ELSEVIER

Evolutionary aspects of whole-genome biology

Russell F Doolittle

A decade of access to whole-genome sequences has been increasingly revealing about the informational network relating all living organisms. Although at one point there was concern that extensive horizontal gene transfer might hopelessly muddle phylogenies, it has not proved a severe hindrance. The melding of sequence and structural information is being used to great advantage, and the prospect exists that some of the earliest aspects of life on Earth can be reconstructed, including the invention of biosynthetic and metabolic pathways. Still, some fundamental phylogenetic problems remain, including determining the root — if there is one — of the historical relationship between Archaea, Bacteria and Eukarya.

Addresses
Department of Chemistry & Biochemistry, University of California, San Diego, La Jolla, CA 92093-0314, USA

Corresponding author: Doolittle, Russell F (rdoolittle@ucsd.edu)

Current Opinion in Structural Biology 2005, **15**:248–253

This review comes from a themed issue on
Sequences and topology
Edited by Steven E Brenner and Anna Tramontano

Available online 25th April 2005

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2005.04.001

Introduction

Historically, the primary goal of the study of molecular evolution is to reconstruct past events in a way that explains the present living world. Ultimately, if the evidence has not been overly blurred by time, all trails should lead back to a common ancestral cell type. Over the years, macromolecular sequence information has been applied effectively towards this end, even in the face of major complications resulting from vastly unequal rates of change along different lineages, horizontal transfers of genes and gene clusters, and numerous other distractions. That these efforts have succeeded as well as they have must be regarded as a major triumph.

Although the enterprise has been ongoing for half a century, it's only during the past decade that whole-genome sequences have been available [1]; the question needs to be asked how this resource has affected the quest. In a word, immensely. Not only are organism connections at all levels being better established, but the full extent of the proliferation of gene families and the protein structures that underlie cellular divergences is also being greatly extended. In this brief review, I attempt to highlight some of the most impressive advances that whole-genome studies have contributed to our views of evolution.

Gene recognition

From its beginning, the whole-genome enterprise depended heavily on the premise that most genes would be readily identified by computer analysis alone. The basis of this hope was that most — if not all — extant genes are descendants from a smaller ancestral population that has been expanded by gene duplication. As such, identifications would be made by comparison with known genes and gene products whose functions had been determined experimentally. Lurking beyond the simple hope that a function would be assigned to every identified gene was an even more optimistic view: that it would eventually be possible to match every putative gene product with a known homologous three-dimensional structure.

It was somewhat disappointing, then, to find that, among the first half-dozen microbial genomes to be completed, almost half of all open reading frames (ORFs) were URFs (the 'U' denotes 'unidentified') [2]. After more than a century of study by biochemists and microbial geneticists, how could there be so many unrecognizable genes? Were these unrecognizable ORFs the consequence of anomalously accelerated rates of sequence change? Were all these ORFs really true genes? In fact, many of them were quite short and might not even be expressed. But others fell well within the size range of average genes and were found also in other genomes, suggesting they were indeed authentic.

The trend continued as more whole genomes were reported, the fraction of ORFs without identifiable counterparts in sequence databases hardly declining [3,4]. The absolute number grew to the point at which there were 20 000 ORFans (formerly URFs) gathered and cataloged in a single microbial database [5•,6•,7]. On the bright side, the fraction seems now to be diminishing as searching regimens improve [8,9]. For example, the use of a fold recognition algorithm is making connections that were previously missed when only sequences were being considered [10•,11]. Still, it remains mysterious how these ORFans have become so different from their alleged nearest relatives.

Whole-genome trees

The appearance of the first several whole-genome sequences quickly led to attempts to reconstruct

phylogenetic trees based on them. Every possible derivation of the information was put to use, it seemed, and whole-genome trees were generated based on sequence [12–15] and gene content [16,17], as well as on structural attributes (treated separately below). Not surprisingly, there was a degree of incongruity among trees made by different strategies, although combinations of various methods led to consensus trees that seemed reasonable [18,19]. Nonetheless, there were problems and surprises.

Horizontal gene transfers

One complication that arose during attempts to generate whole-genome phylogenies had to do with a substantial number of genes that, when examined by standard phylogenetic methods, were clearly out of order with the parental genome trees. This was particularly true with regard to prokaryotic genomes. The simplest explanation was that the genes had been transferred horizontally. At first glance, the unexpectedly large number of such occurrences seemed to make the prospect of establishing a stable phylogeny unlikely [20]. With time and further reflection, however, it became clear that phylogenetic construction was not going to be seriously impaired, so long as one remained alert to the very real prospect of some genes or gene clusters having been acquired laterally rather than by vertical descent [21–23,24•].

In many cases, horizontal gene transfers are extremely interesting on their own [25]. Indeed, many reflect some of the most innovative adaptations in all of biology — including bacterial photosynthesis and nitrogen fixation, both probably having resulted from a series of cellular 'barters' [26•,27•]. Such transfers are not restricted to single genes; genes, operons and much more massive assemblies are commonly exchanged among prokaryotes. Sometimes, as in the case of 'pathogenicity islands', they can involve virtual armadas of biological warfare agents [4].

Gene loss

Another phenomenon that became 'visible' when every gene in a genome could be counted was the loss of genes along different lineages [28]. It was found, for example, that almost 400 genes were lost along each of the lineages leading from the common ancestor of fission and budding yeasts [29]. In the pre-genome era, most biologists regarded gene loss as a truly calamitous event that could affect the entire interaction network within a cell. That so many genes could be lost in such a relatively short time implied a much more fluid genome than had been anticipated. Significantly, a large fraction of horizontally transferred genes is quickly lost [30•]. If the moment of opportunity is not exploited, the event is doomed. In the final analysis, there is good reason to think that gene loss is the principal determinant of gene content [31•,32].

Table 1

Size and gene content of some reduced genomes.

Organism	Genome size (Mb)	Genes
Nanoarchaeum equitans	0.49	536
Mycoplasma genitalium	0.58	484
Buchnera aphidicola	0.64	596
Chlamydia trachomatis	1.04	895
Rickettsia prowazekii	1.11	835
Mycobacterium leprae	3.2	1604
Encephalitozoon cuniculi	2.5	1996
Cryptosporidium parvum	9.1	3807

Reduced genomes

A variety of whole-genome sequences have been determined for parasitic organisms that have adapted to an existence with severely reduced genomes. These organisms have jettisoned much of their own enzymatic equipment and live off the metabolic resources of their host cells. The phenomenon occurs in all three superkingdoms (Table 1). Examples among Bacteria include familiar organisms such as Mycoplasmas [33], *Chlamydia* [34], *Buchnera* [35] and *Rickettsia* [36], the last named being especially interesting because of a kinship with likely antecedents of mitochondria. Adaptations to parasitic existence are idiosyncratic, different parasite genomes losing different sets of genes on their way to dependence [4]. In this regard, a fascinating situation is afforded by *Mycobacterium leprae*, a bacterium that is still in the process of decay and that still contains a slew of inactive genes on their way to random oblivion or elimination [37].

A very different situation exists in a unique archaeal parasite, *Nanoarchaeum equitans*, already firmly established with a minimal genome [38]. *N. equitans* gains most of its metabolic needs from another archaeon, *Ignicoccus*, to which it is obligately attached [38]. By contrast, *Ignicoccus* can live very well without the parasite. In passing, it is worth noting that — although archaeal endosymbionts (as opposed to parasites) are common in eukaryotic cells, especially in amoebic cells [39] — archaeal pathogens have never been found.

Reduced parasitic genomes are also found among the Eukarya, the first of which to be determined was the microsporidium *Encephalitozoon cuniculi*, an obligate intracellular pathogen [40]. The apicomplexan *Cryptospora parvum*, although not as small as *E. cuniculi* (Table 1), is interesting in that even its organelles (plastids and mitochondria) have lost their DNA [41•].

Minimal genomes

The initial reports of small bacterial genome sequences led to speculation as to what would constitute a minimal set of genes [42]. Reduced genomes are prisoners of their history, having descended from more complex

circumstances, and the minimal sets of genes they may contain are clearly different from the minimum needed by a free-living organism. In this regard, it is of interest that systematic gene inactivation experiments have shown that a free-living bacterium such as *Bacillus subtilis*, which ordinarily has about 4100 genes, only requires 271 of them to survive [43]. A recent theoretical analysis of what would be the minimal set of genes for a free-living bacterium arrived at the similar number of 206 [44•].

These numbers must be regarded skeptically. They are based on what we know about modern and sophisticated organisms. It's the primitive ones that need to be reckoned. The first step to take backwards in time may be to identify the gene content of the last common ancestor of all current life: the number of genes involved in that hypothetical case usually being thought to be 1000–2000, if only because that is how many genes are found in the smallest extant free-living organisms. The real challenge will be to determine the gene content of the earliest cells, which, at some point, must have been very small indeed.

Introns, splicing and the origin of eukaryotes

Quite apart from it being more difficult to identify genes in eukaryotes because of the intronic disruption of coding regions, we might also ask what effect whole-genome studies have had on the long-standing 'introns early-introns late' debate. In fact, it seems to have provided ammunition for both sides. That introns are gained and lost by modern eukaryotic genomes at a confounding rate seems indisputable [45,46], the gains being regarded by the 'introns late' school as clear support for their side [47]. Contrarily, the fact that losses seem to outnumber gains is used as an argument to support the 'introns early' position [48]. It seems to me that 'early or late' should not have been made the crux of the argument; rather, the focus should be on those original claims that extant exons are vestigial remnants of the earliest proteins [49]. A full appreciation of the history of protein structures requires that these matters be settled more fully [50].

Refining relationships

Whole-genome projects are definitely improving the overall quality of phylogenies, chipping away, for example, at the thorny problem of how the major bacterial phyla are related at the deepest levels [18]. Archaeal phylogenies need similar study, as evidenced by ribosomal trees usually not being congruent with whole-genome trees made on the basis of non-ribosomal attributes. That such contradictions can be overcome when sufficient data are brought to bear is shown by the recent successful clustering of very distantly related amoebas into a clade that conforms perfectly to classical biology [51•,52].

The interactome

Over and beyond tracing the history of organisms and their proteins, there has been significant progress made in determining which gene products interact with each other and how the general outlines of metabolism evolved. Previously, interactions between macromolecules had to be determined experimentally, the yeast two-hybrid system having become the strategy of choice for finding interacting macromolecules [53]. The initial *in silico* tactic was simply to look for genes that were adjacent or very near to each other more often than would be expected [54,55]. Methods have improved to the point at which they now rival experimental approaches [56]. The interaction of its proteins and other macromolecules is the backbone of the cell's machinery, upon which its survival depends.

Melding sequence and structural attributes

From the start, the whole-genome project was the beneficiary of remarkable advances in structural biology. During the past decade, the number of three-dimensional structures in the Protein Data Bank (PDB) [57] has swelled to more than 20 000 entries. Concurrently, the SCOP (Structural Classification of Proteins) database [58,59] has been parsing the PDB structures into their constituent domains — more than 50 000 in version 1.65 — all sorted hierarchically by structural type. The process begins by assigning each domain to one of the six major groups (all-α, all-β, α/β, α+β, membrane associated and small). In version 1.65, these have been further grouped into 800 folds, followed by a further subdivision into 1294 superfamilies. The fold level is defined as comprising domains with similar arrangements of secondary structure elements but which may not necessarily reflect common ancestry. The superfamily level of folds is defined as those folds for which there is good evidence of common ancestry, even if sequence similarity is not obvious. The family level is reserved for those members for which sequence similarity reflecting common ancestry is obvious.

The bridge to whole-genome biology is currently provided by the SUPERFAMILY database [60,61]. Thus, every ORF from every whole genome is searched against the SCOP holdings in an effort to match it with a known structure. Most current search procedures rely on sensitive hidden Markov models [62], the effectiveness of which is attested to by more than 60% of ORFs from 174 completed genomes having been assigned to domain superfamilies [61].

Protein folds and whole genomes

It didn't take long for structural biologists to tally up the putative domain structures in the newly determined whole genomes [63–69,70••,71]. These studies have taken several directions, but all managed to count the different folds in the various superkingdoms and to show

how phylogenies could be rendered from them. These studies also provided data about the relative abundances of different kinds of protein domains overall, there being general agreement that α/β domains are the most common structural element. There was much agreement, also, that the distribution of structural features is in general accord with the tripartite nature of the 'tree of life', Archaea, Bacteria and Eukarya having distinguishable fold contents [63–69,70••,71].

The tripartite tree of life

Nonetheless, the conundrum of how that triumvirate evolved remains. Broadly speaking, there are three general schools of thought on the matter. First, there are those who feel that a divergence leading, on the one hand, to a cytoskeleton-containing pre-eukaryote, on the other, to prokaryotes (including the ancestors of both Archaea and Bacteria) occurred very early. Subsequently, a series of phagocytic events (in which pre-eukaryotes engulfed prokaryotes) gave rise to modern eukaryotes [72–75]. Another group feels that the invention of the Eukarya was the result of a chimeric merging of a member of the Bacteria with an archaeon, the latter assuming the form of the nucleus in the new entity [76,77]. A third interpretation is that all three lines originated at a time when the totality of life on Earth was a simple community of heterogeneous cells that shared only the ability to synthesize proteins with a ribosomal machine and that freely exchanged genetic material [78–80]. Will whole-genome biology be able to answer this question? Let us be hopeful.

Pushing backwards in time

Disputes about the three major domains of life aside, progress is being made concerning events that must have occurred well before the last common ancestor, especially with regard to protein structures. Aravind *et al*. [81•] have reported a convincing analysis of more than a dozen kinds of nucleotide-binding domains that occur in all living organisms ('Rossmannoid domains'). They constructed a phylogenetic tree depicting the evolution of the various types from a common structure that logically pre-dates the last common ancestor.

In an even bolder maneuver, Caetano-Anolles and Caetano-Anolles [70••] constructed phylogenetic trees of all known protein folds. Using strict cladistic principles and straightforward measurements of fold usage as a fundamental character for a hierarchal reconstruction, they showed how α/β proteins originated first and have given rise to the other main protein classes. Moreover, they were able to depict logical scenarios for the evolution of known folds within each of the main classes. Although the underlying premise of this approach — that ''redundancy is a favored evolutionary outcome'' — might be challenged, these are enormously interesting studies that deserve much scrutiny and discussion. It can't be a coincidence that α/β domains are the most predominant type of structure in contemporary proteins and that the glycolytic pathway is composed of proteins that are almost exclusively made from such domains.

Conclusions

The first decade of whole-genome biology has been exciting; it has taught us a great deal about how genomes evolve. But there is much more to come and the next decade should teach us a lot more. There is definite promise that the clarity of seeing backwards in time will improve. I feel confident that the notion of the last common ancestor will be revalidated and — I hope — the controversy over the root of the three-superkingdom triad finally settled. One approach might be to construct trees based on 'interactome' comparisons. For example, actin is well known to have a multitude of interactants. It should be possible to reconstruct the history of how these various associations accumulated, during the course of time, in parallel with eukaryotic attributes such as phagocytosis. In the end, we may be able to glimpse behind the curtain that separates us from life before the last common ancestor, perhaps even into an ancient RNA-protein world.

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. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al.*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd**. *Science* 1995, **269**:496-512.

2. Doolittle RF: **A bug with excess gastric acidity**. *Nature* 1997, **388**:515-516.

3. Doolittle RF: **Microbial genomes opened up**. *Nature* 1998, **392**:339-342.

4. Doolittle RF: **Microbial genomes multiply**. *Nature* 2002, **416**:697-700.

5. • Siew N, Fischer D: **The ORFanage: an ORFan database**. *Nucleic Acids Res* 2004, **32**:D281-D283.
Web-available tallies of various classes of unique ORFs.

6. • Siew N, Fischer D: **Analysis of singleton ORFans in fully sequenced microbial genomes**. *Proteins* 2003, **53**:241-251.
A systematic categorization of ORFans from 60 microbial genomes.

7. Siew N, Fischer D: **Twenty thousand ORFan microbial protein families for the biologist?** *Structure* 2003, **11**:7-9.

8. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C: **Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods**. *J Mol Biol* 1998, **284**:1201-1210.

9. Brent MR, Guigo R: **Recent advances in gene structure prediction**. *Curr Opin Struct Biol* 2004, **14**:264-272.

10. • Siew N, Fischer D: **Structural biology sheds light on the puzzle of genomic ORFans**. *J Mol Biol* 2004, **342**:369-373.
Combing the PDB for ORFans can provide a route to the identification of others.

11. Fischer D: **3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor**. *Proteins* 2003, **51**:434-441.

12. Tekaia F, Lazcano A, Dujon B: **The genome tree as revealed from whole proteome comparisons**. *Genome Res* 1999, **9**:550-557.

13. House CH, Fitz-Gibbon ST: **Using homolog groups to create a whole-genome tree of free-living organisms: an update**. *J Mol Evol* 2002, **54**:539-547.

14. Wolf YI, Rogozin IB, Grishin NV, Koonin EV: **Genome trees and the tree of life**. *Trends Genet* 2002, **18**:472-479.

15. Bansal AK, Meyer TE: **Evolutionary analysis by whole-genome comparisons**. *J Bacteriol* 2002, **184**:2260-2272.

16. Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene content**. *Nat Genet* 1999, **21**:108-110.

17. Korbel JO, Snel B, Huynen MA, Bork P: **SHOT: a web server for the construction of genome phylogenies**. *Trends Genet* 2002, **18**:158-162.

18. Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV: **Genome trees constructed using five different approaches suggest new major bacterial clades**. *BMC Evol Biol* 2001, **1**:8.

19. Snel B, Bork P, Huynen MA: **Genomes in flux: the evolution of archaeal and proteobacterial gene content**. *Genome Res* 2002, **12**:17-25.

20. Doolittle WF: **Phylogenetic classification and the universal tree**. *Science* 1999, **284**:2124-2129.

21. Glansdorff N: **About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal**. *Mol Microbiol* 2000, **38**:177-185.

22. Gogarten JP: **Gene transfer: gene swapping craze reaches eukaryotes**. *Curr Biol* 2003, **13**:R53-R54.

23. Gogarten JP, Doolittle WF, Lawrence JG: **Prokaryotic evolution in light of gene transfer**. *Mol Biol Evol* 2002, **19**:2226-2238.

24. • Bapteste E, Boucher Y, Leigh J, Doolittle WF: **Phylogenetic reconstruction and lateral gene transfer**. *Trends Microbiol* 2004, **12**:406-411.
The perils of discounting horizontal gene transfers in constructing phylogenies are emphatically put forth.

25. Jain R, Rivera MC, Lake JA: **Horizontal gene transfer accelerates genome innovation and evolution**. *Mol Biol Evol* 2003, **20**:1598-1602.

26. • Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE: **Whole-genome analysis of photosynthetic prokaryotes**. *Science* 2002, **298**:1616-1620.
Comparisons of the proteins involved in photosynthesis in all five bacterial phyla that are capable of photosynthesis show a history of frequent exchange and exploitation.

27. • Raymond J, Siefert JL, Staples CR, Blankenship RE: **The natural history of nitrogen fixation**. *Mol Biol Evol* 2004, **21**:541-554.
Unexpectedly, components of the critical enzyme nitrogenase share ancestry with enzymes responsible for synthesizing photosynthetic pigments, uniting the phenomena of nitrogen fixation and photosynthesis.

28. Roelofs J, Van Hastert PJM: **Genes lost during evolution**. *Nature* 2001, **411**:1013-1014.

29. Aravind L, Watanabe H, Lipman DJ, Koonin EV: **Lineage specific loss and divergence of functionally linked genes in eukaryotes**. *Proc Natl Acad Sci USA* 2000, **97**:11319-11324.

30. • Liu Y, Harrison PM, Kunin V, Gerstein M: **Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes**. *Genome Biol* 2004, **5**:R64.
Analysis shows that many horizontal gene transfers are doomed to random oblivion.

31. • Kunin V, Ouzounis CA: **The balance of driving forces during genome evolution in prokaryotes**. *Genome Res* 2003, **13**:1589-1594.
A thorough characterization of 51 prokaryotic genomes reveals that gene loss is the most important factor in shaping gene content.

32. Krylov DM, Wolf YI, Rogozin IB, Koonin EV: **Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution**. *Genome Res* 2003, **13**:2229-2235.

33. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM *et al.*: **The minimal gene complement of *Mycoplasma genitalium***. *Science* 1995, **270**:397-403.

34. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q *et al.*: **Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis***. *Science* 1998, **282**:754-759.

35. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H: **Genome sequence of the endocellular bacterial symbion of aphids *Buchnera sp* APS**. *Nature* 2000, **407**:81-86.

36. Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG: **The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria**. *Nature* 1998, **396**:133-140.

37. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D *et al.*: **Massive gene decay in the leprosy bacillus**. *Nature* 2001, **409**:1007-1011.

38. Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, Barnstead M, Beeson KY, Bibbs L, Bolanos R, Keller M *et al.*: **The genome of *Nanoarchaeum equitans*: insights into archaeal evolution and derived parasitism**. *Proc Natl Acad Sci USA* 2003, **100**:12984-12988.

39. Lee JL, Soldo AT, Riesser W, Lee MJ, Jeon KW, Gortz H-D: **The extent of algal and bacterial endosymbioses in protozoa**. *J Protozool* 1985, **32**:391-403.

40. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P *et al.*: **Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi***. *Nature* 2001, **414**:450-453.

41. • Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S *et al.*: **Complete genome sequence of the apicomplexan, *Cryptosporidium parvum***. *Science* 2004, **304**:441-445.
The characterization of the genome of an extraordinary parasitic organism.

42. Mushegian AR, Koonin EV: **A minimal gene set for cellular life derived by comparison of complete bacterial genomes**. *Proc Natl Acad Sci USA* 1996, **93**:10268-10273.

43. Kobayashi SD, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P *et al.*: **Essential *Bacillus subtilis* genes**. *Proc Natl Acad Sci USA* 2003, **100**:4678-4683.

44. • Gil R, Silva FJ, Pereto J, Moya A: **Determination of the core of a minimal bacterial gene set**. *Microbiol Mol Biol Rev* 2004, **68**:518-537.
A heroic effort to estimate the minimum number of genes required for a hypothetical free-living bacterium.

45. Sverdlov AV, Babenko VN, Rogozin IB, Koonin EV: **Preferential loss and gain of introns in 3′ portions of genes suggests a reverse-transcription mechanism of intron insertion**. *Gene* 2004, **338**:85-91.

46. Coghlan A, Wolfe KH: **Origins of recently gained introns in *Caenorhabditis***. *Proc Natl Acad Sci USA* 2004, **101**:11362-11367.

47. Logsdon JM: **Worm genomes hold the smoking guns of intron gain**. *Proc Natl Acad Sci USA* 2004, **101**:11195-11196.

48. Roy SW, Gilbert W: **The pattern of intron loss**. *Proc Natl Acad Sci USA* 2005, **102**:713-718.

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

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

51. • Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Durufle L, Gaasterland T, Lopez P, Muller M, Phillipe H: **The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba***. *Proc Natl Acad Sci USA* 2002, **99**:1414-1419.
Genomics and classical biology converge to reveal an important history.

52. Cavalier-Smith T: **A revised six-kingdom system of life**. *Biol Rev Camb Philos Soc* 1998, **73**:203-266.

53. Fields S, Song OK: **A novel genetic system to detect protein-protein interactions**. *Nature* 1989, **340**:245-246.

54. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact**. *Trends Biochem Sci* 1998, **23**:324-328.

55. Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences**. *Science* 1999, **285**:751-753.

56. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, Marcotte EM: **Protein interaction networks from yeast to human**. *Curr Opin Struct Biol* 2004, **14**:292-299.

57. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**:235-242.

58. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**:536-540.

59. Andreeva A, Howarth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2004: refinements integrate structure and sequence family data**. *Nucleic Acids Res* 2004, **32**:D226-D229.

60. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure**. *J Mol Biol* 2001, **313**:903-919.

61. Madera M, Vogel C, Kummerfield SK, Chothia C, Gough J: **The superfamily database in 2004: additions and improvements**. *Nucleic Acids Res* 2004, **32**:D235-D239.

62. Karplus K, Barrett C, Hughey R: **Hidden Markov models for detecting remote protein homologies**. *Bioinformatics* 1998, **14**:846-856.

63. Gerstein M: **A structural census of genomes: comparing eukaryotic, bacterial and archaeal genomes in terms of protein structure**. *J Mol Biol* 1997, **274**:562-576.

64. Gerstein M: **Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census**. *Proteins* 1998, **33**:518-534.

65. Gerstein M, Hegyi H: **Comparing genomes in terms of protein structure: surveys of a finite parts list**. *FEMS Microbiol Rev* 1998, **22**:277-304.

66. Wolf YI, Brenner SE, Bash PA, Koonin EV: **Distribution of protein folds in the three superkingdoms of life**. *Genome Res* 1999, **9**:17-26.

67. Lin J, Gerstein M: **Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels**. *Genome Res* 2000, **10**:808-818.

68. Hegyi H, Lin J, Gerstein M: **Structural genomics analysis: characteristics of atypical, common, and horizontally transferred folds**. *Proteins* 2002, **47**:126-141.

69. Qian J, Luscombe NM, Gerstein M: **Protein family and fold occurrence in genomes: power-law behavior and evolutionary model**. *J Mol Biol* 2001, **313**:673-681.

70. •• Caetano-Anolles G, Caetano-Anolles D: **An evolutionarily structured universe of protein architecture**. *Genome Res* 2003, **13**:1563-1571.
An ambitious attempt to uncover the early evolution of all protein folds and their distribution in the present world.

71. Yang S, Doolittle RF, Bourne P: **Phylogeny determined by protein domain content**. *Proc Natl Acad Sci USA* 2005, **102**:373-378.

72. Hartman H: **The origin of the eukaryotic cell**. *Speculations Sci Technol* 1984, **7**:77-81.

73. Hartman H, Fedorov A: **The origin of the eukaryotic cell: a genomic investigation**. *Proc Natl Acad Sci USA* 2002, **99**:1420-1425.

74. Doolittle RF: **Searching for the common ancestor**. *Res Microbiol* 2000, **151**:85-89.

75. Doolittle RF, York AL: **Bacterial actins? An evolutionary perspective**. *Bioessays* 2002, **24**:293-296.

76. Horiike T, Hamada K, Shinozawa T: **Origin of eukaryotic cell nuclei by symbiosis of Archaea in Bacteria supported by the newly clarified origin of functional genes**. *Genes Genet Syst* 2002, **77**:369-376.

77. Rivera MC, Lake JA: **The ring of life provides evidence for a genome fusion origin of eukaryotes**. *Nature* 2004, **431**:152-155.

78. Kandler O: **Cell wall biochemistry and three-domain concept of life**. *Syst Appl Microbiol* 1994, **16**:501-509.

79. Koga Y, Kyuragi T, Nishihara M, Sone N: **Did archaeal and bacterial cells arise independently from noncellular precursors? A hypothesis stating that the advent of membrane phospholipids with enantiomeric glycerophosphate backbones caused the separation of the two lines of descent**. *J Mol Evol* 1998, **46**:54-63.

80. Woese C: **On the evolution of cells**. *Proc Natl Acad Sci USA* 2002, **99**:8742-8747.

81. • Aravind L, Mazumder R, Vasudeven S, Koonin EV: **Trends in protein evolution inferred from sequence and structure analysis**. *Curr Opin Struct Biol* 2002, **12**:392-399.
Among other things, a scenario is depicted suggesting the early diversification of the Rossmann fold.