Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 10268–10273, September 1996
Evolution

A minimal gene set for cellular life derived by comparison of complete bacterial genomes

ARCADY R. MUSHEGIAN AND EUGENE V. KOONIN*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

Communicated by Clyde Hutchinson, University of North Carolina, Chapel Hill, NC, May 17, 1996 (received for review March 11, 1996)

ABSTRACT **The recently sequenced genome of the parasitic bacterium *Mycoplasma genitalium* contains only 468 identified protein-coding genes that have been dubbed a minimal gene complement [Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., *et al.* (1995) *Science* 270, 397–403]. Although the *M. genitalium* gene complement is indeed the smallest among known cellular life forms, there is no evidence that it is the minimal self-sufficient gene set. To derive such a set, we compared the 468 predicted *M. genitalium* protein sequences with the 1703 protein sequences encoded by the other completely sequenced small bacterial genome, that of *Haemophilus influenzae*. *M. genitalium* and *H. influenzae* belong to two ancient bacterial lineages, i.e., Gram-positive and Gram-negative bacteria, respectively. Therefore, the genes that are conserved in these two bacteria are almost certainly essential for cellular function. It is this category of genes that is most likely to approximate the minimal gene set. We found that 240 *M. genitalium* genes have orthologs among the genes of *H. influenzae*. This collection of genes falls short of comprising the minimal set as some enzymes responsible for intermediate steps in essential pathways are missing. The apparent reason for this is the phenomenon that we call nonorthologous gene displacement when the same function is fulfilled by nonorthologous proteins in two organisms. We identified 22 nonorthologous displacements and supplemented the set of orthologs with the respective *M. genitalium* genes. After examining the resulting list of 262 genes for possible functional redundancy and for the presence of apparently parasite-specific genes, 6 genes were removed. We suggest that the remaining 256 genes are close to the minimal gene set that is necessary and sufficient to sustain the existence of a modern-type cell. Most of the proteins encoded by the genes from the minimal set have eukaryotic or archaeal homologs but seven key proteins of DNA replication do not. We speculate that the last common ancestor of the three primary kingdoms had an RNA genome. Possibilities are explored to further reduce the minimal set to model a primitive cell that might have existed at a very early stage of life evolution.**

The sequences of two small genomes of parasitic bacteria, *Haemophilus influenzae* and *Mycoplasma genitalium*, have been reported recently (1, 2). There is a qualitative difference between complete bacterial genomes and any sequences, including viral and organellar genomes, that have been available before. However small, a cellular gene set has to be self-sufficient in the sense that cells generally import metabolites but not functional proteins; therefore, they have to rely on their own gene products to provide housekeeping functions. Analysis of protein sequences encoded in the first two complete genomes based on this simple notion resulted in the theoretical reconstruction of unknown bacterial functional systems (3, 4). Here we systematically compare the *M. genitalium* and *H. influenzae* protein sequences in an attempt to define the minimal gene set that is necessary and sufficient for supporting cellular life.

M. genitalium that has a 0.58 megabase genome, with only 468 protein-coding genes, has been proclaimed the minimal gene complement (2). However, while this is the cellular life form with the smallest known number of genes, there is no evidence that it is indeed minimal. Clearly, the *M. genitalium* genes are sufficient to support a functioning cell but there is no indication as to what fraction of them is necessary.

M. genitalium and *H. influenzae* belong to Gram-positive and Gram-negative bacteria, respectively (5), and are likely to be separated from their last common ancestor by at least 1.5 billion years of evolution (6). *H. influenzae* is also a parasitic bacterium with a relatively small genome that is 1.83 megabases long and contains about 1700 protein-coding genes; its evolution apparently included a number of gene elimination events (1, 3). Therefore, the genes that are conserved in these two bacteria are almost certainly essential for cellular function and are likely to approximate the minimal gene set.

The original analysis of the *H. influenzae* and *M. genitalium* proteins included only the most obvious sequence similarities and the respective functional assignments (1, 2). We performed an in-depth reanalysis of the *H. influenzae* and *M. genitalium* protein sequences (3, 4) using the strategy developed in the recent studies on the *Escherichia coli* genome (7, 8). Here we use the results of a detailed comparison of *M. genitalium* and *H. influenzae* proteins to derive and characterize the minimal gene set compatible with modern-type cellular life. We then discuss possible directions of a further reduction that may be undertaken to model a primordial cell.

MATERIALS AND METHODS

Sequences and Data Bases. The nucleotide sequences of the *H. influenzae* and *M. genitalium* genomes were from refs. 1 and 2, respectively. The gene complement of each of the bacteria was reevaluated. It has been reported that *H. influenzae* possesses 1727 protein-coding genes (1). By merging overlapping open reading frames that apparently belong to the same gene and that have been separated because of frameshifts, and by eliminating short genes whose existence could not be corroborated, we have arrived at a set of 1703 predicted genes (3). The *M. genitalium* genome has been reported to contain 470 protein-coding genes (2). Our analysis detected 468 genes, two of which have been missed in ref. 2, while four of the open reading frames reported in ref. 2 could not be confirmed in our study. In addition, coding regions for two genes were extended.

All data base screening was against the protein and nucleotide versions of the daily updated nonredundant sequence data base maintained at the National Center for Biotechnology Information.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: Ndk, nucleoside diphosphate kinase; PTS, sugar phosphotransferase.

*To whom reprint requests should be addressed at: National Center for Biotechnology Information, National Library of Medicine, Building 38A, National Institutes of Health, Bethesda, MD 20894. e-mail: koonin@ncbi.nlm.nih.gov.

The information on biochemical pathways was primarily from refs. 9–14, with additional data from the PUMA database (http://www.mcs.anl.gov/home/compbio/PUMA).

Protein Sequence Analysis. A hierarchical strategy for protein sequence analysis at genome scale is described in detail elsewhere (8). Briefly, the initial data base screening was done with the BLASTP program (15), and the resulting "hits" were classified according to their taxonomic origin with the BLATAX program (8). The alignments with scores greater than 90 indicate biologically relevant relationships (8). The alignments with lower scores were further explored by analysis of conserved motifs with the CAP and MOST programs (16). Those proteins, for which significant sequence similarity was not detected by the BLASTP search and motif analysis, were searched with the TBLASTN program against the nucleotide data base translated in six frames (17). The proteins, for which no homologs were detected by these procedures, were subjected to an additional data base search with a highly sensitive version of the FASTA program (18). Low complexity regions in protein sequences, i.e., segments enriched in one or several amino acid residues that frequently produce spurious hits in data base searches, were detected and masked with the SEG program (19). A modified version of SEG was used to predict nonglobular domains in proteins (19).

Distinguishing Orthologs from Paralogs. Orthologs are genes related by vertical descent and are responsible for the same function in different organisms, in contrast to paralogs, which are homologs related by duplication and have similar but not identical functions (20). We considered the following criteria necessary and sufficient to consider two genes from the compared organisms orthologs: (*i*) the similarity between the given pair of protein sequences has to be at least several percentage points higher than that between each of these proteins and any other protein from the second organism; (*ii*) the two proteins have to be more closely related to each other than to homologs from phylogenetically more distant organisms; and (*iii*) the sequences of the two proteins should align through most of their lengths (3, 21).

RESULTS AND DISCUSSION

Reanalysis of *M. genitalium* and *H. influenzae* Protein Sequences. A detailed analysis of the sequence conservation in *M. genitalium* and *H. influenzae* proteins to predict the likely function for as many of them as possible is a prerequisite for deriving the minimal gene set. Using the combination of computer approaches outlined under *Materials and Methods*, we assessed the relevance of even the weakest observed sequence similarities, considerably moving up the limit of functional prediction compared with the original publications (Table 1; refs. 3 and 4). On the basis of sequence conservation, a function was assigned with a varying degree of precision to about 80% of the gene products in each of the bacteria (Table 1). Among the remaining proteins that did not show significant sequence conservation, particularly those of *M. genitalium*, many were predicted to contain large nonglobular domains. Some of these proteins may be involved in the adhesion of the bacteria to the host cells (14) but the absence of orthologs for these genes in *H. influenzae* or in other bacteria, for which sequence information is available, makes it unlikely that any of them perform universal functions.

Table 1. *M. genitalium* and *H. influenzae* gene products: Functional prediction and sequence conservation

Prediction and/or conservation	No. of proteins (% of total)	
	M. genitalium	*H. influenzae*
Functional prediction allowing inclusion in one of the broad functional categories (3, 4)	313 (66)	1079 (64)
General functional prediction (e.g., an enzymatic activity)	94 (11)	330 (19)
Sequence conservation only	17 (4)	177 (10)
No functional prediction; no sequence conservation	44 (9)*	128 (7)

*The majority of the *M. genitalium* proteins in this category were predicted to contain large nonglobular domains and may be implicated in the interaction between the bacterium and host cells (4).

The Minimal Gene Set: The List of Orthologs Adjusted for Nonorthologous Gene Displacement, Functional Redundancy, and Parasite-Specific Genes. Our logic in deriving the minimal gene set was straightforward. It is unlikely that any genes, except those that are indispensable for cell function, could have been conserved through the 1.5 billion years or more separating *H. influenzae* and *M. genitalium* from their last common ancestor, given that the evolution in both lineages has been replete with gene elimination (1–4). Therefore, the orthologous genes of the two bacteria should comprise the core of the minimal gene set. We compared the 468 *M. genitalium* protein sequences to the 1703 *H. influenzae* sequences with BLASTP and examined the results case by case using the criteria listed under *Materials and Methods* to identify orthologs. We found that 240 genes appear to be orthologous in *H. influenzae* and *M. genitalium* (Table 2).

It is likely that most, if not all, of the *M. genitalium*/*H. influenzae* orthologs are necessary for cell function but taken together, they are not sufficient. Inspection of the ortholog set identified a number of missing links in essential pathways (Table 2; see also below). The reason for these gaps in the emerging minimal gene set is the phenomenon that we called nonorthologous gene displacement, that is, the presence of nonorthologous (paralogous or unrelated) genes for the same function in different organisms (22). Striking examples are phosphoglycerate mutase, an intermediate enzyme in the glycolytic pathway, and nucleoside diphosphate kinase (Ndk), the terminal enzyme of DNA and RNA precursor biosynthesis. Adding the genes involved in nonorthologous displacement to the orthologs results in an apparently self-sufficient gene set that seems to encode all the systems necessary to support a simple cell and is likely to approximate the minimal set (Table 2).

Generally, in the nonorthologous gene displacement situations, it is unclear which of the two genes involved belongs to the minimal set. We choose to include in Table 2 the 22 respective genes from *M. genitalium*, the simpler of the two "parents" of the minimal gene set. At least in two instances, this choice seemed to be justified by additional considerations. *M. genitalium* does not encode RNase H (2), and we hypothesize that the respective function is performed by the product of the MG262 gene, a putative 5′-3′ exonuclease that is homologous to the exonuclease domain of DNA polymerase I and distantly related to RNase H. This gene appears to be a better candidate for inclusion in the minimal gene set than RNase H as it is actually present both in *M. genitalium* and in *H. influenzae*, either as a stand-alone protein or as a distinct domain of DNA polymerase I.

In the case of Ndk, we observed that *M. genitalium* does not encode a homolog of this enzyme, which is highly conserved in other organisms, including *H. influenzae* (23). We detected two candidates for the role of a novel Ndk among the *M. genitalium* proteins that had no orthologs in *H. influenzae*. Both of these proteins (MG264 and MG268) are distantly related to other nucleoside and nucleoside monophosphate kinases containing the wide-spread P-loop motif (24), a protein class already represented by a number of species in the emerging minimal gene set. Thus, it seemed logical to include in the minimal set one of these proteins rather than the *H. influenzae* Ndk, given also that in *E. coli* Ndk is dispensable (23).

For two apparently essential enzymatic reactions, namely the formation of dUMP from dUTP, catalyzed by dUTPase (Dut) in other organisms, including *H. influenzae*, and folylpolyglutamate synthesis, catalyzed by folylpolyglutamate synthase (FolC), we could not identify obvious candidates among *M. genitalium* pro-

Table 2. The minimal gene set deduced by comparing *M. genitalium* and *H. influenzae* protein sequences

Functional system	Minimal set: *M. genitalium*/*H. influenzae* orthologs, supplemented with *M. genitalium* proteins*	Proteins†	*M. genitalium* genes that are not in minimal gene set	Proteins†
Translation	Virtually complete translation apparatus, including RNA- and protein-modification enzymes (pseudoU-synthetases, methyltransferases); nonorthologous displacements: glycyl- and prolyl-tRNA synthetases (MG251 and MG283); no tRNA-nucleotidyltransferase; MG336 is a putative tRNA-Glu into tRNA-Gln.	**95** 005, 012, 021, 026, 035, 036, 070, 081–083, 087–090, 092, 093, 104, 106, 113, 126, 136, 142–143, 150–169, 172–176, 178, 182, 194, 195–198, 209, 232, 234, 251–253, 257, 258, 266, 283, 292, 325, 334, 336, 345, 346, 361–363, 363a, 365, 375, 378, 417, 418, 424–426, 433, 435, 444–446, 451, 455, 462, 463, 465, 466	An S6 modification protein (RimK paralog), a helicase, a putative pseudoU synthase.	**3** 011, 308, 370
Replication	Virtually complete replication apparatus but no RNase H; nonorthologous displacements: a candidate primer-removing nuclease (MG262) and histone-like chromosome condensation protein (MG353).	**18** 001, 003, 004, 091, 094, 122, 203, 204, 244, 250, 254, 259, 261, 262, 351, 353, 420, 469	Truncated paralog of primase; several helicases; one of the two DNA polymerase III-like genes.	**7** 007, 010, 018, 031, 140, 298, 470
Recombination and repair	Rudimentary repair system (4): no photorepair, reduced system of excision repair, one DNA polymerase (MG261) may be involved in both replication and repair, MG262 (exonuclease) probably replaces the exonuclease activity of DNA polymerase I and RNase H.	**8** 073, 097, 206, 262a, 339, 358, 359, 421	DNA modification enzymes; orthologs of two UV protection proteins, bacterial DinB and plant DRT102.	**5** 184, 235, 360, 396, 438
Transcription	Four RNA polymerase subunits including a single sigma factor. Three transcription factors but no termination factor rho; no helix-turn-helix transcription regulators; ppGpp synthetase may mediate global transcription regulation.	**9** 054, 141, 177, 249, 278, 282, 340, 341, 367	RNA polymerase subunit RpoE.	**1** 022
Chaperone functions	All groups of chaperones, except for HSP90‡; PMSR (MG408) and a conserved domain (MG449), which may be functionally associated with PMSR, are fused into a single protein in *H. influenzae*.	**13** 019, 201, 238, 239, 297, 305, 355, 392, 393, 408, 448, 449, 457	Two DnaJ paralogs.	**2** 002, 200
Nucleotide metabolism	Synthesis of ribonucleotides from ribose, ATP, and nitrous bases; ribonucleotide reduction coupled with thioredoxin oxidation. dTMP synthesis occurs by thymidylate-synthase-mediated methylation of dUMP. This requires a phosphohydrolase activity replacing the missing *dut* gene (one candidate is MG125, which is predicted to possess hydrolase activity). Nitrous base transport system uncertain; the spermidine-putrescine uptake proteins are the candidates. Nonorthologous displacements: cytidine deaminase (MG052), phosphoribosyl pyrophosphate synthase, (MG058), and a putative novel nucleoside diphosphate kinase (MG268; see text).	**23** 006, 030, 042–045, 049, 052, 058, 102, 107, 124, 127, 171, 227, 229, 231, 268, 276, 330, 382, 434, 458	Thymidine kinase, thymidine phosphorylase, a putative nucleotide kinase.	**4** 034, 051, 264
Amino acids metabolism	MG336 is a putative aminotransferase that might convert tRNA-Glu into tRNA-Gln; ATP-dependent transport systems for oligopeptides and for amino acids; several proteases.	**7** 077–080, 180, 336, 391	Five putative permeases; a periplasmic binding protein.	**11** 020, 067, 068, 183, 225, 226, 294, 324, 326, 388, 461
Lipid metabolism	Synthesis of lipids and phospholipids from glycerol and exogeneous fatty acids; acyl carrier protein implied in fatty acids uptake.	**6** 114, 212, 287, 293, 333, 437	A HALO-family acyltransferases; a protein related to PlsX of *E. coli*; a putative glycerophosphoryl diester phosphodiesterase.	**3** 344, 368, 385
Energy	Complete Embden-Meyerhof pathway, conversion of pyruvate into coenzyme A, and further into acetate; uptake and utilization of hexoses, penthoses, and trioses; H^+ ATPase. Non-orthologous displacements: fructose-biphosphate aldolase (MG023), glycerol uptake facilitator protein (MG033), pyruvate dehydrogenase (MG273 and MG274), phosphoglucomutase (MG053), phosphoglycerate mutase (MG430).	**34** 023, 033, 038, 050, 053, 063, 111, 112, 119, 120, 187, 215, 216, 271, 272, 273, 274, 275, 299, 300, 301, 357, 398–407, 430, 431	A PTS system; six permeases; eukaryotic-type protein phosphatase and kinase; a putative monocarboxyamide-specific amidase; an ATP-binding enzyme (putative sugar kinase); L-lactate dehydrogenase M-chain, one more dehydrogenase.	**17** 039, 041, 061, 062, 069, 099, 108, 109, 121, 128, 129, 188, 189, 302, 303, 429, 460
Coenzyme metabolism and utilization	NAD, FAD, and SAM are synthesized from exogenous monomers; ligation of lipoate onto proteins; MG383 (NAD synthase) and MG270 (lipoate-protein ligase) are displaced by nonorthologous proteins in *H. influenzae*. Four enzymes for turnover of folate derivatives. Minimal gene set codes for eight enzymes that use dinucleotide coenzymes (two or three use FAD, the others use NAD or NADP), two of each: lipoate-, pyridoxal-, thiamine-, and folate-dependent enzymes, and three SAM-utilizing methyltransferases§	**8** 013, 047, 145, 228, 245, 270, 383, 394		**0**

Table 2. (*Continued*)

Functional system	Minimal set: *M. genitalium/H. influenzae orthologs, supplemented with M. genitalium* proteins*	Proteins†	*M. genitalium* genes that are not in minimal gene set	Proteins†
Exopolysaccharides	Transketolase¶; galactowaldenase; glycosyltransferases involved in cell wall biosynthesis in *H. influenzae*; they may take part in capsule biosynthesis in *M. genitalium*.	**8** 059, 060, 066, 086, 118, 210, 224, 453	A glycosyltransferase and a dTDP-dehydrohamnose reductase.	**2** 025, 037
Uptake of inorganic ions	Phosphate permease, Na^+ ATPase, cationic/metal ATPases.	**5** 065, 071, 322, 410, 411	A protein similar to phosphate transport regulator.	**2** 323, 409
Secretion, receptors	Signal recognition particle, preprotein translocase.	**5** 015, 048, 072, 170, 297	Parasite-specific proteins: cytadherence-accessory proteins, hemolysin, adhesins, sialoglycoprotease, multidrug-resistance-type ATPases.	**17** 014, 046, 146, 190–192, 303, 312–318, 386, 423, 467
Other conserved proteins	Proteins with ATP-, GTP-, NAD-, FAD-, and SAM-binding domains; one permease; three proteins with uncharacterized conserved regions.	**18** 008, 009, 024, 056, 125, 138, 221, 222, 247, 265, 295, 332, 335, 347, 379, 380, 384, 387		
Other proteins of *M. genitalium* not conserved in *H. influenzae*			Functional prediction, at least in general terms, available for 72 proteins: NTPases, methyltransferases, acyltransferases, metal-binding proteins, integral membrane proteins, coiled-coil proteins presumably involved in cell-cell interactions	**139**
Total protein-coding genes		**256** (**234** *M. genitalium/H. influenzae* orthologs, **22** nonorthologous displacements)		**212** (six orthologs)

PTS, sugar phosphotransferase; PMSR, peptide methionine sulfoxide reductase.

*In all cases of nonorthologous gene displacement, the respective proteins of *M. genitalium* were included in the minimal set for consistency (see text for further discussions).

†Boldface numbers are the number of proteins in a group. Proteins of *M. genitalium* are designated as in ref. 2 (with the exception of MG262a and 363a, proteins that have not been identified in ref. 2), with MG omitted for brevity. Products of genes MG193, MG413, MG416, and MG468 were missing from the original GenBank data base submission (2), and we could not confirm their existence. Underline indicates the cases of nonorthologous gene displacement; double underline indicates *M. genitalium/H. influenzae* orthologs excluded from the minimal gene set.

‡In addition to characterized families of chaperones, the minimal gene set includes representatives of three newly described classes of putative chaperone-like proteins (MG409, MG448, and MG449; unpublished work).

§In *M. genitalium*, serine hydroxymethyltransferase (MG394) is probably involved in folate metabolism (even though it has been associated with the amino acid metabolism in ref. 2). There is no folylpolyglutamate synthase.

¶This category includes transketolase (MG066) that is probably involved in heptose biosynthesis and is unlikely to have a bioenergetic role, contrary to the functional assignment in ref. 2.

teins. A predicted hydrolase without a known function may be a candidate for a novel Dut activity; the list of orthologs also includes three proteins without a known activity that may contain a candidate enzyme of folylpolyglutamate synthesis (Table 2). Alternatively, novel enzymes with these activities may be eventually found among *M. genitalium* proteins without orthologs in *H. influenzae;* that is, there should be two additional cases of nonorthologous gene displacement.

Is the derived set of 262 genes, comprised by the 240 orthologs together with the 22 cases of nonorthologous replacement, not only sufficient but also necessary for a cell existence? We examined this collection of genes for possible functional redundancy and for genes that might be specific for parasitic bacteria, with the understanding that this aspect of our analysis is speculative. We identified only two cases of apparent redundancy but both involved functional systems of major importance. First, we removed from the minimal gene set thymidine kinase (TK), an enzyme of the salvage pathway of thymidine triphosphate biosynthesis. Thymidine triphosphate is synthesized in two pathways: (*i*) salvage from thymine via thymidine, with three phosphorylation steps, and (*ii*) *de novo* from UMP, with two phosphorylation steps (Fig. 1). *H. influenzae* does not encode thymidine phosphorylase (*deoA* gene product, which is present in *M. genitalium*) and therefore the complete salvage pathway does not seem to belong in the minimal gene set. We suggest that the only pathway of thymidine triphosphate formation encoded by the minimal gene set is from UMP. This assignment requires a nonorthologous displacement of Dut (Table 2 and see above). Under this scheme, thymidine kinase becomes superfluous, and we eliminated it from the minimal set, in spite of the fact that the *tk* genes of *H. influenzae* and *M. genitalium* are orthologous.

The second, perhaps most controversial example of apparently redundant orthologous genes involves the PTS system. Even though the three proteins comprising the fructose PTS system are orthologous in *H. influenzae* and *M. genitalium*, PTS seems to be an unlikely part of the minimal gene set as it has been detected only in a subset of bacteria (25). We postulated that the minimal gene set encodes only membrane ATPases and permeases comprising an ATP-dependent system for sugar transport (Table 2); the subsequent phosphorylation of sugars is catalyzed by sugar kinases (MG063 in the minimal gene set), which are ubiquitous and ancient enzymes (26).

The list of orthologs included only two genes, encoding a sialoglycoprotease and a hemolysin, that are apparently parasite-specific. These genes were also removed from the minimal set.

Thus, the construction of the minimal gene set included three distinct steps: (*i*) detection of orthologs among the *H. influenzae*

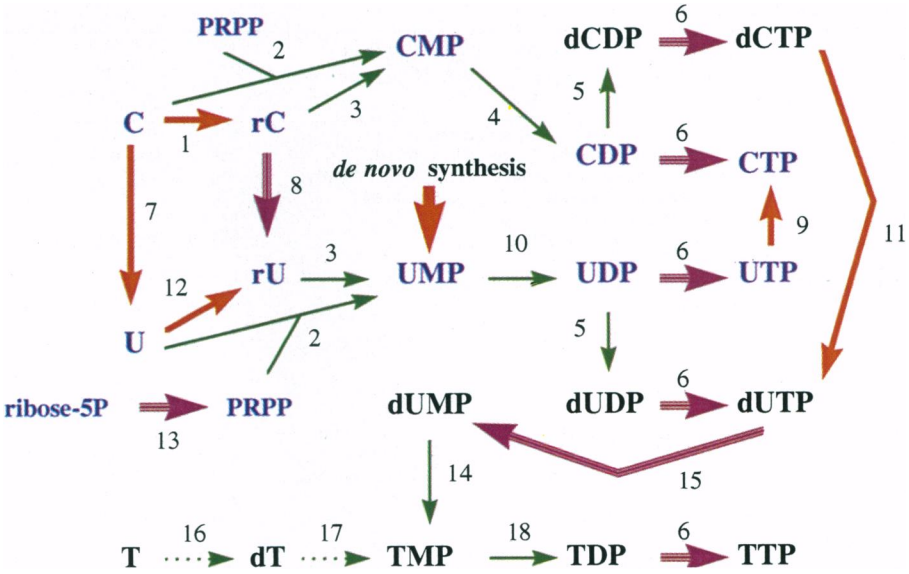

FIG. 1. Pyrimidine nucleotide biosynthesis: the steps of reduction. Red arrows indicate reactions occurring in most organisms but not in *M. genitalium* or in a putative cell with the minimal gene set. Solid green arrows indicate reactions catalyzed by known, orthologous enzymes in *M. genitalium* and in *H. influenzae* and accordingly included in the minimal set. Dotted green arrows indicate reactions identified in *M. genitalium* but not included in the minimal set. Purple arrows indicate reactions that in *M. genitalium* and in *H. influenzae* are catalyzed by nonorthologous enzymes (nonorthologous displacement). Deoxynucleosides and deoxynucleotides are shown in black. The reactions leading to them would be eliminated from a hypothetical primitive cell with an RNA genome (see text). The enzymes for each of the reactions are indicated by numbers: 1, pyrimidine nucleoside phosphorylase; 2, uracil phosphoribosyltransferase; 3, uridine kinase; 4, cytidylate kinase; 5, ribonucleotide diphosphate reductase; 6, nucleoside diphosphate kinase; 7, cytosine deaminase; 8, cytidine deaminase; 9, CTP synthase; 10, uridylate kinase; 11, dCTP deaminase; 12, uridine phosphorylase; 13, phosphoribosyl pyrophosphate synthase; 14, thymidylate synthase; 15, dUTPase; 16, thymidine phosphorylase; 17, thymidine kinase; 18, thymidylate kinase. The *de novo* synthesis of uridine phosphate includes five enzymatic steps.

and *M. genitalium* genes; (*ii*) filling gaps in the biochemical pathways by amending the set of orthologs with genes involved in nonorthologous displacement; (*iii*) elimination of genes that appeared to be functionally redundant or parasite-specific. We believe that the remaining 256 genes are a close approximation of the minimal gene repertoire that is necessary and sufficient for the existence of a modern-type cell.

The Main Features of the Minimal Gene Set. The derived minimal gene set defines a hypothetical simple cell with the following principal features (Table 2):

1. A nearly complete system of translation. However, there is no gene for tRNA nucleotidyltransferase (all 3′-terminal CCA sequences in *M. genitalium* tRNAs are genome encoded), glutaminyl-tRNA synthetase (tRNA-Gln is formed from tRNA-Glu in *M. genitalium* and other Gram-positive bacteria; ref. 27), and some modification enzymes.

2. A virtually complete DNA replication machinery, with the notable exception of RNase H (see above).

3. A rudimentary system of recombination and repair (4). The repair DNA polymerase is a special issue as *M. genitalium* does not encode DNA polymerase I, which is most likely responsible for the repair synthesis in *H. influenzae*, but encodes two α subunits of DNA polymerase III, one of which is orthologous to *H. influenzae* DnaE and the other is orthologous to PolC of Gram-positive bacteria (28). In the construction of the minimal gene set, we assumed that one DNA polymerase III may be responsible for both replicative and repair DNA synthesis.

4. Transcription apparatus that consists of four RNA polymerase subunits, including a single σ factor, which is the only protein with a helix-turn-helix domain in the minimal set, and three transcription factors, but no termination factor Rho. Virtually no proteins for transcription regulation, with the exception of guanosine tetraphosphate synthetase (SpoT; ref. 29), which is likely to be involved in global transcription regulation (4).

5. A surprisingly large set of chaperone-like proteins, including two poorly characterized ones, namely peptide methionine sulfoxide reductase and another domain whose specific function is not known but that is associated with peptide methionine sulfoxide reductase in a variety of proteins from different organisms (unpublished observations). These proteins represent an unusual case of nonorthologous displacement as they are encoded by two distinct genes in *M. genitalium* but by a single gene in *H. influenzae*.

6. Completely anaerobic intermediary metabolism virtually reduced to glycolysis and substrate phosphorylation.

7. No amino acid biosynthesis, with the single exception of the transamination of glutamyl-tRNA into glutaminyl-tRNA (see footnote to Table 2 concerning serine hydroxymethyltransferase).

8. No *de novo* nucleotide biosynthesis but complete salvage pathways for all bases except thymine (see above).

9. Lipid biosynthesis limited to condensation of fatty acids with glycerol; no fatty acid synthesis.

10. Only eight enzymes requiring complex cofactors, other than mono- and dinucleotides, and eight enzymes involved in cofactor biosynthesis.

11. A machinery for protein export.

12. A limited repertoire of metabolite transport systems including membrane ATPases and permeases, presumably with a broad specificity.

We tentatively kept in the minimal set 18 proteins with unassigned functions. All of them are orthologous and highly conserved in *M. genitalium* and *H. influenzae*, and 15 contain known motifs that allow the prediction of their activity. These genes may account for as yet uncharacterized but essential functions.

Inevitably, a hypothetical cell with the minimal gene set is a simplified derivative of *M. genitalium*. The remaining *M. genitalium* genes encode, along with proteins with large nonglobular domains implicated in parasite–host interaction, a variety of globular proteins with diverse activities, including a number of enzymes (Table 2). None of these proteins, however, appeared to have a unique function that is likely to be essential and is not already represented by a functionally related gene in the minimal set.

Fig. 1 illustrates the reduction of the pyrimidine biosynthesis pathways from the general scheme typical of most organisms to the abridged version in *M. genitalium* and to the further simplification in a hypothetical organism with the minimal genome.

The analysis described here involved two bacterial genomes. Have we constructed a minimal set of genes for cellular life in general or for bacteria only? We believe that it is a general minimal set with specifically bacterial solutions for certain functions. Among the 256 proteins comprising the minimal set, 71% show significant similarity to eukaryotic or archaeal proteins as opposed to 48% among all *M. genitalium* proteins.

Thus, the hypothetical minimal genome is enriched in universally conserved proteins. The subset of proteins without eukaryotic or archaeal homologs, along with some that are likely to evolve too fast to reveal ancient conserved regions (e.g., permeases and several small ribosomal proteins), includes at least two conspicuous groups. These are the eight subunits of the H^+ ATPase that in eukaryotes are represented by organellar genes and eight key proteins of DNA replication, namely two subunits of DNA polymerase III, initiator ATPase (DnaA), helicase (DnaB), primase (DnaG), NAD-dependent DNA ligase, and single-stranded DNA-binding proteins (Ssb and Dbh).

Using a completely different approach to estimate the minimal genome size required for life, Itaya (30) has determined the percentage of randomly selected genetic loci in *Bacillus subtilis* that could be disrupted without loss of viability. The estimated minimal genome size of 318 kbp in these experiments corresponds to 254 genes in the case of *M. genitalium* (the average gene length *ca.* 1.25 kb; ref. 2) and is remarkably close to the size of our minimal gene set.

Life Beyond the Minimal Set? It appears unlikely that the minimal gene set derived from the comparison of *M. genitalium* and *H. influenzae* can be significantly reduced without dramatically affecting functional systems that are essential for any extant cell, such as the translation or the replication machinery. As a matter of speculation, one can imagine, however, how the minimal set could be simplified to model a primitive cell, in which such essential systems might have been significantly simpler than they are in modern cells.

Such further reduction may proceed in several directions. (*i*) Examine pathways requiring complex cofactors and eliminate those of them that can be bypassed without the use of the cofactors. (*ii*) Eliminate the remaining regulatory genes. (*iii*) Delineate paralogs and replace at least the most highly conserved families with a single, presumably multifunctional "founder." (*iv*) Apply the parsimony principle (31): those systems and genes that are not found in both bacteria and eukaryotes or both bacteria and archaea are unlikely to come from a primitive cell.

A detailed implementation of these approaches that may be relevant for modeling very early stages of life evolution has to be deferred until completion of genome sequences of archaea and eukaryotes (at the time of revision of this manuscript, the complete sequence of the yeast genome is already available). Nevertheless, as the great majority of eukaryotic protein families are represented in the sequence data bases (32), some of the results may be foreseen, and those that may emerge from a consistent application of the parsimony principle appear to be striking. In particular, given the absence of eukaryotic or archaeal homologs of the key proteins of bacterial DNA replication (see above), it seems likely that the last common ancestor of the three primary kingdoms had an RNA genome. Evidently, this would also entail the absence of pathways for DNA precursor biosynthesis (Fig. 1).

Furthermore, as archaea have significant deviations in the enzymology of the upstream reactions of glycolysis (33), this ancestor might have had an ultimately simplified intermediary metabolism using trioses and pentoses as the source of energy through conversion to glyceraldehyde 3-phosphate, which is oxidized to pyruvate in the downstream reactions of glycolysis.

The consequences of complete elimination of paralogy from the minimal gene set also may be dramatic. The most obvious example is the reduction of the number of aminoacyl-tRNA synthetases, ultimately to two species with a broad specificity (34).

It has to be kept in mind that not only reduction but also certain additions to the minimal gene are likely to be required to produce a realistic model of a primitive cell. The most important of such additions may be a simple system for photo- or chemoautotrophy. It may become possible to glean the essential features of such systems from complete genome sequence of autotrophic organisms. Eventually, the backwards extrapolation from the minimal gene set may lead close to the origin of life itself.

Availability of the Results. The annotated minimal gene set and the detailed results of *M. genitalium* and *H. influenzae* genome analysis are available via the World Wide Web at http://www.ncbi.nlm.nih.gov/Complete_Genomes and by anonymous ftp at ncbi.nlm.nih.gov in the directories repository/MINSET, repository/HIN, and repository/MG.

We are grateful to M. Boguski, P. Bork, M. S. Gelfand, M. Griep, D. Landsman, D. J. Lipman, P. A. Pevzner, J. Reizer, K. E. Rudd, E. E. Selkov, T. G. Senkevich, and C. Woese for critical reading of the manuscript and helpful discussions, and to R. L. Tatusov for programming and valuable contributions to data analysis.

1. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al.* (1995) *Science* **269,** 496–512.
2. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., *et al.* (1995) *Science* **270,** 397–403.
3. Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayes, W., Borodovsky, M., Rudd, K. E. & Koonin, E. V. (1996) *Curr. Biol.* **6,** 279–291.
4. Koonin, E. V., Mushegian, A. R. & Rudd, K. E. (1996) *Curr. Biol.* **6,** 404–416.
5. Olsen, G. J., Woese, C. R. & Overbeek, R. (1994) *J. Bacteriol.* **176,** 1–6.
6. Doolittle, R. F., Feng, D.-F., Tsang, S., Cho, G. & Little, E. (1996) *Science* **271,** 470–477.
7. Koonin, E. V., Tatusov, R. L. & Rudd, K. E. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 11921–11925.
8. Koonin, E. V., Tatusov, R. L. & Rudd, K. E. (1996) *Methods Enzymol.* **266,** 295–322.
9. Neidhardt, F. C., eds. (1996) *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology* (Am. Soc. Microbiol., Washington, DC).
10. Razin, S. (1991) in *The Prokaryotes*, eds. Balows, A., Truper, H. G., Dworkin, M., Harder, W. & Schleifer, K. H. (Springer, New York), pp. 1937–1959.
11. Hoiseth, S. (1991) in *The Prokaryotes*, eds. Balows, A., Truper, H. G., Dworkin, M., Harder, W. & Schleifer, K. H. (Springer, New York), pp. 3304–3330.
12. Kilian, M. & Biberstein, E. L. (1984) *Bergey's Manual of Systematic Bacteriology* (Williams & Wilkins, Baltimore), pp. 558–569.
13. Rasin, S. & Freundt, E. A. (1984) *Bergey's Manual of Systematic Bacteriology* (Williams & Wilkins, Baltimore), pp. 740–770.
14. Kahane, I. & Horowitz, S. (1993) *Subcell. Biochem.* **20,** 225–242.
15. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
16. Tatusov, R. L., Altschul, S. F. & Koonin, E. V. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 12091–12095.
17. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6,** 119–129.
18. Pearson, W. R. (1996) *Methods Enzymol.* **266,** 227–257.
19. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266,** 554–573.
20. Fitch, W. M. (1970) *Syst. Zool.* **19,** 99–106.
21. Bork, P., Ouzounis, C., Casari, G., Schneider, R., Sander, C., Dolan, M., Gilbert, W. & Gillevet, P. M. (1995) *Mol. Microbiol.* **16,** 955–967.
22. Koonin, E. V., Mushegian, A. R. & Bork, P. (1996) *Trends Genet.*, in press.
23. Lu, Q., Zhang, X., Almaula, N., Mathews, C. K. & Inouye, M. (1995) *J. Mol. Biol.* **254,** 337–341.
24. Saraste, M., Sibbald, P. R. & Wittinghofer, A. (1990) *Trends Biochem. Sci.* **15,** 430–434.
25. Romano, A. H. & Saier, M. H., Jr. (1992) in *The Evolution of Metabolic Function,* ed. Mortlock, R. P. (CRC, Boca Raton, FL), pp. 171–204.
26. Bork, P., Sander, C. & Valencia, A. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 7290–7294.
27. Strauch, M. A., Zalkin, H. & Aronson, A. I. (1988) *J. Bacteriol.* **170,** 916–920.
28. Koonin, E. V. & Bork, P. (1996) *Trends Biochem. Sci.* **21,** 128–129.
29. Condon, C., Squires, C. & Squires, C. L. (1995) *Microbiol. Rev.* **59,** 623–645.
30. Itaya, M. (1995) *FEBS Lett.* **362,** 257–260.
31. Benner, S. A., Ellington, A. D. & Tauer, A. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 7054–7058.
32. Green, P., Lipman, D., Hillier, L., Waterston, R., States, D. & Claverie, J. M. (1993) *Science* **259,** 1711–1716.
33. Danson, M. J. & Hough, D. W. (1992) *Biochem. Soc. Symp.* **58,** 7–21.
34. Eriani, G., Delarue, M., Poch, O., Gangloff, J. & Moras, D. (1990) *Nature (London)* **347,** 203–206.