duplications were recognized as one of the main forces in the generation of diverse gene families and the creation of new functional capabilities. New tools developed to search data banks for homologous sequences, and an increased availability of reliable three-dimensional structural information led to the recognition that proteins with diverse functions can belong to the same superfamily. Analyses of the evolution of these superfamilies promises to provide insights into early evolution but are complicated by several important evolutionary processes. Horizontal transfer of genes can lead to a vertical spread of innovations among organisms, therefore finding a certain property in some descendants of an ancestor does not guarantee that it was present in that ancestor. Complete or partial gene conversion between duplicated genes can yield phylogenetic trees with several, apparently independent gene duplications, suggesting an often surprising parallelism in the evolution of independent lineages. Additionally, the breakup of domains within a protein and the fusion of domains into multifunctional proteins makes the delineation of superfamilies a task that remains difficult to automate.

Abbreviations

PSI-BLAST—position-specific iterative basic local alignment search tool.

Introduction – homology and the evolution of protein families

During the early evolution of life, gene duplications are considered to have allowed for the rapid diversification of enzymatically catalyzed reactions and an increase in genome size [1], and provided material for the invention of new enzymatic properties, the diversification of cytoskeletal elements and more complex regulatory and developmental patterns (e.g. [2] [3] [4] [5]). Homology refers to two structures or sequences that evolved from a single ancestral structure or sequence. To classify the different types of homology, Fitch [6] introduced the terms 'orthology' and 'paralogy'. Orthologous structures or sequences in two organisms are homologs that evolved from the same feature in their last common ancestor but they do not necessarily retain their ancestral function. The evolution of orthologs reflects organismal evolution — molecular systematics has, therefore, traditionally been concerned with comparing orthologous sequences. In contrast, homologs whose evolution reflects gene duplication events are called paralogs. For example, the β chain of hemoglobin is a paralog of the hemoglobin α chain and of myoglobin as they evolved from the same ancestral globin gene through repeated gene-duplication events.

In eukaryotes, many gene duplications appear to have been generated in an autochthonous fashion (i.e. within a single line of descent). A possible mechanism for their generation is miss-pairing and miss-sorting of alleles during cell division. Examples for these putatively autochthonous gene duplications are the histone and rRNA gene families but many smaller gene families in eukaryotes might also belong in this group (e.g. V-ATPase A and B subunit encoding genes in plants [7]; globin genes in birds and mammals [8]). Horizontal gene transfer in prokaryotes and symbiosis in eukaryotes, however, can result in paralogs being present in the same genome. Using genes that encode resistance to antibiotics as a paradigm, Fitch [6] coined the term 'xenology' for homologs that were acquired by an organism via horizontal gene transfer. The archaeal/vacuolar type sodium-pumping ATPases found in some Gram-positive bacteria in addition to their proton-pumping F-ATPase, probably represent xenologs. To separate the transfer of single genes from events that involve the fusion of complete genomes Gogarten [9] introduced the term 'synology' to denote homologs that occur in a single organism by the fusion of two independent lines of descent. Bacterial genes brought into the eukaryotic cell via the mitochondrial endosymbiont provide an example of synologous sequences.

In criticism of these concepts, Williams and Embley [10] pointed out that the different categories cannot be clearly identified from the available data; however, the increasing number of diverse genome sequences available allows testing of ideas that a few years ago were considered untestable. The suggestion that a chimera is at the root not only of the eukaryotic but also of the archaeal domain can serve to illustrate this point [11] [12]; vertical inheritance alone is insufficient to explain the inferred molecular phylogenies [13] [14••]. The many bacterial genes found not only in eukaryotic [15] [16] but also in archaeal genomes [17] suggest horizontal gene transfer. Analyses of multiple molecular markers indicates that this transfer of bacterial genes into the archaeal domain did not occur in a single event but involved at least a few independent interdomain transfers [12] [14••] [18]. However — especially when considering ancient gene divergences, including those in eukaryotes — it is often impossible to discriminate autochthonous duplications from those that involved horizontal transfer. This review describes recent methodological advances in the delineation of families of paralogous genes. We discuss the use of paralogous genes for rooting the tree of life and address problems in phylogenetic reconstruction that result from long-branch attraction and from partial gene conversions between paralogous genes. Finally, we review the current discussion on the role of gene and genome duplications in molecular innovation and speciation.

Delineation of protein families

Conventional databank searches (e.g. FASTA [19] or BLAST [basic local alignment search tool]) [20] compare a single sequence to those stored, and find and align those sequences that give the best match to the query sequence. Although convergent evolution has been observed [21] [22], protein sequence space (defined by all possible permutations of amino acid sequences) is so large that any significant similarity found with standard search tools can safely be assumed to be as a result of homology (i.e. shared evolutionary ancestry) [23] [24]. Sadly, the reverse is not true: sequences that in standard searches show no significant similarity might nevertheless be homologs the sequences of which have diverged too much for the similarity to be detected. For example the vacuolar, archaeal and bacterial F-ATPases are homologous multi-subunit enzymes; they have very similar quaternary structures, and the homology between some of the subunits can be demonstrated readily (e.g. [25] [26]). The subunits that form the stalk of these ATPases, however, have no dectectable similarity among the different ATPase types. Nevertheless, it is a safe assumption that the ancestor of these ATPases already had a stalk that connected the ATP hydrolyzing subunits to the proton translocating proteolipid, both of which show detectable homology between the different ATPase types [26]. Selection pressure on stalk subunits was insufficient to retain significant primary structural similarity over billions of years of evolution.

Some protein families, the members of which have very divergent functions, were already established using conventional tools. For example, the nucleotide-binding subunits of the F-ATPase are homologous not only to the nucleotide-binding subunits of other ion-translocating ATPases, they also are homologous to an ATPase involved in the assembly of bacterial flagella [27] and to a bacterial transcription termination factor [28]. Using conventional tools, however, the detection of superfamilies is restricted to protein folds that are under sufficient selective pressure to retain recognizable sequence motifs.

Using conventional tools, Riley and Labedan [29] [30] [31][32••] analyzed families of paralogs encoded in the *Escherichia coli* genome. Of the 3996 putative proteins longer than 79 amino acids encoded in the *E. coli* genome, 66% were members of paralogous pairs or groups and, using their approach, 1367 proteins had no detectable homolog in *E. coli* — that is, the majority of *E. coli* proteins evolved through gene duplication [32••]. To avoid the artificial extension of families through multidomain proteins, Labedan and Riley dissected the paralogous genes into modules [31]. In the complete *E. coli* genome, they found 873 families of modules, in addition to 1367 siglets. Given their very conservative approach, these numbers certainly underestimate the contribution of gene duplications to genome evolution.

In some instances, homology between proteins with very divergent functions was suggested by comparison of their three-dimensional structures and not by the similarity of the primary sequence. Brenner *et al.* [33] demonstrated that conventional sequence comparison is often insufficient for detecting homologs. An interesting example regarding the detection of homology on the basis of three-dimensional information is that of **D**-alanine:**D**-alanine ligase and glutathione synthetase [34]. The structures of these proteins — as well as that of the two ATP-binding domains of carbamoyl phosphate synthetase, and succinyl-CoA synthetase [35] — are so similar that there can be little doubt that they all evolved from the same ancestral protein. Nevertheless, pairwise comparisons of the primary sequences fail to confirm the homology.

The development of new search tools that utilize information from more than a single sequence [36] [37] is revolutionizing our understanding of protein families. In particular, PSI (position-specific iterative) BLAST [38], a program that performs iterated BLAST searches and successively builds and uses a pattern of conserved residues, is providing amazing results. For example, after only a few iterations using the **D**-alanine:**D**-alanine ligase as the initial query, the program not only recovered the homologs that were already detected on the basis of three-dimensional information, it also identified many other proteins which have a similar ATP-binding domain, for example urea amidolyase, tubulin–tyrosine ligase, malate–CoA ligase, and ATP–citrate lyase [39]. Wolf and co-workers [40••] used PSI BLAST to study the distribution of protein folds in completely sequenced genomes. 284 distinct folds were defined using available tertiary structures and used to search the proteins encoded in 14 different genomes. Folds could be predicted for 21% of *Caenorhabditis elegans*, 24% of *Saccharomyces cerevisiae*, 39% of *Mycoplasma genitalium* and 28% of *E. coli* proteins. About one tenth of the

proteins with successful fold prediction were recognized as multidomain proteins. Surprisingly the fraction of multidomain proteins was not found to be different in the different domains of life. Using a similar approach, Teichmann *et al*. [41] found that 114 identified protein superfamilies comprise ~41% of the protein-encoding sequences in the *M. genitalium* genome. At first sight, these numbers might seem disappointingly small considering that a much more powerful search algorithm was employed but one has to keep in mind that only folds with known structures were included and membrane proteins, which have not been crystallized, were not included (although they are known to form some of the most extensive families of paralogs [31] [42] [43]).

The error rate for false positives in PSI BLAST searches was estimated from double assignments to be ~2% [33] [40••]. The rate for false negatives is more difficult to assess, it certainly depends on the sequence used for the initial query [44••]. The present limitation in using PSI BLAST is that it only works when some similar sequences are found in the initial search. Using different initial query sequences that belong to the same superfamily does not always retrieve the same nor the complete set of matching sequences [33] [40••] [44••]. Thus the delineation of a superfamily requires much human intervention, using different query sequences and applying some judgement in removing suspected false positives. Nevertheless, this technique has significantly extended our knowledge of diverse protein families that evolved from ancient gene duplications [45••]. PSI BLAST critically depends on the amount and diversity of recognized homologs. Therefore, the process of fold identification using PSI BLAST or related algorithms can be expected to improve as more structural information and more related sequences become available [44••].

Ancient gene duplications and the root of the tree of life

Ancient gene duplications allow the evolution of protein families to be traced back to the time before the last common ancestor of all organisms [46] [47]. One outcome of these studies was the placement of the last common ancestor in the tree of life. If paralogs were already present in the last common ancestor, the phylogenies for each of the paralogs should be identical except for instances of horizontal gene transfer and additional later duplication events. The branch that connects the two sets of ancient paralogs (i.e. the branch along which the gene duplication occurred) indicates the placement of the last common ancestor (compare Fig. 1a,d). Using this approach, analyses of ATPases [48], elongation factors [49] [50] [51], aminoacyl tRNA synthase [52], and the nucleotide-binding site of signal recognition particles [53] placed the root of the tree of life outside of each of the three domains on the central branch leading to the bacteria, thus joining the archaea and the eukaryotic nucleocytoplasmic component as sister taxa. Many analyses, especially those of the replication, transcription and translation machinery [54] [55] support the closer relationship of archaea and eucarya; in contrast, other characters support a fundamental split between pro- and eukaryotes [56] [57] [58] [59]. The rooting of the tree of life in the central bacterial branch should be treated as a working hypothesis and not as established fact [60] [61]. Two considerations need to be kept in mind.

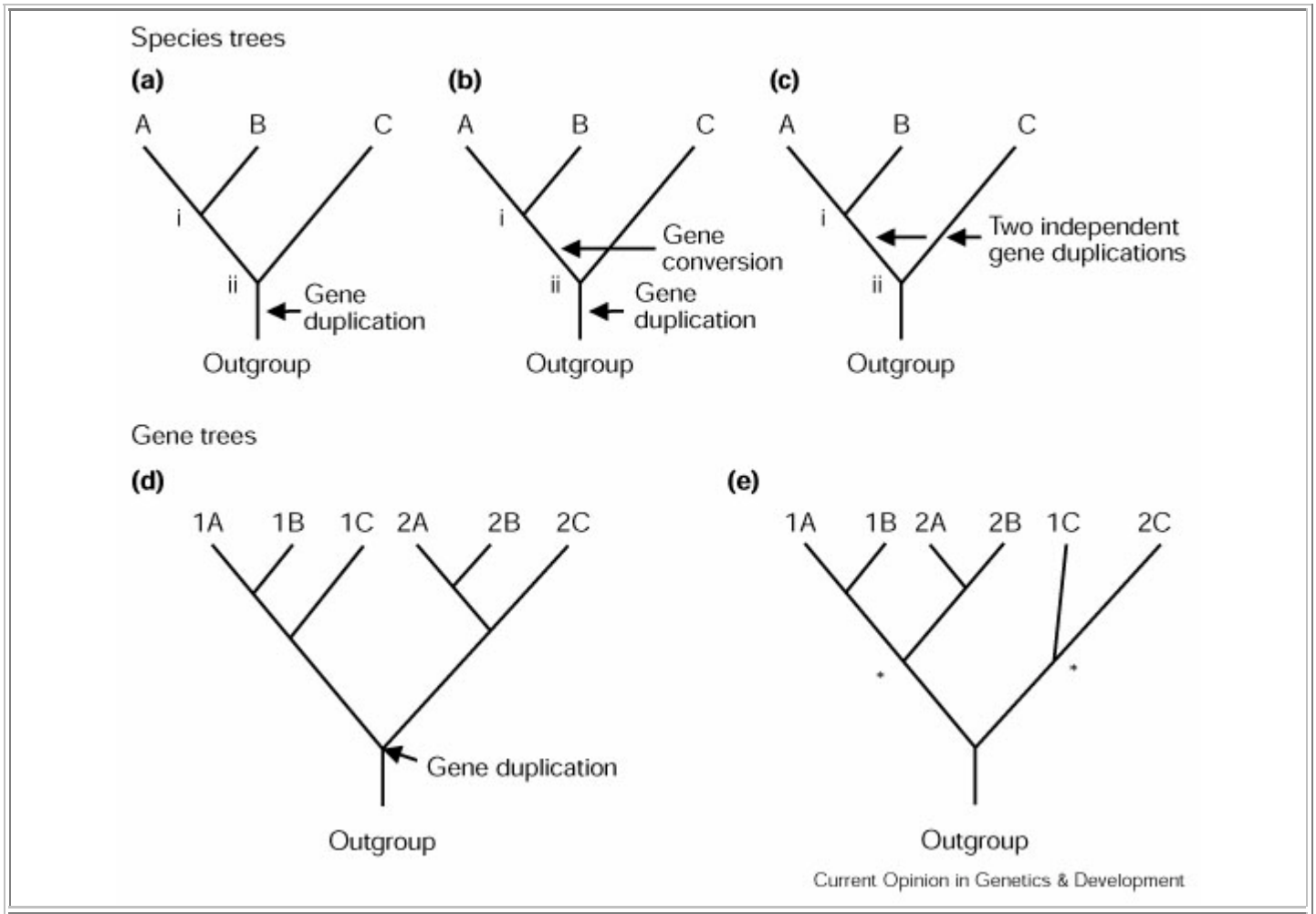

Figure 1 Gene duplication and gene conversion events in phylogenetic reconstruction. The trees depict the evolution of three species named A, B, and C that contain two paralogous genes labeled 1 and 2. If the gene duplication that gave rise to the paralogs occurred before the deepest speciation event (node ii; species tree in **[a]**), the interpretation of the molecular phylogeny **(d)** is straightforward: the central node reflects the gene duplication, each of the speciation events (node i and ii) is reflected in each of the subtrees for the paralogs. The reconstruction of the evolution of paralogous genes, however, often yields phylogenies as depicted in **(e)**. This molecular phylogeny contains two nodes (indicated with asterisks) that apparently do not reflect speciation events. Two interpretations are possible: either two independent gene duplication events took place **(c)**, or a single gene duplication event took place before the speciation indicated by node II followed by at least one gene conversion event along the branch connecting node i and ii **(b)**. (See text for further discussion and examples.)

Return to text reference [1] [2]

First, the paralogs used to root the tree of life are usually joined in the longest central branch for each of the paralogs This is exactly the position the root would be attracted to by long-branch attraction, a well known artifact in phylogenetic reconstruction [62] [63]. The available analyses leave room for alternative explanations but some of these appear less likely than others. For example, moving the root to inside the group of archaea and eukaryotes is incompatible with derived structural features of the vacuolar and archaeal ATPases [25] [48]; whereas moving the root to inside the bacterial domain would be compatible with these features.

Second, organismal evolution is not tree-like. Horizontal gene transfer between organisms and the fusion of independent lines of descent transform organismal evolution into a net [13] [64] [65]. Whereas it is possible to define a last common ancestor for each molecular phylogeny (though it might be difficult to place the ancestor with any degree of confidence), the different molecular last common ancestors were not necessarily present in the same organism [14••]. Extracting the organismal net-like phylogeny from the many different, often ill resolved and incompatible molecular phylogenies remains a formidable task. The approach to consider genes with low substitution rates that were only infrequently transferred as a 'backbone' to organismal evolution and to

map major transfer events onto this tree is a possible approach [64]; however, the selection of backbone molecules, although justifiable, introduces some personal bias into the reconstruction.

The recognition of horizontal gene transfer as an important evolutionary process [13] [65] [66] [67] opens the door to very different scenarios describing organismal evolution, including the origin of eukaryotes. Although some features of the eukaryotes are shared with the archaea [68], others are bacterial. Current hypotheses at present include that the mitochondrial ancestor contributed all of the bacterial genes found in early eukaryotes [69], that an earlier symbiosis between an archaeon and a bacterium gave rise to the eukaryotic nucleocytoplasm [70] [71] [72] [73], or that the eukaryotic cell evolved as an independent lineage that did not survive into the present but the remnants of which might be recognized in some features of the eukaryotic cell [74] [75].

Gene and genome duplications, speciation, and innovation in eukaryotes

Gene duplications can occur on a gene by gene basis, resulting in tandem repeats, or they can result from the duplication of larger chromosomal regions or from polyploidization, including allopolyploidization — the retention of two different sets of chromosomes after hybridization between closely related species. The two completely sequenced eukaryotic genomes provide indications for both tandem and whole genome duplication. In their analyses of duplicated genes in the *C. elegans* genome, Semple and Wolfe [76••] found 40% of the genes to be members of duplicated gene pairs or families. As expected for autochthonous duplications they found an excess — beyond a random distribution of the gene pairs in the genome — of duplicated genes that resided on the same chromosome. 13% of the gene pairs were separated by fewer than five intervening genes. The authors found few small regional duplications but no evidence for large-scale or even genome duplications. In contrast, the *S. cerevisiae* genome [15] contains many duplicated chromosomal regions. It was suggested that these resulted from a genome duplication in yeast (tetraploidization) followed by rearrangement and deletion of superfluous duplicated genes [77]. Using a model with 70–100 reciprocal translocations and which retained 8% of the original genes in duplicate, Seoighe and Wolfe [78] were able to model the distribution of repeated gene clusters in the yeast genome.

Genome duplications and, in particular, allopolyploidization are frequent events in plant evolution. Ehrendorfer [79] estimated that at least 50% of the Cormophytes ('higher plants') and nearly all cultivated plants evolved in conjunction with allopolyploidization. Two complete genome duplications were postulated to have occurred during the early evolution of vertebrates (e.g. [3] [80] [81] [82] [83] [84] [85]) but this idea, mainly derived from mapping of gene clusters, remains controversial [86] [87••]. Hughes reconstructed the evolution of nine protein families important in development and also discusses the phylogenies of four additional protein families (see [86]). Of the resulting 13 phylogenies, only one was compatible with the postulated two successive genome duplications early in vertebrate evolution. Gene duplications undoubtedly played a key role in the evolution of developmental pathways, but it remains unclear to which extent these duplications occurred on a gene by gene basis or as the result of genome duplications. Complete or partial genome duplications, interspecies hybridization, and large-scale genomic rearrangements have the long recognized potential to explain genetic isolation and rapid formation of new species. These processes provide ample duplicated genes to evolve new functions and/or differential regulation of expression (e.g. [2] [88]). Although it is clear that these processes play an important role in some instances, their reconstruction and often even their unambiguous detection remains an ongoing challenge in genomic analyses [89].

Convergence versus concerted evolution

The analysis of molecular data often indicates several independent gene duplication events in closely related lines of descent (e.g. vacuolar ATPase subunits in plants and algae [7] [90]; archaeal DNA polymerases [91]; interferons [92]). In these cases, one explanation is to consider each branching event as a gene duplication and

to interpret the many parallel gene duplication events as convergent evolution. However, it is well known that after a duplication event gene conversion can maintain a high level of similarity between two copies of a gene present in the same genome, especially if the two genes remain in close proximity within the genome (see [93]). During a gene conversion event, part of one gene is copied into the other copy of the gene. The lengths of the gene conversion tracts — that is, the stretch of sequence that is actually replaced with sequence from the other copy — ranges from a few dozen to a couple of hundred nucleotides [94] [95] [96]. On the basis of a reconstructed molecular phylogeny and without further information, it is impossible to decide between the two possibilities: parallel-gene duplications versus a single-gene duplication followed by a gene conversion in one or both of the descending lineages (Fig. 1).

The evolution of type I interferons provides a good example of the problems involved: in mammals, these interferons are divided into subfamilies denoted as alpha, beta, delta, tau and omega. The largest of these subfamilies are the alpha interferons: in human, over a dozen paralogous genes are known to encode alpha interferons. The deepest split in the mammalian interferon I family is between the beta interferons and all of the other mammalian interferons [92]. When different avian type I interferons are included in a phylogenetic analysis together with the mammalian interferons, all of the mammalian interferons cluster together, suggesting that the so-called beta interferons in birds (IFN2) are not orthologs to the mammalian beta interferons but that in birds, alpha and beta interferons evolved from a separate and independent gene duplication [92]. However, the two chicken interferons (ChIFN1 and ChIFN2) share some features with the mammalian alpha and beta interferons, respectively. In particular, their regulation in response to viral and non viral stimuli corresponds to that of the mammalian alpha and beta interferons [96], suggesting that ChIFN1 should be considered an alpha and ChIFN2 a beta interferon. A possible reconciliation between the functional data and the phylogenetic reconstruction is the incorporation of partial gene conversion events. In this scenario, the gene duplication that gave rise to the alpha and beta interferons predates the divergence between birds and mammals. The response characteristics of the alpha and beta interferons had already been established in the last common ancestor of birds and mammals. However, at least in one of the two lineages, a partial gene conversion event homogenized most of the coding regions of the two paralogs, yielding a phylogenetic reconstruction that appears to indicate two independent gene duplications.

Conclusions

Increasing molecular evidence underscores the importance of gene duplication in evolution. Recognition of gene conversion and horizontal transfer as complicating factors will improve our understanding of the evolutionary processes that generated biochemical, cell biological and developmental innovations. Increasingly powerful search tools and the recognition that related proteins can diverge on the primary structural level, yet still reveal their relatedness in three-dimensional structure, will continue to help expand our understanding of protein superfamilies as new data arrives leading to a more complete picture of the fundamental role of paralogy in organismal evolution.

Acknowledgements

We thank the NASA Exobiology Program for support and Phillip Marcus for stimulating discussions on interferons.

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. Lazcano M, Miller S:
How long did it take for life to begin and evolve to cyanobacteria?
J Mol Evol 1994, **39**: 546–554. [Cited by]
Return to citation reference [1]

2. Haldane JBS: *The Causes of Evolution.* London: Longwood Green, 1933,
Return to citation reference [1] [2]

3. Ohno S: *Evolution by Genome Duplication.* Berlin: Springer Verlag, 1970,
Return to citation reference [1] [2]

4. Tartof KD:
Redundant genes.
Annu Rev Genet 1975, **9**: 355–385. [MEDLINE] [Cited by]
Return to citation reference [1]

5. Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T et al.:
Comparison of the complete protein sets of worm and yeast: orthology and divergence.
Science 1998, **282**: 2022–2028. [Full text] [MEDLINE] [Cited by]
Return to citation reference [1]

6. Fitch WS:
Distinguishing homologous from analogous proteins.
Syst Zool 1970, **19**: 99–113. [MEDLINE] [Cited by]
Return to citation reference [1] [2]

7. Gogarten JP, Starke T, Kibak H, Fichmann J, Taiz L:
Evolution and isoforms of V-ATPase subunits.
J Exp Biol 1992, **172**: 137–147. [Cited by]
Return to citation reference [1] [2]

8. Czelusniak J, Goodman M, Hewett-Emmett D, Weiss ML, Venta PJ, Tashian RE:
Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes.
Nature 1982, **298**: 297–300. [MEDLINE] [Cited by]
Return to citation reference [1]

9. Gogarten JP:
Which is the most conserved group of proteins? Homology — orthology, paralogy, xenology and the fusion of independent lineages.
J Mol Evol 1994, **39**: 541–543. [MEDLINE] [Cited by]
Return to citation reference [1]

10. Williams DM, Embley TM:
Microbial diversity: domains and kingdoms.
Annu Rev Ecol Syst 1996, **27**: 569–595. [Cited by]
Return to citation reference [1]

11. Koonin EV, Mushegian AR, Galperin MY, Walker DR:
Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel function and suggests a chimeric origin for the archaea.
Mol Microbiol 1997, **25**: 619–637. [MEDLINE] [Cited by]
Return to citation reference [1]

12. Gogarten JP, Hilario E, Olendzenski L: **Gene duplications and horizontal gene transfer during early evolution.** In *Evolution of Microbial Life.* Edited by Roberts DL, Sharp P, Alderson G, Collins M. Society for General Microbiology, 1996, **54**: 267–292.
Return to citation reference [1] [2]

13. Hilario E, Gogarten JP:
Horizontal transfer of ATPase genes — the tree of life becomes a net of life.
Biosystems 1993, **31**: 111–119. [MEDLINE] [Cited by]

Return to citation reference [1] [2] [3]

14. •• Doolittle WF:
Phylogenetic classification and the universal tree.
Science 1999, **284**: 2124–2129. [Full text] [MEDLINE] [Cited by]
This paper provides a chronicle on the recent paradigm change that has led to the incorporation of horizontal inheritance into out picture of early evolution.
Return to citation reference [1] [2] [3]

15. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M et al.:
Life with 6000 genes.
Science 1996, **274**: 546, 563–567. [Cited by]
Return to citation reference [1] [2]

16. Martin W, Schnarrenenberger C:
The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis.
Curr Genet 1997, **32**: 1–18. [MEDLINE] [Cited by]
Return to citation reference [1]

17. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD et al.:
Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*.
Science 1996, **273**: 1058–1073. [MEDLINE] [Cited by]
Return to citation reference [1]

18. Olendzenski L, Gogarten JP: **Gene duplications and horizontal gene transfer.** In *Thermophiles: The Key to Molecular Evolution and the Origin of Life?* Edited by Adams M, Weigel J. London: Taylor and Francis, 1998, 165–176.
Return to citation reference [1]

19. Pearson WR, Lipman DJ:
Improved tools for biological sequence comparison.
Proc Natl Acad Sci USA 1988, **85**: 2444–2448. [MEDLINE] [Cited by]
Return to citation reference [1]

20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ:
Basic local alignment search tool.
J Mol Biol 1990, **215**: 403–410. [MEDLINE] [Cited by]
Return to citation reference [1]

21. Gandbhir M, Rasched I, Marliere P, Mutzel R:
Convergent evolution of amino acid usage in archaebacterial and eubacterial lineages adapted to high salt.
Res Microbiol 1995, **146**: 113–120. [MEDLINE] [Cited by]
Return to citation reference [1]

22. Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ:
Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species.
Proc Natl Acad Sci USA 1999, **96**: 3578–3583. [Full text] [MEDLINE] [Cited by]
Return to citation reference [1]

23. Lipman DJ, Pearson WR:
Rapid and sensitive protein similarity searches.
Science 1985, **227**: 1435–1441. [MEDLINE] [Cited by]
Return to citation reference [1]

24. Doolittle RF:
Convergent evolution: the need to be explicit.
Trends Biochem Sci 1994, **19**: 15–18. [Cited by]
Return to citation reference [1]

25. Zimniak L, Dittrich P, Gogarten JP, Kibak H, Taiz L:
The cDNA sequence of the 69 kDa subunit of the carrot vacuolar H+-ATPase: homology to the beta-chain of F0F1-ATPases.

J Biol Chem 1988, **263**: 9102–9112. [MEDLINE] [Cited by]
Return to citation reference [1] [2]

26. Hilario E, Gogarten JP:
The prokaryote to eukaryote transition reflected in the evolution of the V/F/A-ATPase catalytic and proteolipid subunits.
J Mol Evol 1998, **46**: 703–715. [MEDLINE] [Cited by]
Return to citation reference [1] [2]

27. Vogler AP, Homma M, Irikura VM, Macnab RM:
***Salmonella typhimurium* mutants defective in flagellar filament regrowth and sequence similarity of FliI to F0F1, vacuolar, and archaebacterial ATPase subunits.**
J Bacteriol 1991, **173**: 3564–3572. [MEDLINE] [Cited by]
Return to citation reference [1]

28. Opperman T, Richardson JP:
Phylogenetic analysis of sequences from diverse bacteria with homology to the *Escherichia coli* rho gene.
J Bacteriol 1994, **176**: 5033–5043. [MEDLINE] [Cited by]
Return to citation reference [1]

29. Labedan B, Riley M:
Widespread protein sequence similarities: origins of *Escherichia coli* genes.
J Bacteriol 1995, **177**: 1585–1588. [Full text] [MEDLINE] [Cited by]
Return to citation reference [1]

30. Labedan B, Riley M:
Gene products of *Escherichia coli*: sequence comparisons and common ancestries.
Mol Biol Evol 1995, **12**: 980–987. [MEDLINE] [Cited by]
Return to citation reference [1]

31. Labedan B, Riley M:
Protein evolution viewed through *Escherichia coli* protein sequence introducing the notion of a structional segment of homology, the module.
J Mol Biol 1997, **268**: 857–868. [Full text] [MEDLINE] [Cited by]
Return to citation reference [1] [2] [3]

32. •• Labedan B, Riley M: **Genetic inventory: *Esherichia coli* as a window on ancestral proteins.** In *Organization of the Prokaryotic Genome.* Edited by Charlebois R. Washington DC: ASM Press, 1999, 311–329.
Using DARWIN, a ml matching program, the authors analyzed all *E. coli* proteins. Assessing matches of >80 amino acids with PAM scores or 250 or better, they found that most scores were in the range of 120–200, indicating divergence at about the same time; recent divergences were rare. This was interpreted to mean that there were windows in evolution when proteins could duplicate and diverge and that perhaps functional constraints hinder further divergence among related proteins. 65% of sequences longer than 79 amino acids were found to be paralogous. Grouping modules into families yielded 3717 paralogous modules: 1173 unimodular paralogous genes, 2629 paralogous proteins, and 873 families (with the same number of assumed ancestral genes).
Return to citation reference [1] [2]

33. Brenner SE, Chothia C, Hubbard TJ:
Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships.
Proc Natl Acad Sci USA 1998, **95**: 6073–6078. [Full text] [MEDLINE] [Cited by]
Return to citation reference [1] [2] [3]

34. Fan C, Moews PC, Shi Y, Walsh CT, Knox JR:
A common fold for peptide synthetases cleaving ATP to ADP: glutathione synthetase and D-alanine:D-alanine ligase of *Escherichia coli*.
Proc Natl Acad Sci USA 1995, **92**: 1172–1176. [MEDLINE] [Cited by]
Return to citation reference [1]

35. Matsuda K, Mizuguchi K, Nishioka T, Kato H, Go N, Oda J:
Crystal structure of glutathione synthetase at optimal pH: domain architecture and structural similarity with other proteins.
Prot Eng 1996, **9**: 1083–1092.
Return to citation reference [1]

36. Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF:
Protein sequence similarity searches using patterns as seeds.
Nucleic Acids Res 1998, **26**: 3986–3990. [MEDLINE] [Cited by]
Return to citation reference [1]

37. Retief JD, Lynch KR, Pearson WR:
Panning for genes — a visual strategy for identifying novel gene orthologs and paralogs.
Gen Res 1999, **9**: 373–382.
Return to citation reference [1]

38. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ:
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
Nucleic Acids Res 1997, **25**: 3389–3402. [MEDLINE] [Cited by]
Return to citation reference [1]

39. Galperin MY, Koonin EV:
A diverse superfamily of enzymes with ATP-dependent carboxylate-amine/thiol ligase activity.
Protein Sci 1997, **12**: 2639–2643. [Cited by]
Return to citation reference [1]

40. •• Wolf Y, Brenner SE, Bash PA, Koonin EV:
Distribution of protein folds in the three superkingdoms of life.
Gen Res 1999, **9**: 17–26.
14 proteomes were compared using PSI-BLAST to a database of protein fold identifiers. Folds for 20–30% of the proteins could be identified immediately. Distribution of folds were found to differ greatly between prokaryotes and eukaryotes with folds related to housekeeping genes (e.g. P-loop NTPases) being most common in prokaryotes, whereas eukaryotes had a predominance of domains involved in regulatory functions. The distribution of multidomain proteins is consistent with a model of origin by random protein combinations.
Return to citation reference [1] [2] [3]

41. Teichmann SA, Park J, Chothia C:
Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements.
Proc Natl Acad Sci USA 1998, **95**: 14648–14663. [Full text] [Cited by]
Return to citation reference [1]

42. Saier MH Jr, Eng BH, Fard S, Garg J, Haggerty DA, Hutchinson WJ, Jack DL, Lai EC, Liu HJ, Nusinew DP et al.:
Phylogenetic characterization of novel transport protein families revealed by genome analyses.
Biochim Biophys Acta 1999, **1422**: 1–56. [MEDLINE] [Cited by]
Return to citation reference [1]

43. Paulsen IT, Sliwinski MK, Saier MH Jr:
Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities.
J Mol Biol 1998, **277**: 573–592. [Full text] [MEDLINE] [Cited by]
Return to citation reference [1]

44. •• Aravind L, Koonin EV:
Gleaning non-trivial structural, functional, and evolutionary information about proteins by iterative database searches.
J Mol Biol 1999, **287**: 1023–1040. [Full text] [MEDLINE] [Cited by]
Demonstrates the usefulness PSI-BLAST in using sequence data to detect homology between proteins whose homology was previously only detectable by three-dimensional structure comparison. The authors point out the importance of using appropriate starting sequences for detection of these subtle similarities. Examples of superfamilies of protein folds detected using this method include HSP70/actin, ATP and AND dependent ligases, the ACT ligand binding domain, and a β propeller domain in recombinase subunits.
Return to citation reference [1] [2] [3]

45. •• Koonin EV, Tatusov RL, Galperin MY:
Beyond complete genomes: from sequence to structure and function.
Curr Opin Struct Biol 1998, **8**: 355–363. [Full text] [MEDLINE] [Cited by]
Recent review discussing improvements in delineating families of paralogs in completely sequenced genomes. Includes a table showing examples of protein superfamilies recently expanded as a result of improved techniques and additional

data.
Return to citation reference [1]

46. Schwarz RM, Dayhoff MO:
Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts.
Science 1978, **199**: 395–403. [MEDLINE] [Cited by]
Return to citation reference [1]

47. Gogarten JP, Taiz L:
Evolution of proton pumping ATPases: rooting the tree of life.
Photosyn Res 1992, **33**: 137–146.
Return to citation reference [1]

48. Gogarten JP, Kibak H, Dittrich P, Taiz L, Bownam EJ, Bowman B, Manolson M, Poole R, Date T, Oshima T et al.:
The evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes.
Proc Natl Acad Sci USA 1989, **86**: 6661–6665. [MEDLINE] [Cited by]
Return to citation reference [1] [2]

49. Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T:
Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes.
Proc Natl Acad Sci USA 1989, **86**: 355–359. [Cited by]
Return to citation reference [1]

50. Cammarano P, Palm P, Creti R, Ceccarelli E, Sanangelantoni AM, Tiboni O:
Early evolutionary relationships among known life forms inferred from elongation factor EF-2/EF-G sequences: phylogenetic coherence and structure of the archaeal domain.
J Mol Evol 1992, **34**: 396–405. [MEDLINE] [Cited by]
Return to citation reference [1]

51. Baldauf SL, Palmer JD, Doolittle WF:
The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny.
Proc Natl Acad Sci USA 1996, **93**: 7749–7754. [MEDLINE] [Cited by]
Return to citation reference [1]

52. Brown JR, Doolittle WF:
Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications.
Proc Natl Acad Sci USA 1995, **92**: 2441–2445. [MEDLINE] [Cited by]
Return to citation reference [1]

53. Gribaldo S, Cammarano P:
The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery.
J Mol Evol 1998, **47**: 508–516. [MEDLINE] [Cited by]
Return to citation reference [1]

54. Marsh TL, Reich CI, Whitelock RB, Olsen GJ:
Transcription factor IID in the Archaea: sequences in the *Thermococcus celer* genome would encode a product closely related to the TATA-binding protein of eukaryotes.
Proc Natl Acad Sci USA 1994, **91**: 4180–4184. [MEDLINE] [Cited by]
Return to citation reference [1]

55. Rowlands T, Baumann P, Jackson SP:
The TATA-binding protein: a general transcription factor in eukaryotes and archaebacteria.
Science 1994, **264**: 1326–1329. [MEDLINE] [Cited by]
Return to citation reference [1]

56. Pesole G, Gissi C, Lanave C, Saccone C:
Glutamine synthetase gene evolution in bacteria.
Mol Biol Evol 1995, **12**: 189–197. [MEDLINE] [Cited by]
Return to citation reference [1]

57. Koonin EV, Mushegian AR, Galperin MY, Walker DR:

Comparison of archaeal and bacterial genomes: computer analysis of portein sequences predicts novel functiona and suggests a chimeric origin for the archaea.
Mol Microbiol 1997, **25**: 619–637. [MEDLINE] [Cited by]
Return to citation reference [1]

58. Gupta RS, Singh B:
Cloning of the HSP70 gene from *Halobacterium marismoruti*: relatedness of archaebacterial HSP70 to its eubacterial homologs and a model of the evolution of the HSP70 gene.
J Bacteriol 1992, **174**: 4594–4605. [MEDLINE] [Cited by]
Return to citation reference [1]

59. Gupta RS, Golding GB:
Evolution of the HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria and eukaryotes.
J Mol Evol 1993, **37**: 573–582. [MEDLINE] [Cited by]
Return to citation reference [1]

60. Forterre P, Benanchenhou-Lafha N, Confalonieri F, Duguet M, Elie C, Labedan B:
The nature of the last universal ancestor and the root of the tree of life, still open questions.
Biosystems 1993, **28**: 15–32. [MEDLINE] [Cited by]
Return to citation reference [1]

61. Benanchenhou-Lafha N, Forterre P, Labedan B:
Evolution of glutamate dehydrogenase genes: evidence for two paralogous protein families and unusual branching patterns of the archaebacteria in the universal tree of life.
J Mol Evol 1993, **36**: 335–346. [MEDLINE] [Cited by]
Return to citation reference [1]

62. Felsenstein J:
Cases in which parsimony and compatibility methods will be positively misleading.
Syst Zool 1978, **27**: 401–410. [Cited by]
Return to citation reference [1]

63. Huelsenbeck PJ:
Performance of phylogenetic methods in simulation.
Syst Biol 1995, **44**: 17–48. [Cited by]
Return to citation reference [1]

64. Gogarten JP:
The early evolution of cellular life.
Trends Ecol Evol 1995, **10**: 147–151. [Cited by]
Return to citation reference [1] [2]

65. Woese C:
The universal ancestor.
Proc Natl Acad Sci USA 1998, **95**: 6854–6859. [Full text] [MEDLINE] [Cited by]
Return to citation reference [1] [2]

66. Lawrence JG, Roth JR:
Selfish operons: horizontal transfer may drive the evolution of gene clusters.
Genetics 1996, **143**: 1843–1860. [MEDLINE] [Cited by]
Return to citation reference [1]

67. Lorenz MG, Wackernagel W:
Bacterial gene transfer by natural genetic transformation in the environment.
Microbiol Rev 1994, **58**: 563–602. [MEDLINE] [Cited by]
Return to citation reference [1]

68. Keeling PJ, Doolittle WF:
Archaea: narrowing the gap between prokaryotes and eukaryotes.
Proc Natl Acad Sci USA 1995, **92**: 5761–5764. [MEDLINE] [Cited by]
Return to citation reference [1]

69. Doolittle WF: **Some aspects of the biology of cells and their possible evolutionary significance.** In *Evolution of Microbial Life.* Edited by Roberts DL, Sharp P, Alderson G, Collins M. Society for General Microbiology, 1996, **54**: 1–21.
Return to citation reference [1]

70. Zillig W, Palm P, Klenk HP: **A model of the early evolution of organisms: the arisal of the three domains of life from the common ancestor.** In *The Origin and Evolution of the Cell.* Edited by Hartman H, Matsuno K. Singapore: World Scientific, 1992, 163–182.
Return to citation reference [1]

71. Martin W, Muller M:
The hydrogen hypothesis for the first eukaryote.
Nature 1998, **392**: 37–41. [Full text] [MEDLINE] [Cited by]
Return to citation reference [1]

72. Searcy D: **Origins of mitochondria and chloroplasts from sulfur-based symbioses.** In *The Origin and Evolution of the Cell.* Edited by Hartman H, Matsuno K. Singapore: World Scientific, 1992, 47–78.
Return to citation reference [1]

73. Lake JA, Rivera MC:
Was the nucleus the first endosymbiont?
Proc Natl Acad Sci USA 1994, **91**: 2880–2881. [MEDLINE] [Cited by]
Return to citation reference [1]

74. Hartman H:
The origin of the eukaryotic cell.
Spec Sci Technol 1984, **7**: 77–81.
Return to citation reference [1]

75. Sogin ML, Silberman JD, Hinkle G, Morrison HG: **Problems with molecular diversity in the Eukarya.** In *Evolution of Microbial Life.* Edited by Roberts DL, Sharp P, Alderson G, Collins M. Society for General Microbiology, 1996, **54**: 167–184.
Return to citation reference [1]

76. •• Semple C, Wolfe KH:
Gene duplication and gene conversion in the *Caenorhabditis elegans* genome.
J Mol Evol 1999, **48**: 555–564. [MEDLINE] [Cited by]
A comprehensive analysis of duplication and gene conversion for 7394 *C. elegans* genes is presented. Of the genes examined, 40% are involved in duplicated gene pairs. Intrachromosomal or *cis* gene duplications occur approximately twice more often than expected. Gene conversion events are detectable between only 2% of the duplicated pairs. Three recent, regional duplications, each spanning three genes and already having undergone substantial deletions spanning hundreds of base pairs, are described. The relative rates of duplication and deletion may contribute to the compactness of the *C. elegans* genome.
Return to citation reference [1]

77. Wolfe KH, Shields DC:
Molecular evidence for an ancient duplication of the entire yeast genome.
Nature 1996, **387**: 708–713. [Full text] [MEDLINE] [Cited by]
Return to citation reference [1]

78. Seoighe C, Wolfe KH:
Extent of genomic rearrangement after genome duplication in yeast.
Proc Natl Acad Sci USA 1998, **95**: 4447–4452. [Full text] [MEDLINE] [Cited by]
Return to citation reference [1]

79. Ehrendorfer F: *Evolution und Sytematik in Lehrbuch der Botanik,* edn 33. Stuttgart: Gustav Fischer Verlag, 1991,
Return to citation reference [1]

80. Comings DE:
Evidence for ancient tetraploidy and conservation of linkage groups in mammalian chromosomes.
Nature 1972, **238**: 455–457. [MEDLINE] [Cited by]
Return to citation reference [1]

81. Lundin LG:
Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse.

Genomics 1993, **16**: 1–19. [MEDLINE] [Cited by]
Return to citation reference [1]

82. Morizot DC:
Comparative gene mapping evidence for chromosome duplications in chordate evolution.
Isozyme Bull 1986, **19**: 9–10.
Return to citation reference [1]

83. Nadeau JH, Sankoff D:
Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution.
Genetics 1997, **147**: 1259–1266. [MEDLINE] [Cited by]
Return to citation reference [1]

84. Postlethwait JH, Yan YL, Gates MA, Horne S, Amores A, Brownlie A, Donovan A, Egan ES, Force A, Gong Z et al.:
Vertebrate genome evolution and the zebrafish gene map.
Nat Genet 1998, **18**: 345–349. [MEDLINE] [Cited by]
Return to citation reference [1]

85. Pebusque MJ, Coulier F, Birnbaum D, Pontarotti P:
Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution.
Mol Biol Evol 1998, **15**: 1145–1159. [MEDLINE] [Cited by]
Return to citation reference [1]

86. Hughes A:
Phylogenies of developmentally important proteins do not suport the hypothesis of two rounds of genome duplication early in vertebrate history.
J Mol Evol 1999, **48**: 565–576. [MEDLINE] [Cited by]
Return to citation reference [1] [2]

87. •• Skrabanek L, Wolfe KH:
Eukaryote genome duplication — where's the evidence?
Curr Opin Genet Dev 1998, **8**: 694–700. [Full text] [MEDLINE] [Cited by]
Reviews literature showing that although maize, yeast and *Xenopus* have experienced recent whole-genome duplications, Ohno's original hypothesis that a gene duplication occurred in an ancestor of vertebrates is still unproved, largely because of the shortage of gene sequence and map data.
Return to citation reference [1]

88. Goldsmith R: *The Material Basis of Evolution.* New Haven: Yale University Press, 1940,
Return to citation reference [1]

89. Ruddle FH:
Vertebrate genome evolution — the decade ahead.
Genomics 1997, **46**: 171–173. [Full text] [MEDLINE] [Cited by]
Return to citation reference [1]

90. Ikeda M, Konishi K, Kadowaki H, Moritani C, Watanabe Y:
Molecular cloning of cDNAs encoding *Acetabularia acetabulum* V type ATPase, A and B subunits.
Plant Physiol 1996, **111**: 651. [Cited by]
Return to citation reference [1]

91. Edgell DR, Malik S-B, Doolittle WF:
Evidence of independent gene duplications during the evolution of archaeal and eukaryotic family B DNA polymerases.
Mol Biol Evol 1998, **15**: 1207–1217. [MEDLINE] [Cited by]
Return to citation reference [1]

92. Roberts RM, Liu L, Guo Q, Leaman D, Bixby J:
The evolution of type one interferons.
J Interferon Cytokine Res 1998, **18**: 805–816. [MEDLINE] [Cited by]
Return to citation reference [1] [2] [3]

93. Li W: *Molecular Evolution.* Sunderland, Massachusetts: Sinauer Associates, Inc., 1997,
Return to citation reference [1]

94. Yang D, Waldman AS:
Fine-resolution analysis of products of intrachromosomal homeologous recombination in mammalian cells.
Mol Cell Biol 1997, **17**: 3614–3628. [Full text] [MEDLINE] [Cited by]
Return to citation reference [1]

95. Betran E, Rozas J, Navarro A, Barbadilla A:
The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data.
Genetics 1997, **146**: 89–99. [MEDLINE] [Cited by]
Return to citation reference [1]

96. Sweetser DB, Hough H, Whelden JF, Arbuckle M, Nickoloff JA:
Fine-resolution mapping of spontaneous and double-strand break-induced gene conversion tracts in *Saccharomyces cerevisiae* reveals reversible mitotic conversion polarity.
Mol Cell Biol 1994, **14**: 3863–3875. [MEDLINE] [Cited by]
Return to citation reference [1] [2]

97. Sick C, Schultz U, Munster U, Meier J, Kaspers B, Staeheli P:
Promoter structures and differential responses to viral and nonviral inducers of chicken type I interferon genes.
J Biol Chem 1998, **273**: 9749–9754. [Full text] [MEDLINE] [Cited by]

Author Contacts

JP Gogarten and L Olendzenski, Department of Molecular and Cell Biology, 75 North Eagleville Road U-44, University of Connecticut, Storrs, Connecticut 06269, USA. E-mail address for JP Gogarten: gogarten@uconnvm.uconn.edu. E-mail address for L Olendzenski: lco95001@uconnvm.uconn.edu.

Return to author list

Copyright

Copyright © 1999 Current-Opinion.com

Help Feedback BioMedNet Home Go Search Map

© 1998-2000 BioMedNet. All rights reserved.
email: info@biomednet.com.