Evolution of reduced prokaryotic genomes and the minimal cell concept: Variations on a theme

Luis Delaye,[1,2] *and Andrés Moya*[1,3]*

[1]Unidad Mixta de Investigación en Genómica y Salud CSISP-UVEG/Instituto Cavanilles, València, Spain
[2]Facultad de Ciencias, Universidad Nacional Autónoma de México, México
[3]CIBER en Epidemiología y Salud Pública (CIBEResp), Spain

Prokaryotic genomes of endosymbionts and parasites are examples of naturally evolved minimal cells, the study of which can shed light on life in its minimum form. Their diverse biology, their lack of a large set of orthologous genes and the existence of essential linage (and environmentally) specific genes all illustrate the diversity of genes building up naturally evolved minimal cells. This conclusion is reinforced by the fact that sometimes the same essential function is performed by genes from different evolutionary origins. Nevertheless, all cells perform a set of life-essential functions however different their cell machinery and environment in which they thrive. An upcoming challenge for biologists will be to discern, by studying differences and similarities in current biodiversity, how cells with reduced genomes have adapted while retaining all basic life-supporting functions.

Keywords: comparative genomics; endosymbiosis; genome reduction; minimal cell; paleome

Introduction

What is the minimum number of functions necessary and sufficient for life? To address this question two different, but complementary, approaches have been followed.[1] On the one hand, bottom-up approaches, which are usually contextualized under the origin of life problem, aim to construct living systems from relatively simple molecular precursors; on the other hand, top-down efforts aim to simplify extant cells to its minimal expression.

The minimal cell concept states that for a particular kind of cell in a defined (and favorable) environment, there is a minimum number of features or functions necessary to keep the cell alive.[2] However, a minimal cell is only meaningful in relation to a particular environment (and of course, to the kind of cell under study), thus a plethora of minimal cells may exist.[3]

***Correspondence to:** A. Moya, Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Universitat de València, Apartado Postal 22085, 46071 València, Spain.
E-mail: andres.moya@uv.es

From the top-down viewpoint, the search for a minimal cell usually involves the engineering of a reduced genome through successive rounds of gene deletion.[4] Genes to be deleted can be identified through comparative genome analysis. Alternatively, a minimal gene set can be proposed by singly inactivating all genes in the genome and identifying those essential for cell survival. Here we state that a third source of information on minimal living systems can be exploited. Genomes from a variety of prokaryotes, whose biology encompasses endosymbionts as well as parasites, are the product of intensive natural genome reduction. These genomes, although not minimal (in that the absence of any one of their genes would render them lifeless), can provide clues to the set of essential genes a hypothetical minimal cell should have.[5,6] The set of essential functions thus identified should then be taken into account for synthetic biology purposes.

Naturally evolved reduced genomes

The smallest sequenced genomes from free-living prokaryotes include an uncultured ocean β-proteobacterium from the clade OM43, strain HTCC2181 with 1,377 genes,[7] followed closely by the cosmopolitan oceanic bacterium *Candidatus* Pelagibacter ubique HTCC1062 with 1,394 genes,[8] the dehalorespirant *Dehalococcoides* sp. BAV1 with 1,436 genes[9] and the hyperthermophilic crenarchaeon *Ignicoccus hospitalis* KIN4/I, with 1,494 genes.[10] It is noteworthy that all these different organisms, belonging to very different clades from the tree of life, and living in clearly different environments, have evolved toward genomes with a similar number of genes. Taking these organisms as hallmarks, it has been proposed that the minimum number of genes for a free-living prokaryote should be ~1,400.[10,11]

Genome sizes for host-associated prokaryotes are much smaller. Figure 1 shows the G + C content as a function of number of protein-coding genes for those genomes sequenced from parasites and endosymbionts with fewer than 1,400 protein-coding genes. As shown, there is a trend toward smaller G + C content associated to genome reduction. Nevertheless, the genome of *Candidatus Hodgkinia*

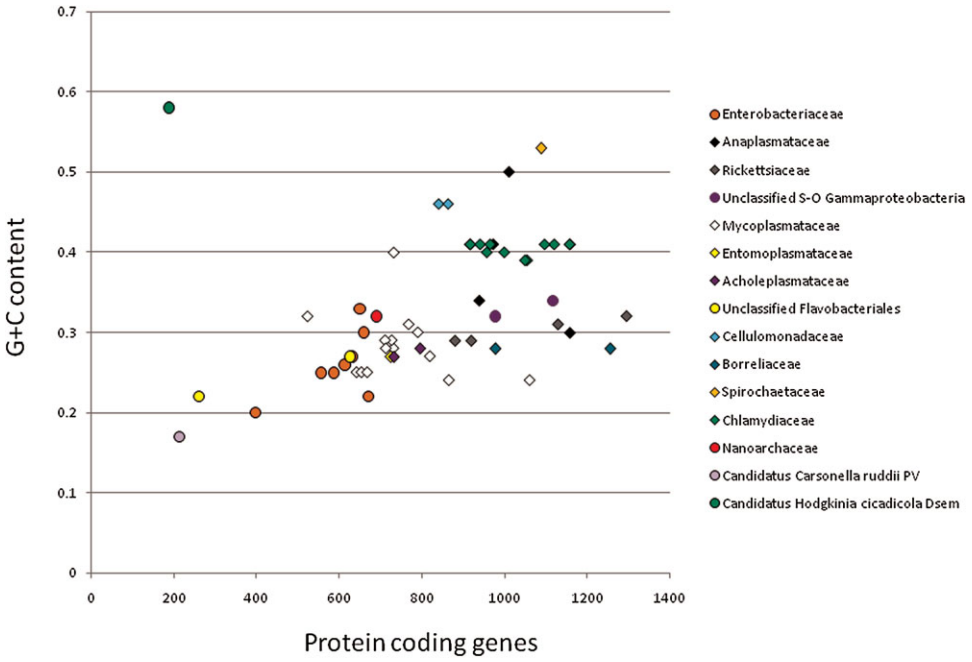

Figure 1. G + C content as a function of protein-coding genes in diverse prokaryotic genomes sequenced from endosymbionts (circles) and parasites (squares). Prokaryotic genomes are grouped at the family level.

cicadicola Dsem shows that low G + C content is not a necessary property of minimal cells.[12]

The smallest genomes are the outcome of endosymbiosis.[13] A process of metabolic complementation may be one of the reasons behind such small genomes. One such case is exemplified by the endosymbionts of the aphid *Cinaria cedri* where the two resident bacteria are indispensable to supply tryptophan to their host.[14] The γ-proteobacteria *Buchnera aphidicola* BCc contains a plasmid with *trpEG* genes to biosynthesize tryptophan, while *Candidatus* Serratia symbiotica codes for the rest of the genes of the pathway on its chromosome (*trpDCBA*). The split of the tryptophan operon into two different cells must have important consequences for its regulation by attenuation. Therefore, it should be explored if this situation is evolutionary stable in the sense that it provides some particular advantage to the host (*i.e.*, an increased tryptophan provision). A similar situation is found in the symbiotic systems of the xylem-feeding glassy-winged sharpshooter (*Homalodisca coagulata*)[15] and in the cicada *Diceroprocta semicincta*.[16] Both insects have a bacteroidetes endosymbiont (*Sulcia muelleri*), and in both cases this endosymbiont provides metabolic routes for eight out of ten amino acids (arginine, phenylalanine, tryptophan, lysine, threonine, isoleucine, leucine, and valine) that cannot be synthesized by the host. The remaining two amino acids (methionine and histidine) are provided by another endosymbiont. These are a γ-proteobacterium (*Baumannia cicadellinicola*) in the case of the sharpshooter and an α-proteobacterium (*H. cicadicola*) in the case of the cicada. Metabolic integration through differential gene loss seems to be an important process in the evolution of endosymbiotic systems in insects. It remains to be seen if metabolic compartmentalization provides by itself evolutionary advantages to the hosts, or if it represents only transitional steps toward the replacement of one endosymbiont by another. Nevertheless, the synthetic biology community should learn from these systems the advantages of engineering systems with more than one kind of cell.

What is it called: Endosymbiont or organelle?

At what point should we stop calling a cell an endosymbiont and start calling it an organelle? One criterion proposed to distinguish a cell form an organelle is whether or not some active proteins in the cytosol of the entity under discussion are coded in the genome of the host.[17] For instance, this criterion was used when comparing the genomes of *Haemophilus influenza* and *Mycoplasma genitalium* to derive a theoretical minimal gene set to sustain life: ''*however small, a cellular gene set has to be self-sufficient in the sense that cells generally import metabolites and not functional proteins*''.[18]

It should be noted that the above definition does not contemplate the evolutionary origin of the genes coding for

proteins targeted from the nucleus to the organelle (*i.e.*, if the genes in question have a different evolutionary origin other than the ancestral endosymbiont that evolved into an organelle). Nevertheless, it provides a useful criterion related to the level of genetic integration between a host and its guest.

Accordingly, the lack of several important genes involved in DNA replication, transcription and translation in *Candidatus Carsonella ruddii* (considered the primary endosymbiont of the psyllid *Pachpsylla venusta*) has led to the proposal that this entity has lost the status of cell to become an organelle.[19] The same could apply to other prokaryotes with extremely reduced genomes, like *H. cicadicola* and *S. muelleri*, endosymbionts of the cicada *D. semicincta*.

On similar grounds but in the opposite direction, the organelle/cell status of the chromatophore of *Paulinella chromatophora* has been recently discussed. The genome of the resident entity is much larger than those of the aforementioned dwellers of insects, and it has been indicated that it cannot be considered an organelle unless there is evidence of targeted proteins from the host to the chromatophore.[20] However, it has been argued that the above definition does not take into account the gradation that may exists in becoming an organelle, and that the chromatophore is so well integrated to the host physiology that is deserves (at least functionally) the status of organelle.[21] If some of the proteins in the cytosol of endosymbionts are effectively coded in the host genome, caution should be taken while studying these systems as examples of minimal cells.

Decontructing cells, a case study of comparative and experimental biology

M. genitalium, *B. aphidicola*, and *Nanoarchaeum equitans* diverged over several thousand million years (the divergence is particularly large between the two bacteria and the archaea). They have clearly different lifestyles and have some of the smallest genomes known for a parasite, an endosymbiont, and an episymbiont, respectively.[22–24] A key question is whether we can gain new insights into the set of life-essential functions by comparing their genomes and their biology. In principle, shared features of their biology (whether performed by homologous genes or not) should help us identify essential functions.

Before proceeding, it should be noted that none of these organisms are capable of surviving without their hosts. Therefore, the set of genes and functions shared by the three of them will lack several important genes necessary for a minimal free-living lifestyle. It should also be considered that a minimal set of genes and associated functions necessary to sustain life is a different concept than an ancestral gene set.[11] Despite having small genomes, these organisms do not resemble ancient stages of cellular evolution. These prokaryotes (although simple) are the outcome of evolutionary refinements leading to highly specialized adaptations. Moreover, at least in the case of *M. genitalium* and *B. aphidicola*, they have clearly evolved from prokaryotes with larger and more complex genomes.[25,26] Taking into account previous considerations, homologous traits shared by the three of them (setting aside hypothetical cases of horizontal gene transfer), must have evolved very early in the history of life on Earth and certainly belongs to the essential set of genes and/or functions common to all cells.

As shown in Fig. 2, gene comparison reveals that there are only 59 families of genes where each one of them is the best reciprocal hit to each other among the three genomes (49 of them are strict best reciprocal hits, while 10 families follow a slightly different schema where two of the proteins are the best reciprocal hit to the same third one). Most of these genes participate directly in translation (ribosomal proteins and aminoacyl-tRNA synthetases; Table 1). This result is in agreement with other comparative analyses.[27] It is well known that the so-called informational genes (those involved in transcription and translational cell machinery) tend to be conserved across largest phylogenetic distances.[27] A cautionary note concerning the kind of homology relationship among these genes is that despite best reciprocal hits to each other, the genes inside each one of these families are not necessarily orthologous. This is because differential secondary gene loss of paralogous genes is a possibility that cannot be ruled out. In fact, proteins belonging to some of the families shown in Table 1 have members with different domain architectures, indicating that substantial evolutionary change

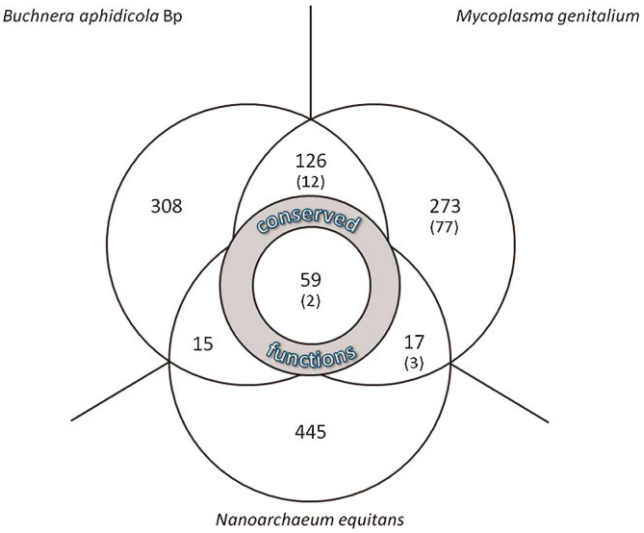

Figure 2. Phylogenetic distribution of putative orthologous genes identified as best reciprocal hits between *M. genitalium*, *B. aphidicola* Bp, and *N. equitans*. In parenthesis the number of essential genes when singly deleted in the genome of *M. genitalium*.

Table 1. Conserved genes among *M. genitalium*, *B. aphidicola* Bp, and *N. equitans* identified as reciprocal best hits. The product name follows the annotation of *B. aphidicola*; (*) Protein products annotated with a name other than that of *B. aphidicola* Bp; proteins differing in domain structure inside each family are shown in bold; nonessential genes when singly deleted in the genome of *M. genitalium* are shown in grey boxes

Function	*B. aphidicola*	*M. genitalium*	*N. equitans*	Product name
DNA replication	bbp425	MG_419	**NEQ170***	*dnaX*; DNA polymerase III subunit gamma
Transcription	bbp034	MG_340	**NEQ503**	*rpoC*; DNA-directed RNA polymerase
	bbp035	MG_341	**NEQ156**	*rpoB*; DNA-directed RNA polymerase
tRNA processing	bbp185	MG_182	**NEQ333**	*truA*; tRNA pseudouridine synthase A
Ribosomal proteins	bbp213	MG_070	NEQ508	*rpsB*; 30S ribosomal protein S2
	bbp461	MG_157	**NEQ481**	*rpsC*; 30S ribosomal protein S3
	bbp443	MG_311	**NEQ247**	*rpsD*; 30S ribosomal protein S4
	bbp450	MG_168	NEQ388	*rpsE*; 30S ribosomal protein S5
	bbp471	MG_088	NEQ242	*rpsG*; 30S ribosomal protein S7
	bbp453	MG_165	NEQ274	*rpsH*; 30S ribosomal protein S8
	bbp353	MG_417	NEQ446	*rpsI*; ribosomal protein S9
	bbp468	MG_150	NEQ083	*rpsJ*, nusE; 30S ribosomal protein S10
	bbp444	MG_176	NEQ069	*rpsK*; 30S ribosomal protein S11
	bbp472	MG_087	NEQ058	*rpsL*; 30S ribosomal protein S12
	bbp445	MG_175	NEQ467	*rpsM*; 30S ribosomal protein S13
	bbp458	MG_160	NEQ326	*rpsQ*; 30S ribosomal protein S17
	bbp463	MG_155	NEQ480	*rpsS*; 30S ribosomal protein S19
	bbp038	MG_082	NEQ546	*rplA*; 50S ribosomal protein L1
	bbp464	MG_154	NEQ361	*rplB*; 50S ribosomal protein L2
	bbp467	MG_151	NEQ433	*rplC*; 50S ribosomal protein L3
	bbp455	MG_163	NEQ093	*rplE*; 50S ribosomal protein L5
	bbp039	MG_081	NEQ101	*rplK*; 50S ribosomal protein L11
	bbp354	MG_418	NEQ207	*rplM*; 50S ribosomal protein L13
	bbp457	MG_161	NEQ092	*rplN*; 50S ribosomal protein L14
AA-tRNA biosynthesis	bbp103	MG_021	**NEQ457**	*metG*; methionyl-tRNA synthetase
	bbp119	**MG_375**	**NEQ177**	*thrS*; threonyl-tRNA synthetase
	bbp123	MG_194	NEQ505	*pheS*; phenylalanyl-tRNA synthetase α-chain
	bbp124	MG_195	**NEQ479**	*pheT*; phenylalanyl-tRNA synthetase β-chain
	bbp139	MG_345	**NEQ230**	*ileS*; isoleucyl-tRNA synthetase
	bbp224	MG_378	NEQ208	*argS*; arginyl-tRNA synthetase
	bbp267	**MG_035**	NEQ102	*hisS*; histidyl-tRNA synthetase
	bbp290	MG_005	NEQ308	*serS*; seryl-tRNA synthetase
	bbp329	MG_113	NEQ535	*asnC*; asparaginyl-tRNA synthetase
	bbp331	MG_334	**NEQ252**	*valS*; valyl-tRNA synthetase
	bbp431	MG_253	**NEQ055**	*cysS*; cysteinyl-tRNA synthetase
	bbp478	MG_126	NEQ115	*trpS*; tryptophanyl-tRNA synthetase
	bbp364	**MG_292**	**NEQ211**	*alaS*; alanyl-tRNA synthetase
	bbp221	MG_283	NEQ210	prolyl-tRNA synthetase
	bbp395	**MG_266**	NEQ239	*leuS*; leucyl-tRNA synthetase
Translation factors	bbp469	MG_451	NEQ082	*tuf*; elongation factor Tu
	bbp470	MG_089	NEQ543	*fusA*; elongation factor G
	bbp340	**MG_142**	**NEQ498**	*infB*; translation initiation factor IF-2
	bbp180	MG_024	**NEQ463**	*ychF*; GTP-dependent nucleic acid-binding
Translation associated	bbp212	MG_172	NEQ399	*map*; methionine aminopeptidase
Metabolism	bbp291	MG_102	NEQ491	*trxB*; thioredoxin reductase
Transport	bbp295	MG_526	NEQ421*	*znuC*; high-affinity zinc uptake system
	bbp394	**MG_146***	**NEQ189***	*corC*; magnesium and cobalt efflux protein
Cell division	bbp194	**MG_224**	NEQ133	*ftsZ*; cell division protein FtsZ
	bbp345	MG_457	**NEQ186***	*hflB*; cell division protein FtsH
Protein degradation	bbp421	MG_239	**NEQ349**	*lon*; ATP-dependent protease La
	bbp055	MG_046	NEQ493	*gcp*; putative endopeptidase
	bbp332	MG_391*	NEQ412*	*pepA*; cytosol aminopeptidase
Genetic information processing	bbp325	MG_009	NEQ456	*ycfH*; putative deoxyribonuclease
Oxidative phosphorylation	bbp008	MG_399	**NEQ263***	*atpD*; F0F1 ATP synthase subunit beta
Protein folding	bbp021	MG_392	NEQ141*	*groEL*; chaperonin GroEL
Other	**bbp352**	**MG_384**	**NEQ157***	*obgE*; GTPase ObgE
	bbp327	MG_132	NEQ519	*ycfF*; HIT-like protein
	bbp162	MG_259	NEQ238	*hemK*; hypothetical protein
	bbp082	MG_351	NEQ461	*ppa*; inorganic pyrophosphatase

have occurred since their divergence, thus compromising simple assignment of function by homology.

Notably, metabolism is represented by just one single enzyme (thioredoxin reductase, *trxB*). The lack of more conserved metabolic proteins could be the outcome of two non-exclusive phenomena. On one hand, much of prokaryotic adaptation and evolutionary innovation seems to happen at the metabolic level (*i.e.*, prokaryotes are biochemically the most diverse set of organisms); and on the other, these three organisms may have lost several pathways in their evolutionary path toward their host-associated lifestyle.[28]

In contrast to metabolism, protein degradation is represented by three gene families. These are annotated as: (i) ATP-dependent Lon protease involved in degradation of short-lived regulatory and abnormal proteins in *Escherichia coli*; (ii) a probable *O*-sialoglycoprotein endopeptidase; and (iii) a cytosol (leucyl) aminopeptidase, presumably involved in the processing and regular turnover of intracellular proteins. Although the annotation of some of these proteins should be confirmed by experimental approaches (like the case of the probable *O*-sialoglycoprotein endopeptidase), degradation is clearly a conserved function among these three prokaryotes. The reason why these proteins involved in degradation are conserved in these diverse organisms is not clear. Nevertheless, it has been suggested that by selectively degrading components of the cell, proteins involved in degradation make the necessary room for evolutionary novelties.[29] The role of degradation process in the cell, as an essential function, clearly deserves more attention.

The next question we should ask ourselves is why these particular genes, rather than others, have remained relatively unchanged with respect to other gene families, despite billions of years of divergence. Evolution by natural selection indicates that there will be differential reproduction and survival of variants following adaptation to their environment. Accordingly, changes in cell transcription and translation machinery have clearly been penalized by natural selection. One possible explanation for this pattern could be that the high level of integration and complexity on these subcellular systems has made them less evolvable, the so-called complexity hypothesis.[30] Whatever the reason, it seems that natural selection has favored relative stasis in the so-called informational machinery of the cell, while enhancing innovation and adaptive change in other parts of the genome.

These two components of the genome are referred to as the paleome and the cenome, respectively.[31] As shown in Fig. 2, the intersection of the three genomes clearly corresponds to the paleome, while a certain proportion of the unshared genes belongs to the cenome. Of course the ability (the resolution) to accurately detect genes belonging to the cenome is related to the phylogenetic sample (the denser the better) and on the level of phylogenetic relatedness of the genomes in question.

By looking at the set of experimentally determined essential genes of *M. genitalium*,[32] the smallest proportion of essential genes corresponds to those conserved among two or three genomes (Fig. 2). It is known that essential genes tend to be phylogenetically conserved;[33] nevertheless, deviating from this pattern there are two conserved genes in the three genomes that prove nonessential when singly deleted (Table 1). One of them, *ycfH* is a putative deoxyribonuclease from the HIT superfamily, whose homolog is also known to be nonessential when singly deleted in *E. coli* (or when both of them are deleted).[34] This gene together with the other nonessential conserved gene *ychF*, a GTP-dependent nucleic acid-binding protein that might be involved in the translation,[35] shows that the apparent lack of essentiality, according to single-gene deletion criteria, is not enough to exclude a gene from a functional minimal gene set.[36] These two genes may not prove essential when singly deleted under laboratory conditions, but their absence within their natural environment would seriously compromise cell fitness. This example illustrates how in order to better define a set of essential genes, single-gene deletion experiments should be accompanied by comparative genome analysis to detect other important cell components. Accordingly, once several different species have been screened for essential genes, and conserved proteins have been identified through comparative genome analysis, comparative biology of essential functions in these different cells could then shed light on the set of essential functions. This would, in turn, make it possible to draw up a catalog of essential subcellular systems.

This simple analysis might underestimate the number of homologous genes among the three cells (*e.g.*, highly divergent homologous genes can go undetected by sequence similarity searches at the level of primary structure; or other homologs, which differ from the best reciprocal hit to a gene, are not identified). However, the low number of putative orthologous genes among the three genomes is remarkable given that the three cells perform all the basic functions required for a simple bacterial cell[33] [*i.e.*, biosynthesis of a membrane enclosing the elements necessary to synthesize proteins that carry out reactions required for: (i) the duplication and inheritance of DNA-based genetic information, (ii) the division of the compartment, and (iii) the provision of energy]. The lack of many orthologous proteins among these three reduced genomes is an indication that all the basic life-giving functions can be performed by different non-homologous genes.

Beyond homologous genes, the search for conserved functions

As mentioned above, genes conserved in *M. genitalium*, *B. aphidicola*, and *N. equitans* very likely belong to the ancient

set of genes inherited from the last common ancestor (LCA) of all life, and are certainly important for cells in general. However, there may be other levels (other than gene comparison) at which we could compare cells to gather clues about the minimal features necessary for life. One such level of comparison is to look for the same biochemical reaction (or a more complex cellular process) coded by unrelated enzymes (or set of enzymes). One such process is DNA replication, where the central replicative enzyme is not universally conserved (together with other important components of replication machinery). Bacteria use DNA polymerase from family C, while eukarya and archaea use B family DNA polymerases.[37] Despite the lack of conservation of the replicase, DNA polymerization is clearly a central function required for present day life. Despite controversies regarding the chemical nature of the genome of the LCA,[38] the similarities of DNA replication among cells and the homology between certain components of the DNA replication machinery[39] suggest that DNA-based genomes may have evolved prior to the existence of the LCA.

A similar case concerns membrane phospholipids.[40] Archaea possess phospholipids that generally comprise isoprenoid ethers built on sn-glycerol-1-phosphate (G1P), while bacterial and eukaryal membrane phospholipids are fatty acid esters linked to sn-glycerol-3-phosphate (G3P). The two key dehydrogenase enzymes that produce G1P and G3P, G1PDH and G3PDH, respectively, are not homologous.

These are two well-known examples of important cellular functions that are not universally conserved. The lack of homology among the enzymes participating in these functions can sometimes be explained by cases of non-orthologous gene displacement, or alternatively by convergence; to choose between the two scenarios, parsimony criteria should be implemented. Regarding minimal cell reconstruction, the systematic analysis of shared functions that are conserved across wide phylogenetic distances (performed by unrelated enzymes) can also shed light on the set of essential functions. In Fig. 2, these functions are indicated by the grey circle surrounding the set of homologous genes to the three genomes.

Search for common principles of life or the blueprint and common ancestry

It has been suggested that behind the concept of the minimal cell lies the idea that there is a set of basic principles common to all life. This approach is similar, in spirit, to physical sciences where the search for universal and unifying principles of natural phenomena is one of the main goals. However, when looking for principles common to life, it should be remembered that living beings are related by common ancestry.[41] This makes it difficult to distinguish between those features of a cell that correspond to such principles (if they exist), and those features that evolved contingently and are present in all cells due to necessity and/or common ancestry. If historical contingence and common ancestry is so important to our understanding of cells, then the search for those principles common to all life must necessarily consider these questions.[42]

Conclusions and prospects

In 1818 Mary Shelley wrote the novel Frankenstein,[43] in which she considered the possibility of creating life by delivering an electric shock to a body assembled from human parts of different origin. Implicit in her novel is the suggestion that life requires properly organized parts sparked by some form of energy. Top-down approaches to minimal cells aim to identify those components essential for life. It remains to be seen whether it is possible to join those systems in a single entity and trigger it into life by using some free-energy form. There is an unbroken chain reaching from present-day cells back to the first living systems. If we ever aim to engineer truly minimal living systems, it is crucial to understand the historical course of events run by molecules performing purely chemical processes up to the first living systems. Comparative analyses of sequenced genomes from parasites and endosymbionts show high diversity despite being small. Underlying this diversity, there is the lack of a large set of orthologous genes among these reduced genomes. However, similarities at the molecular level, like DNA-based genomes or the genetic code, indicate that all life on Earth is related by common ancestry.

Metabolic complementation of endosymbiotic bacteria, such as *B. aphidicola* BCc and *Candidatus* S. symbiotica to provide tryptophan to their host, or the striking case of metabolic convergence from the symbiotic systems of the xylem-feeding glassy-winged sharpshooter and in the cicada *D. semicincta* suggests new avenues of research for synthetic biology, that of engineering microbial systems composed by different kinds of cells.

Next challenge for biologists will be to identify which cell components are necessary and sufficient for life in different, naturally evolved (nearly) minimal cells and discover, by comparative analysis, how the same basic functions for life are performed by a diversity of, sometimes unrelated, genes.

Acknowledgments: Financial support was provided by grants BFU2009-12895-C02-01/BMC (Ministerio de Ciencia e Innovación, Spain), the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement number 212894 and Prometeo/2009/092 (Conselleria d'Educació, Generalitat Valenciana, Spain) to A. M.

References

1. **Ruiz-Mirazo K, Peretó J, Moreno A.** 2004. A universal definition of life: Autonomy and open-ended evolution. *Orig Life Evol Biosph* **34**: 323–46.
2. **Peterson SN, Fraser CM.** 2001. The complexity of simplicity. *Genome Biol* **2**: comment2002.1–2002.8.
3. **Huynen M.** 2000. Constructing a minimal genome. *Trends Genet* **16**: 116.
4. **Pósfai G, Plunkett G III, Fehér T,** *et al.* 2006. Emergent properties of reduced-genome *Escherichia coli*. *Science* **312**: 1044–46.
5. **Gil R, Silva FJ, Peretó J,** *et al.* 2004. Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol Rev* **68**: 518–37.
6. **Klasson L, Andersson SGE.** 2004. Evolution of minimal-gene-sets in host-dependent bacteria. *Trends Microbiol* **12**: 37–43.
7. **Giovannoni SJ, Hayakawa DH, Tripp HJ,** *et al.* 2008. The small genome of an abundant coastal ocean methylotroph. *Environ Microbiol* **10**: 1771–82.
8. **Giovannoni SJ, Tripp HJ, Givan S,** *et al.* 2005. Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**: 1242–5.
9. **He J, Ritalahti KM, Yang KL,** *et al.* 2003. Detoxification of vinyl chloride to ethene coupled to growth of an anaerobic bacterium. *Nature* **424**: 62–5.
10. **Podar M, Anderson I, Makarova KS,** *et al.* 2008. A genomic analysis of the archaeal system *Ignicoccus hospitalis-Nanoarchaeum equitans*. *Genome Biol* **9**: R158.
11. **Islas S, Becerra A, Luisi PL,** *et al.* 2004. Comparative genomics and the gene complement of a minimal cell. *Orig Life Evol Biosph* **34**: 243–56.
12. **McCutcheon JP, McDonald BR, Moran NA.** 2009. Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genet* **5**: e1000565.
13. **Moya A, Gil R, Latorre A,** *et al.* 2009. Toward minimal bacterial cells: Evolution vs. design. *FEMS Microbiol Rev* **33**: 225–35.
14. **Gosalbes MJ, Lamelas A, Moya A,** *et al.* 2008. The striking case of tryptophan provision in the cedar aphid *Cinara cedri*. *J Bacteriol* **190**: 6026–9.
15. **McCutcheon JP, Moran NA.** 2007. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proc Natl Acad Sci USA* **104**: 19392–7.
16. **McCutcheon JP, McDonald BR, Moran NA.** 2009. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proc Natl Acad Sci USA* **106**: 15394–9.
17. **Cavalier-Smith T, Lee JJ.** 1985. Protozoa as hosts for endosymbioses and the conversion of symbionts into organelles. *J Protozool* **32**: 376–9.
18. **Mushegian AR, Koonin EV.** 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci USA* **93**: 10268–73.
19. **Tamames J, Gil R, Latorre A,** *et al.* 2007. The frontier between cell and organelle: Genome analysis of *Candidatus* Carsonella ruddii. *BMC Evol Biol* **7**: 181.
20. **Theissen U, Martin W.** 2006. The difference between organelles and endosymbionts. *Curr Biol* **16**: R1016–7.
21. **Keeling PJ, Archibald JM.** 2008. Organelle evolution: what's in a name? *Curr Biol* **18**: R345–7.
22. **Fraser CM, Gocayne JD, White O,** *et al.* 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**: 397–3.
23. **van Ham RC, Kamerbeek J, Palacios C,** *et al.* 2003. Reductive genome evolution in *Buchnera aphidicola*. *Proc Natl Acad Sci USA* **100**: 581–6.
24. **Randau L, Münch R, Hohn MJ,** *et al.* 2005. *Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5′- and 3′-halves. Nature* **433**: 537–41.
25. **Woese CR, Maniloff J, Zablen LB.** 1980. Phylogenetic analysis of the mycoplasmas. *Proc Natl Acad Sci USA* **77**: 494–8.
26. **Comas I, Moya A, González-Candelas F.** 2007. From phylogenetics to phylogenomics: the evolutionary relationships of insect endosymbiotic gamma-Proteobacteria as a test case. *Syst Biol* **56**: 1–16.
27. **Koonin EV.** 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* **1**: 127–36.
28. **Delmotte F, Rispe C, Schaber J,** *et al.* 2006. Tempo and mode of early gene loss in endosymbiotic bacteria from insects. *BMC Evol Biol* **6**: 56.
29. **Danchin A.** 2009. Myopic selection of novel information drives evolution. *Curr Opin Biotechnol* **20**: 504–8.
30. **Jain R, Rivera MC, Lake JA.** 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA* **96**: 3801–6.
31. **Danchin A.** 2009. Bacteria as computers making computers. *FEMS Microbiol Rev* **33**: 3–26.
32. **Glass JI, Assad-Garcia N, Alperovich N,** *et al.* 2006. Essential genes of a minimal bacterium. *Proc Natl Acad Sci USA* **103**: 425–30.
33. **Kobayashi K, Ehrlich SD, Albertini A,** *et al.* 2003. Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci USA* **100**: 4678–83.
34. **Wexler M, Sargent F, Jack RL,** *et al.* 2000. TatD is a cytoplasmic protein with DNase activity. No requirement for TatD family proteins in sec-independent protein export. *J Biol Chem* **275**: 16717–22.
35. **Teplyakov A, Obmolova G, Chu SY,** *et al.* 2003. Crystal structure of the YchF protein reveals binding sites for GTP and nucleic acid. *J Bacteriol* **185**: 4031–7.
36. **Fang G, Rocha E, Danchin A.** 2005. How essential are nonessential genes? *Mol Biol Evol* **22**: 2147–56.
37. **Filée J, Forterre P, Sen-Lin T,** *et al.* 2002. Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *J Mol Evol* **54**: 763–73.
38. **Leipe DD, Aravind L, Koonin EV.** 1999. Did DNA replication evolve twice independently? *Nucleic Acids Res* **27**: 3389–401.
39. **Harris JK, Kelley ST, Spiegelman GB,** *et al.* 2003. The genetic core of the universal ancestor. *Genome Res* **13**: 407–12.
40. **Peretó J, López-García P, Moreira D.** 2004. Ancestral lipid biosynthesis and early membrane evolution. *Trends Biochem Sci* **29**: 469–77.
41. **Becerra A, Delaye L, Islas S,** *et al.* 2007. The very early stages of biological evolution and the nature of the last common ancestor of the three major cell domains. *Annu Rev Ecol Evol Syst* **38**: 361–79.
42. **Lazcano A.** 2007. What is life? A brief historical overview. *Chem Biodiversity* **4**: 1–15.
43. **Shelley M.** 1818. Frankenstein; or the Modern Prometheus. United Kingdom, Lackington, Hughes, Harding, Mayor, & Jones.