The minimal genome concept

Arcady Mushegian

Complete genome sequences are becoming available for a large number of diverse species. Quantification of gene content, of gene family expansion, of orthologous gene conservation, as well as their displacement, are now possible – laying the ground for the estimation of the minimal set of proteins sufficient for cellular life. The consensus of computational results suggests a set close to 300 genes. These predictions will be evaluated by engineering of small bacterial genomes.

Addresses
Akkadix Corporation, 11099 North Torrey Pines Road, Suite 200, La Jolla, California 92037, USA;
e-mail: mushegian@akkadix.com

Current Opinion in Gentics & Development 1999, **9**:709–714

0959-437X/99/$ – see front matter

Abbreviation
COGs clusters of othrologous groups

Introduction

The 'minimal genome' approach seeks to estimate the smallest number of genetic elements sufficient to build a modern-type free-living cellular organism. These approaches are related but not identical to, first, studies of existing species with small numbers of genes (reviewed in [1]), second, reconstruction of ancestral genomes (reviewed elsewhere in this issue), and third, modeling the early history of life using knowledge of prebiotic chemistry [2–4]. I further limit the scope of this review to minimal proteomes, ignoring functionally important RNA molecules as well as the physical layout of regulatory signals, coding and non-coding sequences on chromosomes.

Reconstruction of a minimal protein set uses the knowledge of existing genomes to approximate the essential set of survival genes. Completely sequenced genomes have added momentum to the minimal genome approach by providing lists of proteins that are sufficient to sustain the life of a cell, assuming that recruitment of exogenous proteins is unimportant in the present-day species (but see [5] for a vignette). From these empirical lists of sufficient elements, one may proceed to define shorter lists of core players and pathways, by dry-lab comparative sequence analysis combined with wet-lab genome engineering.

Computational analysis of protein sequence similarities

'Connected components'

Estimation of the minimal set of functions conserved in three superkingdoms of life — bacteria, archaea and eukarya — had been attempted even before the advent of fully sequenced genomes, by exhaustive matching of the SWISSPROT database and determining the 'connected components', that is, protein (super)families clustered by pairwise similarity [6]. Connected components that include proteins from two or more superkingdoms may be universally important. 36 connected components (20 enzymes and 16 ribosomal proteins) included representatives of archaea and at least one other superkingdom [6,7]. A picture of relatively complex metabolism began to emerge in these studies, including proteins involved in genome replication and expression, and, in addition, some enzymes of the glycolytic pathway and amino acid salvage [7]. Given the incompleteness of all genomes at the time, perhaps the most important conclusion from this early work is that the computer methods may be applied in a systematic manner to reconstruct fragments of ancient, as well as minimal, metabolism.

From homologs to orthologs

Homology (i.e. descent from a common ancestor) is the basic concept of any evolutionary analysis. In the case of biological sequences, this concept has been refined by contrasting orthology (i.e. homology between genes in two lineages that were derived by speciation) and paralogy (i.e. homology between functionally related genes that were derived by duplication) [8]. Criteria for distinguishing between orthologous and paralogous sequences in the context of complete genomes have been described [9–11], but a number of factors hamper efforts to resolve these two types of homology [11,12].

The 'minimal gene set for cellular life' has been constructed using the complete protein lists of the first two fully sequenced bacterial genomes, *Haemophilus influenzae* and *Mycoplasma genitalium* [13]. A list of orthologous proteins was compiled, with the expectation that it would predominantly contain proteins integral for cell survival, as both parasitic 'parents' themselves evolved towards shedding auxiliary genes. The large evolutionary distance separating Gram-negative and Gram-positive bacteria was also expected to help to enrich with genes of universal importance. Indeed, 244 detected orthologs [13] contained almost no parasitism-specific proteins (two proteins, thought to be in this category, were eliminated from the list but one of those, an endopeptidase prototyped by *Escherichia coli ygjD* gene product, now appears to be conserved in all completely sequenced genomes).

An important result of the analysis was that the same biochemical function might be performed in two genomes by non-orthologous proteins [13,14••,15]. When the

Table 1

Analysis of some non-orthologous gene displacements between *H. influenzae* and *M. genitalium*.

Missing orthologs in *H. influenzae*/ *M. genitalium* minimal gene set	Reason for absence and proposed solution	Status update, 1999
Translation: aminoacyl-tRNA synthetases		
Pro	Paralogous displacement	
Gly	Paralogous displacement	
Gln	No enzyme for charging tRNA with Gln in Gram-positive bacteria; tRNAGlu is aminated, predicted aminotransferase MG336 proposed	Displacement, but not by MG336. Real aminotransferase complex has been characterized [27], consisting of three subunits. *In M. genitalium*, the corresponding genes are: MG099, MG100, and the carboxy-terminal domain of MG098 (probably a stand-alone ORF)
Replication		
RNAaseH	Thought to be missing in *M. genitalium*, exonuclease MG262 proposed	PSI–BLAST searches (A Mushegion, unpublished observations) identify MG199 as the ortholog of *H. influenzae* RNAaseHII (see also [28])
Nucleotide biosynthesis		
Nucleoside diphosphate kinase	Missing in *M. genitalium*, putative kinase MG268 proposed	Prediction stands, PSI–BLAST links MG268 to (deoxy)nucleotide kinases
Glycolysis		
Phosphoglyceromutase	Cofactor-dependent enzyme in *H. influenzae*, unrelated, cofactor-independent enzyme in *M. genitalium*. The latter (MG430) proposed for minimal gene set	Prediction stands. Some genomes code for phosphoglyceromutases of both types, and many code for just one [17]
4 other enzymes	Found in *M. genitalium* but thought to be paralogs based on taxonomic criterion: some *M. genitalium* enzymes were closer to the archaeal homologs, thought to be evolutionary outgroup, than to *H. influenzae* enzymes	Most likely to be orthologs anyway. Taxonomic pattern unexpected in 1996 is typical of most metabolic enzymes in archaea that appear to have bacterial origin [18]
Coenzyme biosynthesis		
Lipoate-protein ligase	Two types of lipoate-protein ligases are prototyped by *E. coli* lplA and lipB enzymes. *H. influenzae* has only lplA ortholog, *M. genitalium* has only lipB (MG270)	Proteins lplA and lipB appear to be distantly related paralogs (A Mushegian, unpublished observations)

*As proposed in [13]. ORF, open reading frame. PSI–BLAST, position-specific iterated basic local alignment search tool.

orthologs found in *H. influenzae* and *M. genitalium* proteomes were superimposed on the metabolic map, several pathways appeared to be present but incomplete. In a number of cases, proteins predicted to have specific biochemical activities could be detected in two 'parent' genomes but these proteins were not orthologous to each other. To supplement these 'missing links', the *M. genitalium* proteins were selected. The resulting set consisted of 256 genes comprising a biologically plausible, if bacterially slanted, metabolism. Nutritional requirements of a minimal genome were quite extensive, including all amino acids, nucleotides, fatty acids and complex coenzymes [13].

Displacement of orthologs

As has been shown in the 'minimal gene set' experiment, proteins with the same activity may not be orthologous. Although the exact number and nature of such non-orthologous displacements between two genomes can now be re-evaluated using more refined methodology (see Table 1), there is increasing evidence that functional equivalence of proteins requires neither sequence similarity nor even common three-dimensional folds [14••,15,16••,17]. Gene displacements involve all classes of enzymes and types of biological processes [15,16••]. They may help to illuminate the evolutionary past of metabolic pathways and are of practical interest, such as for defining species-specific protein targets in pathogenic microorganisms or in agricultural pests [17].

Gene displacements may occur quite frequently. The rate of displacements in the minimal genome is at least 5% ([13]; Table 1). Recent analysis of the citric acid cycle enzymes in bacteria and archaea indicates that, for >25% of relevant *E. coli* enzymes, a displacement or, at least, the absence, can be detected in at least one other species [14••]. Thus, any reconstruction of a minimal proteome that does not consider gene displacements will underestimate the number of essential genes.

In a recent work [9], conservative criteria of orthology have been applied to protein sets encoded by 21 complete microbial genomes, to arrive at a list of 51 genes shared

Table 2

Proteins conserved in bacterial and archaeal genomes.

Gene (Ec)	Function (genomes without the gene)	Functional category
alaS	Ala-tRNA synthetase	Translation
argS	Arg-tRNA synthetase	Translation
ffh	Signal recognition particle GTPase	Protein secretion
ftsY	Signal recognition particle GTPase	Protein secretion
fusA	Translation elongation factor EF-G	Translation
hisS	His-tRNA synthetase	Translation
ileS	Ile-tRNA synthetase	Translation
infB	Protein chain initiation factor 2	Translation
ksgA	Dimethyladenosine transferase	Translation
leuS	Leu-tRNA synthetase	Translation
metG	Met-tRNA synthetase	Translation
nusA	Transcription factor	Transcription
pheS	Phe-tRNA synthetase α chain	Translation
pheT	Phe-tRNA synthetase β chain	Translation
prlA	Preprotein translocase subunit	Protein secretion
pyrH	Uridine 5′-monophosphate kinase	Nucleotide metabolism
recA	Strand exchange protein, ATPase	Repair/recombination
rplA	50S ribosomal subunit protein L1	Translation
rplB	50S ribosomal subunit protein L2	Translation
rplC	50S ribosomal subunit protein L3	Translation
rplD	50S ribosomal subunit protein L4	Translation
rplE	50S ribosomal subunit protein L5	Translation
rplF	50S ribosomal subunit protein L6	Translation
rplJ	50S ribosomal subunit protein L10	Translation
rplK	50S ribosomal subunit protein L11	Translation
rplM	50S ribosomal subunit protein L13	Translation
rplN	50S ribosomal subunit protein L14	Translation
rplO	50S ribosomal subunit protein L15	Translation
rplP	50S ribosomal subunit protein L16	Translation
rplR	50S ribosomal subunit protein L18	Translation
rplV	50S ribosomal subunit protein L22	Translation
rplW	50S ribosomal subunit protein L23	Translation
rplX	50S ribosomal subunit protein L24	Translation
rpoA	RNA polymerase, α-subunit	Transcription
rpoB	RNA polymerase, β-subunit	Transcription
rpoC	RNA polymerase, β′-subunit	Transcription
rpsB	30S ribosomal protein S2	Translation
rpsC	30S ribosomal subunit protein S3	Translation
rpsE	30S ribosomal subunit protein S5	Translation
rpsH	30S ribosomal subunit protein S8	Translation
rpsI	30S ribosomal subunit protein S9	Translation
rpsJ	30S ribosomal subunit protein S10	Translation
rpsK	30S ribosomal subunit protein S11	Translation
rpsM	30S ribosomal subunit protein S13	Translation
rpsQ	30S ribosomal subunit protein S17	Translation
rpsS	30S ribosomal subunit protein S19	Translation
serS	Seryl-tRNA synthetase	Translation
tmk	Thymidylate kinase	Nucleotide metabolism
truA	Pseudouridylate synthase I	Translation
tsf	Elongation factor Ts	Translation
valS	Valyl-tRNA synthetase	Translation
ygjD	Endopeptidase	Protein secretion?
adk	Adenylate kinase (Mth, Mj)	Nucleotide metabolism
cysS	Cysteinyl-tRNA synthetase (Mth, Mj)	Translation
eno	Enolase (Rp)	Energy metabolism (glycolysis)
infA	Initiation factor IF-1 (Tm)	Translation
nusG	Transcription antiterminator (Mg, Mp)	Transcription
pgk	Phosphoglycerate kinase (Rp)	Energy metabolism (glycolysis)
pyrG	CTP synthetase (Mg, Mp)	Nucleotide biosynthesis
ruvB	Holliday junction DNA helicase (Aa)	Repair/recombination
serS	Seryl-tRNA synthetase (Mth, Mj)	Translation
tpiA	Triosephosphate isomerase (Rp)	Energy metabolism (glycolysis)
ycfH	Putative 'PHP family' hydrolase (Aa)	Unknown

Proteins conserved in bacterial and archaeal genomes (data from [9]; B Snel, M Huynen, P Bork, personal communication; with modifications). The species included in the analysis: archaea, *Archaeoglobus fulgidus*, *Methanobacterium thermoautotrophicum* (Mth), *Methanococcus jannaschii* (Mj), *Pyrococcus furiosus*, *Pyrococcus horikoshii*; bacteria, *Aquifex aeolicus* (Aa), *Borrelia burgdorferi*, *Bacillus subtilis*, *Chlamydia psittaci*, *Chlamydia trachomatis*, *Escherichia coli* (Ec), *Haemophilus influenzae*, *Helicobacter pylori* (2 strains), *Mycoplasma genitalium* (Mg), *Mycoplasma penumoniae* (Mp), *Mycobacterium tuberculosis*, *Rickettsia prowazeki* (Rp), *Synechocystis sp.*, *Thermotoga maritima* (Tm), *Treponema pallidum*.

between the *E. coli* genome and all other completely sequenced bacterial and archaeal genomes (Table 2). This list, rich in proteins involved in ribosome formation and protein biosynthesis, is not sufficient to build a coherent cell or even a complete ribosome. In fact, the criteria employed by the authors lead to a bias against orthologs with low pairwise similarity (i.e. fast-evolving and/or short proteins, including some ribosomal proteins [A Mushegian, unpublished observations]). Additional restriction on the list of orthologs was imposed by the inclusion of obligate parasites, with their simplified biochemistry; however, this overly rigorous list expands if displacement of essential genes by non-orthologs is considered. For example, if one- or two-species exceptions are allowed, 11 genes are added to the list, and some of the missing pathways, notably glycolysis and nucleotide salvage, start to take shape (Table 2).

Orthologous groups and phyletic patterns

Differential gene duplication and selective gene loss in distinct evolutionary lineages are factors complicating reconstruction of orthologous relationships between genes [11]. The problem of ortholog detection has been redefined as delineation of orthologous groups — that is, clusters of genes that include orthologs and, additionally, those paralogs that were derived at or after separation of the clades [12]. Important aspects of construction of these clusters of orthologous groups (COGs) are, first, that the elementary unit of a COG is a symmetrical best hit between two proteins in different evolutionary lineages; second, that a COG has to include three or more symmetrical best hits; and third, that lineage-specific compact families of paralogs may be treated as one protein [12]. Notably, instead of the similarity scores, the ranking approach is used for selection of best hits, increasing the sensitivity of homologs' detection when sequence identity is low.

Relevant to the scope of this review is the notion of phyletic patterns — the sets of clades in which COG members are found (http://www.ncbi.nlm.nih.gov/cgi-bin/COG/phytabu). Several distinct clades have been defined within sequenced bacterial genomes, so a COG may either contain bacterial proteins only or include representatives of other superkingdoms of life (archaea and/or eukarya). COGs with the phyletic pattern 'bacteria + archaea + eukarya' contain protein sequences that persevere in evolution and therefore are more likely to be necessary for survival. Using four bacterial clades (γ division and ε division for *Proteobacteria*, Gram-negative bacteria, and blue-green bacteria), one archaeal clade (methanogens) and one eukaryotic clade (fungi), the National Center for Biotechnology Information (NCBI) group constructed 816 COGs, of which 327 contained representatives of all three superkingdoms.

Focusing on groups instead of single orthologs, and on representation of the clades instead of species, one may hope to account for at least some cases of gene loss and displacement. Indeed, many of the *Mycoplasma/Haemophilus*

Table 3

Minimal gene sets deduced by various approaches.

Minimal genome approach [reference]	Estimated number of genes/ proteins	Why this might be an underestimation (overestimation)
Connected components [6,7]	At least 36	Incomplete data for all genomes
Minimal gene set *(Haemophilus/ Mycoplasma)* [13]	256	Extensive elimination of paralogs in small genome of *Mycoplasma*; parasitic lifestyle resulting in multiple auxotrophies in both species; unknown number of gene displacements in addition to several detected cases. (Bacteria-specific solutions for some functions)
Set of universal orthologs (see legend to Table 2)	51	No consideration of non-orthologous gene displacement
Universal phyletic patterns in COGs ('omnipresent COGs') [12]	~320	Superkingdom-specific pathways, such as DNA replication, repair, lipid biosynthesis. (Unknown number of parallel solutions for the same biochemical function)
Random knockouts in *B. subtilis* [26]	300–560	(Extensive paralogy in large bacterial genome)
Precise deletions in *S. cerevisiae* [31]	~1000	(Large number of essential genes involved in eukaryote-specific housekeeping)

displacements shown in Table 1 are evident in the list of 327 'universal' COGs, as pairs of clusters with similar functional annotations (e.g. COGs 0406 and 0696 contain two unrelated phosphoglyceromutases; COGs 0095 and 0031 represent two types of lipoate–protein ligase). It is not known how many of the 327 clusters represent parallel solutions to the same functional requirement. Given that the average number of proteins per COG is 1.2 in the case of small genomes of *Mycoplasmae* [12,13,18], and that the rate of gene displacement may be 5–25% per genome (see above), the number of essential proteins may be estimated at 300–370.

The major, and perhaps obvious, consequence of the phyletic pattern approach is that the resulting gene set is not strongly impacted by the multiple auxotrophies of small parasitic bacteria, unlike the *H. influenzae*/*M. genitalium* minimal gene set. Metabolism reconstructed from the proteins that belong to the omnipresent COGs suggests a species capable of *de novo* biosynthesis of amino acids, nucleotides, complex carbohydrates and some coenzymes, thus depending on a small number of organic precursors in the environment. Additionally, 43 of universal COGs contain conserved proteins for which precise function is not known but the type of biochemical activity can be predicted. Diverse enzymes in this group, such as oxidoreductases, hydrolases, and transferases, may also assist in reducing the nutritional requirements of the minimal genome.

Functions that are missing in minimal genomes are of no less interest than the functions found there. It has been noted that several proteins involved in DNA replication are not orthologous between bacteria and archaea/eukarya. The list includes the catalytic subunit of replicative DNA polymerase, replication initiator ATPase, replicative helicases and other proteins [13,18,19••]. Moreover, very few enzymes of DNA repair are orthologous between the three superkingdoms [20•]. In addition, only a small number of enzymes involved in lipid biosynthesis are orthologous between bacteria/eukarya and archaea. This agrees with the observation that the side chains in archaeal lipids are of isoprenoid origin, in contrast to fatty acid side chains in bacteria/eukarya [21]. Whatever the evolutionary mechanism of these displacements might be, a minimal genome has to be supplemented with several proteins representing necessary functions.

Minimal genome and structural diversity of proteins

Proteins that are distantly related at the sequence level typically have similar three-dimensional folds. Sensitive approaches towards protein fold identification, on the basis of recognition of distant sequence similarities to the proteins with the known structures, have been applied recently to the protein sequence sets encoded by completely sequenced genomes [22,23,24••]. It was found that the overall fold diversity increases logarithmically with the increase in genome size but at a rate higher than what is predicted by purely stochastic sampling from a pool of available folds [24••]. In other words, evolution appears to have selected for the richness of fold repertoire. In this spirit, we compared fold distributions in the complete set of *M. genitalium* proteins and in those proteins that belong to *H. influenzae*/*M. genitalium* minimal gene set. More than 90% of proteins with recognizable folds in *M. genitalium* belong to the minimal set ([24••]; YI Wolf, personal communication). In large part, this is because the conserved, essential proteins are better covered by structural biology approaches than lineage-specific proteins; but this could also be taken as an indication of the limits to which folds can be lost from a small genome, or to the notion of fold diversity threshold as a requirement for supporting cell functions. High diversity of folds in universally conserved ribosomal proteins has been noted [25]. A concerted effort to resolve diverse protein structures, including those of many minimal-genome proteins, has been launched (http://www.structuralgenomics.org), that will help to determine the set of folds necessary for sustaining a genome.

Mutational analysis of bacterial genomes

In a seminal work, Itaya [26] attempted to assess directly the minimal genome size compatible with life by inserting

a selectable marker into near-random locations (e.g. rare restriction sites) in the genome of *Bacillus subtilis*. Among 79 loci tested, only six insertions impaired the ability of *B. subtilis* to form colonies on Luria broth medium. Moreover, simultaneous insertions into 33 *NotI* sites still rendered a viable phenotype. Statistical analysis indicated that the indispensable DNA size is in the range of 318–562 kbp, or, given that the size of an average bacterial open reading frame is close to 1 kbp, of 300–500 genes [26].

A caveat in this analysis is that large microbial genomes may have a relatively high degree of functional redundancy. One way to indirectly estimate redundancy is by assessing the level of paralogy — the fraction of genes that has at least one paralog in the same genome. The paralogy level in *E. coli* and *H. influenzae*, two relatively close species (both in the γ subdivision of Proteobacteria) has been calculated as ~50% and ~33%, respectively, whereas the gene number in *E. coli* is ~2.5 times higher [10]. The ratio of gene numbers is 1.4:1 in two closely related Mycoplasma species, *M. pneumoniae* and *M. genitalium*, and the ratio of their respective levels of paralogy is 1.15:1 (our calculations based on [1,18]).

If these values are of general significance, then the degree of paralogy in phylogenetically close species appears to change approximately as the square root of the gene number ratio. Thus, the paralogy level in Itaya's projected minimal genome [26] would have to be almost three-fold lower than in *B. subtilis*. Conceivably, certain genes that are dispensable in *B. subtilis* because of functional redundancy would become unique and possibly indispensable in a smaller genome. The extent of this 'loss of backup' upon the reduction of genome size is difficult to estimate. The ongoing experiments to systematically knock-out genes and to profile gene expression in the small genomes of *Mycoplasmae* (CA Hutchison III, R Herrmann, personal communication) will address this and other aspects of minimal genome assessment in a direct way.

Conclusions

Distinguishing between orthologs and paralogs, and detecting displacements of orthologs, have been two important advances in evolutionary reconstructions and in estimating the minimal gene set compatible with cellular life. Requirements of orthologous groups instead of strict orthology, and of occurrence in most clades instead of every species, help to account for many cases of gene loss and displacement. In a sense, this is a return to the concept of connected components, although in a much more rigorous manner. These better-defined and more quantitative approaches could not be developed without the information currently available on completely sequenced genomes. Estimations of the minimal proteome size are summarized in Table 3. The cases of non-orthology of DNA replication/repair and lipid biosynthesis in three superkindoms of life await their (separate) evolutionary explanations.

Notes added in proof

1. In the course of an effort to evaluate some of the universally conserved bacterial proteins as drug targets, endopeptidase YgjD was found to be essential in *E. coli* [29].

2. 'Universal protein families' shared by all three domains of life have been described [30•]. Although the emphasis of that work was on the ancestral, rather then minimal, genome, the set of 324 proteins with 246 known biochemical functions is very close to the set of 327 universal COGs (see above).

3. The COG server (http://www.ncbi.nlm.nih.gov/COG) currently presents COGs from 21 species, including four archaea, one eukaryote and 16 bacteria. Notably, despite a large number of the new COGs, there are now only 322 COGs with phyletic pattern 'bacteria + archaea + eukarya', apparently because some previously recognized COGs could be linked together.

4. Recent analysis of a representative collection of yeast null mutants [31] indicated that 17% of them were nonviable, giving the estimate of ~1000 essential genes in the yeast genome. This number is higher than in the case of *B. subtilis* [26]. Although difference in methodologies may have contributed to an underestimation in [26], a more important reason for the difference seems to be the emergence of eukaryote-specific genes involved in sustaining the nucleus, organelles, and cytoplasmic structures, as well as in the cell cycle. Indeed, the ratio of essential to nonessential ORFs is higher for these categories of yeast genes than for those coding for metabolic enzymes [31]. The truly lower bound of the number of essential genes will be defined by engineering small bacterial genomes.

5. Computational definition of shared elements, as well as direct genome engineering, are also applicable to analysis of essential regions on DNA and RNA that do not code for proteins [32,33]. A consequence of the minimal genome composition is that the approximate size of the minimal cell is estimated to be at least 0.14 micrometers, which proved relevant in a recent discussion of 'nanobacteria' [34].

Acknowledgements

I thank Eugene Koonin for mentoring, many stimulating discussions on a variety of topics over the years, and all sorts of support; Jerry Feitelson for critically reading the manuscript; Peer Bork, Patrick Forterre, Richard Herrmann, Clyde Hutchison III, Martin Huynen, Berend Snel and Yuri Wolf for communicating their unpublished results.

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. Herrmann R, Reiner B: ***Mycoplasma pneumoniae* and *Mycoplasma genitalium*: a comparison of two closely related bacterial species.** *Curr Opin Microbiol* 1998, **1**:572-579.

2. Levy M, Miller SL: **The stability of the RNA bases: implications for the origin of life.** *Proc Natl Acad Sci USA* 1998, **95**:7933-7938.

3. Levy M, Miller SL: **The prebiotic synthesis of modified purines and their potential role in the RNA world.** *J Mol Evol* 1999, **48**:631-637.

4. Orgel LE: **The origin of life – a review of facts and speculations.** *Trends Biochem Sci* 1998, **23**:491-495.

5. Heinemann JA: **Genetic evidence of protein transfer during bacterial conjugation.** *Plasmid* 1999, **41**:240-247.

6. Gonnet GH, Cohen MA, Benner SA: **Exhaustive matching of the entire protein sequence database.** *Science* 1992, **256**:1443 1445.

7. Benner SA, Cohen MA, Gonnet GH, Berkowitz DB, Johnsson KP: **Reading the palimpsest: contemporary biochemical data and the RNA world.** In *The RNA World*. Edited by Gesteland RF, Atkins JW. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1993:27-70.

8. Fitch WM: **Distinguishing homologous from analogous proteins.** *Syst Zool* 1970, **19**: 99-113.

9. Huynen M, Bork P: **Measuring genome evolution.** *Proc Natl Acad Sci USA* 1998, **95**:5849-5856.

10. Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, Rudd KE, Koonin EV: **Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*.** *Curr Biol* 1996, **6**:279-291.

11. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y: **Predicting function: from genes to genomes and back.** *J Mol Biol* 1998, **283**:707-725.

12. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278**:631-637.

13. 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.

14. •• Huynen MA, Dandekar T, Bork P: **Variation and evolution of the citric acid cycle: a genomic perspective.** *Trends Microbiol* 1999, **7**:281-291.
In addition to the exploration of citric acid cycle evolution, classification of gene displacement is presented, as well as an interesting attempt to visualize evolution of the pathway.

15. Koonin EV, Mushegian AR, Bork P: **Non-orthologous gene displacement.** *Trends Genet* 1996, **12**:334-336.

16. •• Galperin MY, Walker DR, Koonin EV: **Analogous enzymes: independent inventions in enzyme evolution.** *Genome Res* 1998, **8**:779-790.
A comprehensive collection of functions performed by unrelated enzymes set against the backdrop of genomes, biochemistry, and protein three-dimensional folds. Oxidoreductases appear to be especially prone to displacements, as well as various enzymes of carbohydrate metabolism.

17. Galperin MY, Bairoch A, Koonin EV: **A superfamily of metalloenzymes unifies phosphopentomutase and cofactor-independent phosphoglycerate mutase with alkaline phosphatases and sulfatases.** *Protein Sci* 1998, **7**:1829-1835.

18. Koonin EV, Mushegian AR, Galperin MY, Walker DR: **Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea.** *Mol Microbiol* 1997, **25**:619-637.

19. •• Forterre P: **Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins.** *Mol Microbiol* 1999, **33**:457-465.
The notion of non-orthologous gene displacement used to explore early evolution of DNA replication. A scenario, in which the last common ancestor may have had a DNA genome, with the evidence thereof obscured by gene displacements. See [35••] for an alternative scenario.

20. • Aravind L, Walker DR, Koonin EV: **Conserved domains in DNA repair proteins and evolution of repair systems.** *Nucleic Acids Res* 1999, **27**:1223-1242.
A gallery of previously unrecognized domains in what thought to be a well-studied set of proteins. Minimal requirements for a functioning DNA repair system are emerging from this study.

21. Kates M: **Archaebacterial lipids: structure, biosynthesis and function.** *Biochem Soc Symp* 1992, **58**:51-72.

22. Huynen M, Doerks T, Eisenhaber F, Orengo C, Sunyaev S, Yuan Y, Bork P: **Homology-based fold predictions for *Mycoplasma genitalium* proteins.** *J Mol Biol* 1998, **280**:323-326.

23. Rychlewski L, Zhang B, Godzik A: **Fold and function predictions for *Mycoplasma genitalium* proteins.** *Fold Des* 1998, **3**:229-238.

24. •• Wolf YI, Brenner SE, Bash PA, Koonin EV: **Distribution of protein folds in the three superkingdoms of life.** *Genome Res* 1999, **9**:17-26.
One of the most sensitive procedures for fold recognition and of the most comprehensive surveys of recognized folds thus far, with such interesting observations as convergent elimination of similar subsets of folds in diverse bacterial pathogens of humans.

25. Ramakrishnan V, White SW: **Ribosomal protein structures: insights into the architecture, machinery and evolution of the ribosome.** *Trends Biochem Sci* 1998, **23**:208-212.

26. Itaya M: **An estimation of minimal genome size required for life.** *FEBS Lett* 1995, **362**:257-260.

27. Curnow AW, Hong KW, Yuan R, Kim SI, Martins O, Winkler W, Henkin TM, Soll D: **Glu-tRNAGln amidotransferase: a novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation.** *Proc Natl Acad Sci USA* 1997, **94**:11819-11826.

28. Bellgard MI, Gojobori T: **Identification of a ribonuclease H gene in both *Mycoplasma genitalium* and *Mycoplasma pneumoniae* by a new method for exhaustive identification of ORFs in the complete genome sequences.** *FEBS Lett* 1999, **445**:6-8.

29. Arigoni F, Talabot F, Peitsch M, Edgerton MD, Meldrum E, Allet E, Fish R, Jamotte T, Curchod ML, Loferer H: **A genome-based approach for the identification of essential bacterial genes.** *Nat Biotechnol* 1998, **16**:851-856.

30. • Kyrpides N, Overbeek R, Ouzounis C: **Universal protein families and the functional content of the last common ancestor.** *J Mol Evol* 1999, **49**:413-423.
Reconstruction of the gene content of the ancestral genome, with the obligate for these authors criticism of [13,18]. Characteristically, several crucial points in the latter papers are misrepresented, such as the statement of the difference between minimal and ancestral genomes in [13] and the fraction of proteins in *M. jannaschii* with predicted functions in [18].

31. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H *et al.*: **Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis.** *Science* 1999, **285**:901-906.

32. Itaya M, Tanaka T: **Fate of unstable Bacillus subtilis subgenome: re-integration and amplification in the main genome.** *FEBS Lett* 1999, **448**:235-238.

33. Washio T, Sasayama J, Tomita M: **Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination.** *Nucleic Acids Res* 1998, **26**:5456-5463.

34. Maniloff J: **Nanobacteria: size limits and evidence.** *Science* 1997, **276**:1776.

35. •• Leipe DD, Aravind L, Koonin EV: **Survey and summary: did DNA replication evolve twice independently?** *Nucleic Acids Res* 1999, **27**:3389-3401.
A scenario alternative to [19••], intriguingly suggesting an ancestral genome which resembles the present-day retroid viruses in that it underwent direct and reverse transcription in the course of normal replicative cycle. This work depicts yet another collection of previously unrecognized conserved domains in the major housekeeping proteins. Finally, a timely review of alternative hypotheses on the nature of the ancestral genomes is presented.