Sequencing and analysis of bacterial genomes

Eugene V. Koonin, Arcady R. Mushegian and Kenneth E. Rudd

The complete sequences of two small bacterial genomes have recently become available, and those of several more species should follow within the next two years. Sequence comparisons show that the most bacterial proteins are highly conserved in evolution, allowing predictions to be made about the functions of most products of an uncharacterized genome. Bacterial genomes differ vastly in their gene repertoires. Although genes for components of the translation and transcription machinery, and for molecular chaperones, are typically maintained, many regulatory and metabolic systems are absent in bacteria with small genomes. *Mycoplasma genitalium*, with the smallest known genome of any cellular life form, lacks virtually all known regulatory genes, and its gene expression may be regulated differently than in other bacteria. Genome organization is evolutionarily labile: extensive gene shuffling leaves only very few conserved gene arrays in distantly related bacteria.

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA.

Current Biology 1996, Vol 6 No 4:404–416

Introduction

The start of a new era in genome science can be dated precisely: July 28, 1995. On that day, the paper describing the 1.8 Megabase (Mb) sequence of the *Haemophilus influenzae* genome, the first complete genome sequence for a cellular life form, was published [1]. The complete sequence of the 0.58 Mb genome of *Mycoplasma genitalium* followed in less than three months [2]. The expectation is that, within the next two years, the number of completely sequenced bacterial and archaeal genomes will reach at least ten. Faced with these rapid developments, one is forced to ask whether complete genome sequences provide for a qualitatively new understanding of the genome. In particular, are there major problems that can be addressed with complete genome sequences, but not with partial sequences? We believe that the first two genome sequences already answer these questions with a definite “yes”. We shall review the status of bacterial and archaeal genome sequencing projects, and discuss the strategies for computer analysis of genome sequences, the methodological challenges ahead and the new understanding of genomes that is emerging now that complete sequences are available.

Current status

Genome sequences accumulate from two main sources — first, from individual laboratories, which gather short sequences in the course of functional studies, and second, from genome sequencing projects, which aim to determine long sequences independently of function. Until recently, the small-scale studies produced most of the data. In the last three years, however, the situation has changed dramatically, and now genome projects produce most of the sequence information, largely freeing individual investigators from the task of sequencing. The first bacterial genome project, started in 1991, aimed to sequence the complete *Escherichia coli* genome [3]. Since then, genome projects have been initiated for a variety of bacteria and archaea (Table 1). The genomes being sequenced represent a broad cross-section of the universal phylogenetic tree (Fig. 1). One may thus hope that these genome sequences will provide a revealing, if incomplete, picture of prokaryotic genome diversity.

Several specialized bacterial genome databases are actively maintained. These include four independent databases collecting information on the *E. coli* genome [4–7], two integrated *Bacillus subtilis* genome databases [8,9] and the new databases on the *H. influenzae* and *M. genitalium* genomes maintained by The Institute for Genome Research (TIGR) and accessible *via* the World Wide Web [1,2]. A ‘Genomes’ division of the GenBank database has

Table 1

Status of bacterial and archaeal genome sequencing projects.

Species	Taxonomic division	Genome size (Mb)	Available sequences/ longest contig/ % completed	Projected date of completion	Laboratory/ institution	References
H. influenzae	Gram-negative bacteria/ purple bacteria/ gamma subdivision	1.83	1.83/1.83/100 %	1995	TIGR	[1]
M. genitalium	Low G+C Gram-positive bacteria	0.58	0.58/0.58/100 %	1995	TIGR	[2]
E. coli	Gram-negative bacteria/ purple bacteria/ gamma subdivision	4.7	3.52/1.61/75 %	1996	Lab. Genet., Un. Wisconsin-Madison; Inst. Virus Res., Kyoto Univ.	[3,11]
B. subtilis	Low G+C Gram-positive bacteria	4.17	1.48/0.18/36 %	1997	European consortium; Japanese consortium	[12,13]
Mycoplasma pneumoniae	Low G+C Gram-positive bacteria	0.8	0.48/?/60 %	1996	Un. Heidelberg	[72]
Synechocystis sp.	Cyanobacteria	3.6	1.0/1.0/28 %	1996	Kazusa DNA Res. Institute	[73]
Chlamydia trachomatis	Planctomyces/ chlamydia group	1.04	?	?	Dept. Biochem, Stanford Univ	[74]
Methanococcus jannaschii	Archaea/ euryarchaeota	?	?	1996	TIGR	[75]
Mycobacterium leprae	High G+C Gram-positive bacteria	2.8	1.8/?/64 %	?	Genome Therapeutics Corporation (GTC); Institut Pasteur	[76,77]
Mycobacterium tuberculosis	High G+C Gram-positive bacteria	4.0	0.8/?/20 %	?	GTC	[76]
Methanobacterium thermoautotrophicum	Archaea/ euryarchaeota	1.7	?	1996	GTC	[76]
Synechococcus sp	Cyanobacteria	2.7	?	?	GTC	[76]
Haloferax volcanii	Archaea/ euryarchaeota	?	?	?	GTC	[76]
Methanopyrus kandleri	Archaea/ euryarchaeota	?	?	?	GTC	[76]
Rhodococcus rhodochrous	Gram-negative bacteria/purple bacteria/alpha subdivision	?	?	?	GTC	[76]
Pyrococcus furiosus	Archaea/ euryarchaeota	?	?	?	Ctr. Marine Biotech Un. Maryland-Baltimore	[78]

Sulfolobus solfataricus	Archaea/ crenarchaeota	3.1	?	?	Un. Ottawa;
					Inst. Marine Biosci., Halifax;
					Dalhousie Un.
						http://www.imb.nrc.ca/imb/sulfolob

been established very recently, with the specific purpose of representing complete genome sequences [10]. Furthermore, two integrated computer systems have been recently developed that are specifically designed to store and semi-automatically analyze sequences on a genome scale [11,12].

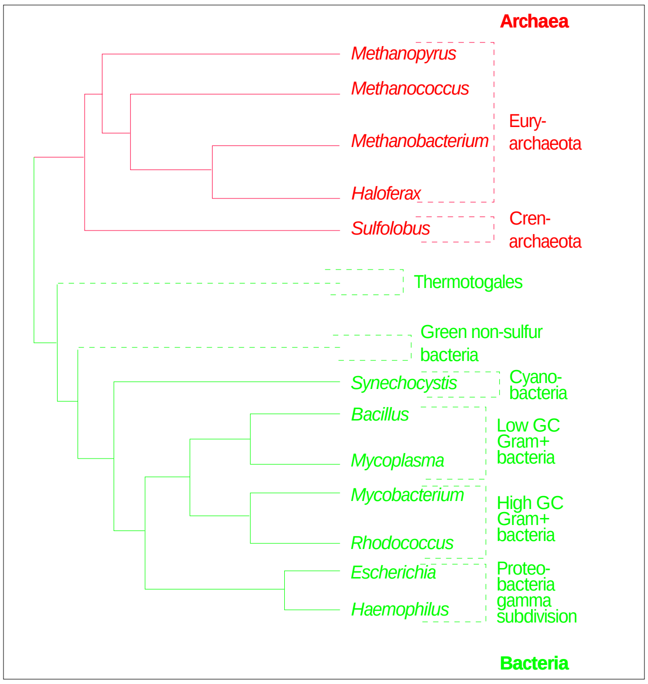
Technical basis and quality

The two projects that were initiated first, but are still not yet finished, aimed to sequence the complete genomes of *E. coli* [3,13] and *B. subtilis* [14,15]. In these projects, the approach has been to sequence subcloned λ inserts of known chromosomal position. The *H. influenzae* [1] and *M. genitalium* [2] genomes, in contrast, were sequenced using a random, shotgun strategy. This revolutionary approach entails sequencing several thousand short clones, with most of the genome sequenced several times, followed by the assembly of contiguous sequences ('contigs') and gap closure using DNA hybridization, the polymerase chain reaction (PCR), and a variety of standard and specifically designed computer programs [1]. The early genome projects used slab gel electrophoresis and autoradiography, but this approach has been superseded by automated DNA sequencers using fluorescently labeled nucleotide analogs.

Unfortunately, there has been little attempt seriously to evaluate sequencing accuracy in these various projects. Comparisons of sequences from the same genome regions determined in independent laboratories may provide rough estimates of accuracy, but as long as the accuracy is not known precisely for any of the copies, this is not a satisfactory criterion. In particular, redundancy does not guarantee the resolution of non-random sequencing errors that occur in specific positions because of compression or other artifacts. However, the use of thermostable polymerases and nucleotide analogs such as dITP has improved the resolution of compressions, and an assessment of the random error rate by re-sequencing seems appropriate. Furthermore, the most serious type of sequencing errors, frameshifts, could be dramatically reduced by co-sequencing the genomes of two closely related organisms, such as *E. coli* and *Salmonella typhimurium*.

The *E. coli* genome project originally claimed an accuracy of about one ambiguity per 600 nucleotides [3]. Recently, rates of one error per 5 000–10 000 bases for *H. influenzae* [1] and one error per 10 000 bases for *M. genitalium* [2] have been estimated. Still, it remains unclear what should be considered a 'final' genome sequence [16]. The *H. influenzae* chromosome sequence — a total of 1 830 137 nucleotides — contains not only 119 ambiguous nucleotides, which is compatible with the above error rate, but also up to 100 frameshifts and nonsense mutations interrupting open reading frames (ORFs). Furthermore, the sequence contains several 'orphan' gene fragments and two long repeats of protein-coding regions whose origin remains uncertain [1,17]. Even though each of the respective sequences has been determined from multiple independent clones, in our opinion it would be premature to conclude that all of these anomalies are real mutations that have accumulated in the *H. influenzae* strain chosen for sequencing. A careful re-sequencing of the frameshifted regions directly from natural and laboratory strain genomic DNA would likely resolve this important issue.

Figure 1



The phylogenetic distribution of bacterial species that are the subjects of genome-sequencing projects. The tree topology, based on 16 S rRNA sequences, is from [79]. The solid lines indicate lineages that include archaea (red) and bacteria (green) whose genomes are being sequenced. Major branches that are not represented in genome-sequencing projects are shown by broken lines.

Sequence analysis strategy

Genome sequencing presents challenges to computer analysis at all levels, from sequence assembly to large-scale genome comparisons (looking for evidence of evolutionary rearrangments, for example). Genome sequence analysis is a multistep process that starts with establishing the maximally accurate assembled sequence, and proceeds through functional predictions to higher-level genome comparisons [18]. We shall focus on the latter steps of this scheme, from gene prediction to genome rearrangement analysis.

Gene prediction

Once a genome sequence has been determined, an immediate task is to identify all the genes. Arguably, the best way to do this is by sequence similarity. Indeed, if a putative protein encoded by an uncharacterized ORF shows statistically significant similarity to another protein of known function, this simultaneously proves beyond reasonable doubt that the ORF in question is a *bona fide* new gene and predicts its likely function [19]. Even if the homolog of the new protein has not been characterized, useful information is produced in the form of conserved motifs that may be important for protein function. The methods of choice for the initial database screening are those, such as BLASTX [20], that translate the query nucleotide sequence in all six reading frames and compare the resulting putative protein sequences to the protein sequence database. Such methods allow the detection of frameshift errors and will not miss even small ORFs if homologs are present in the database.

Systematic use of BLASTX has resulted in the discovery of a number of new bacterial genes in sequences that have been deposited in databases but not fully annotated [21–23]. A problem with this approach, however, is that a sizable fraction of bacterial gene products are not similar to any known proteins. Sequence analysis methods that distinguish between coding and non-coding regions in DNA on the basis of their different statistical properties are therefore indispensable for gene identification. A variety of such methods have been developed [24,25]. Lately, the non-homogeneous Markov models using in-phase hexamer statistics [19,21], and hidden Markov models [22], have proved particularly effective in bacterial gene prediction.

Many of the genes originally predicted by these statistical methods have subsequently proved to be homologous to newly described genes or have been confirmed experimentally, thus supporting the robustness of the prediction methods [26]. Eventually, with the accumulation of new sequences, sequence conservation will become the definitive criterion for gene identification, whereas the contribution of statistical methods will decrease. Nevertheless, it is still likely that some genes will not have identifiable homologs, and statistical and experimental approaches will remain necessary for their detection. Furthermore, even for genes that have homologs, statistical methods of coding-potential analysis will remain useful for localizing frameshifts and choosing among the possible initiation codons.

Functional prediction

A crucial question for the whole-genome sequencing enterprise is: how informative are the sequences? In other words, when the complete genome sequence is available, for what fraction of the gene products will it be possible to reveal evolutionary relationships and predict functions? Fortunately, it turns out that most of the bacterial proteins are highly, or at least moderately, conserved in evolution. The analysis of the sequenced portion of the *Mycoplasma capricolum* genome has revealed significant similarity to proteins in databases for 75 % of the putative gene products [27]. An even higher fraction (85 %) of proteins were found to have statistically significant database matches in our recent analysis of the *E. coli* genome sequence (75 % complete) [28].

Table 2

Sequence conservation in *E. coli* proteins*.

Similarity level	Best ‘hit’ in database	Best BCR ‘hit’†	Best ACR ‘hit’
Highly significant ($p < 10^{-3}$)	2172 (72 %)	1351 (45 %)	833 (28 %)
Twilight zone‡ ($p > 10^{-3}$)	392 (13 %)	462 (15 %)	468 (16 %)
No detectable similarity	446 (15 %)	1197 (40 %)	1709 (56 %)

* The table was constructed from the output of the BLATAX program, which classified the database search results by the taxonomic origin on the ‘hits’ [18].
† Distantly related bacteria were defined as those outside the proteobacteria [79].
‡ The relevance of the ‘hits’ in this category was additionally assessed using motif search and multiple alignment methods [18].

This high level of sequence conservation is not due to trivial similarity to homologs from closely related bacterial species, as about two-thirds of the *E. coli* proteins contain regions conserved at least at the level of distantly related bacteria — ‘bacterial conserved regions’ or BCRs — and over 40 % contain regions shared with eukaryotic or archaeal homologs — ‘ancient conserved regions’ or ACRs. Most of these sequence similarities are detectable with standard database-searching methods, such as BLASTP [29,30]. Nevertheless, additional approaches to similarity analysis, including methods for identifying motifs, produce a significant increase in sensitivity. The contribution of these methods is particularly important for the identification of ACRs (Table 2). Even for proteins with closely related homologs, database screening with conserved motifs frequently provides additional connections to functionally well characterized proteins, although there is always a trade-off between the level of similarity and the precision of functional prediction.

E. coli genes have been studied in great detail, and functional information is available for ~60 % of them [31]. Nevertheless, the remaining 40 % of the available *E. coli* proteins provided a large enough sample — more than 1000 proteins — to assess our ability to predict functions of bacterial proteins from sequences alone. As the functions of these uncharacterized proteins are experimentally determined, the accuracy of homology-based predictions can be critically evaluated. Using database-search methods, such as BLAST and FASTA, motif analysis and multiple alignment methods, we predicted, at least in general terms, the functions of about half of the uncharacterized proteins [18,28]. For *M. capricolum*, with almost no information on protein functions available, the level of functional prediction for 287 putative proteins reached 75 % [27]. Taken together, these studies on two distantly related bacteria prove a crucial point: bacterial genome sequencing will provide a wealth of information on phylogenetic relationships and gene functions; there is no concern that the sequences remain useless strings of letters.

Paralog clusters

It has been long known that some bacterial genes are related to other genes of the same organism ([32–34] and references therein). In other words, they are intraspecies homologs, or paralogs, as opposed to orthologs, which are genes in different organisms related by vertical descent [35]. With most of the *E. coli* genome sequence now available, it is possible to evaluate the actual extent of paralogy in bacteria. We found that about 50 % of *E. coli* genes form clusters of paralogs, defined on the basis of significant pairwise similarity [18,28]; using a different method for sequence comparison, other workers have arrived at similar conclusions [33,34].

Most of the paralog clusters are small, with only two to four members, but there are several large clusters, which typically encode transport and regulatory proteins [28]. The largest cluster, with a projected membership of about 100 genes in the complete *E. coli* genome, includes genes for membrane ATPases involved in active transport of various metabolites. The analysis of paralogous relationships is an important aspect of bacterial genome studies, as evolution by gene duplication is likely to provide the basis of adaptability to diverse and changing environments. Moreover, variation in the extent of paralogy may be one of the major factors accounting for the large differences in bacterial genome size.

Genome comparisons

The availability of the first two complete bacterial genome sequences [1,2] has put to test our ideas of bacterial genome organization, as well as the utility of our approaches and methods for genome analysis. The original paper by the TIGR team [1] includes an analysis of the sequences of all putative gene products, carried out using a variety of computer methods. Special attention was paid to the functional classification of putative proteins according to the categories introduced by Riley [31]. Nevertheless, this analysis was not complete — and was not intended to be. As the authors appropriately noted [1], genome analysis is an on-going process. The application of additional, more sensitive analytical tools, the careful examination of relatively weak sequence similarities, and the accumulation of new sequences in the databases adds new dimensions to the analysis of the newly available complete genome sequences.

Sequence conservation statistics and functional prediction

The most obvious amendments to the initial analysis come in the area of functional prediction. In their published analysis, the TIGR team [1] did not attempt to predict functions of those putative proteins for which the closest relative is an uncharacterized ORF product. This conservative approach allowed functional predictions to be made for 58 % of *H. influenzae* proteins and 68 % of *M. genitalium* proteins. Clearly, more functional predictions can be made if additional, weaker but still statistically highly significant, similarities are taken into account. Such an effort has been undertaken using the GeneQuiz system [11], yielding functional predictions for another 8 % of *H. influenzae* proteins.

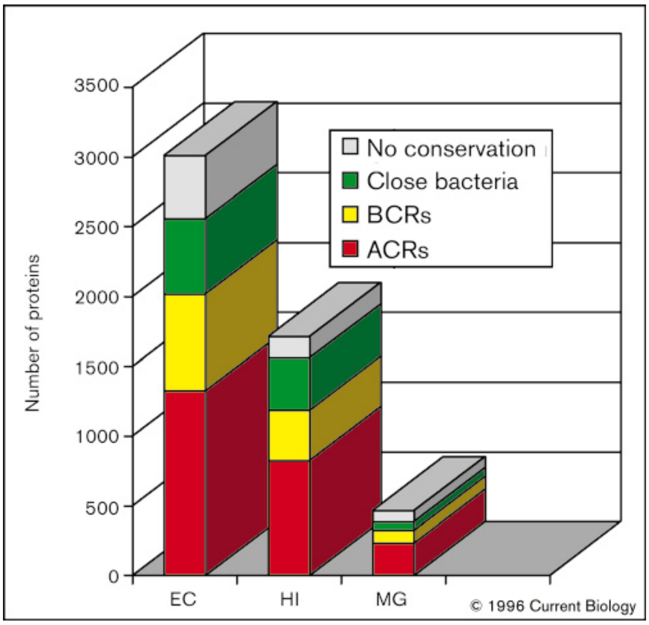
We re-analyzed the sequences of *H. influenzae* and *M. genitalium* using the well characterized set of *E. coli* genes as a reference. The comparison between *H. influenzae* and *E. coli* is addressed in detail elsewhere [17]; here we shall concentrate on the initial conclusions that emerge from the three-way comparison. It has to be emphasized that only a very small fraction of *H. influenzae* and *M. genitalum* proteins have been characterized experimentally, so this analysis is a test of our ability to deduce biology from sequence. In the course of these studies, we had to change a sizable fraction of the original functional assignments — about 10 % for *M. genitalum*, for example — based on the results of our detailed sequence-similarity searches.

The *H. influenzae* and *M. genitalium* protein sequence sets were compared to the non-redundant amino-acid sequence database held at the National Center for Biotechnology Information, using the strategy that has been previously applied to the *E. coli* proteins [18,28]. The results reveal an important and unexpected feature of bacterial gene ensembles: the fractions of proteins containing ACRs and BCRs are very similar for the three bacterial genomes, despite the huge differences between the numbers of proteins they encode (Fig. 2). It seems that the fraction of ACR-containing proteins — about 50 % — may be an important constant in bacterial evolution (even though corroboration from analysis of other bacterial genomes is necessary).

These observation refute one of the possible interpretations of the genome-size reduction — that highly conserved, ‘house-keeping’ genes have been maintained, whereas more variable, ‘luxury’ genes have been lost in the course of the evolution of small genomes. Apparently, bacteria do not adhere to this logic — they maintain the balance between highly conserved and more variable genes even while dramatic changes in genome size are taking place. This may be rationalized as reflecting an equilibrium between the stability of the principal physiological processes and the requirements for environmental adaptability.

In a similar vein, the level of functional prediction is no higher (in fact, it is somewhat lower) for the tiny

Figure 2

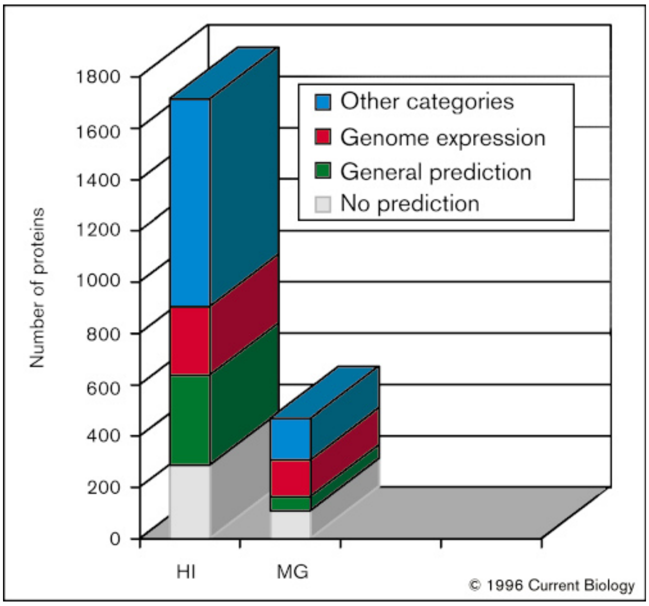


Protein sequence conservation in *E. coli*, *H. influenzae*, and *M. genitalium*. The histogram is organized hierarchically, with BCRs indicating proteins that have detectable homologs in distantly related bacteria but not in eukaryotes or Archaea, and 'close bacteria' indicating proteins that do not have detectable homologs in distantly related bacteria.

M. genitalium genome than it is for the three-times larger *H. influenzae* genome. The extent of functional prediction was significantly increased in our analysis by exploring relatively weak similarities with methods for motif analysis; nevertheless, over 20 % of the *M. genitalium* proteins remain without a predicted function (Fig. 3). The relatively large fraction of *Mycoplasma* proteins with sequences that did not show significant similarity to proteins from other organisms has even been pronounced a measure of our ignorance about the workings of a bacterial cell [36].

This ignorance may, however, not be as dramatic as it seems at first glance. Analysis of the sequences of the 'enigmatic' *M. genitalium* proteins with statistical methods that distinguish between globular and non-globular regions [37,38] shows that most contain large non-globular domains. Many of these appear to have a coiled-coil structure [39]. Adhesins — proteins involved in the adhesion of *M. genitalium* and other bacteria to host cells — have this type of structure ([40] and E.V.K., unpublished observations), so it seems likely that at least some of the uncharacterized *M. genitalium* proteins may also be involved in the interaction of the bacterium with host cells. If this is correct, it may turn out that the small parasitic bacterium dedicates about 20 % of its coding capacity to these anchorage devices. Elucidating the precise functions of these proteins is a challenge for experimentalists.

Figure 3



Predicted fuctions of *H. influenzae* and *M. genitalium* proteins. 'General prediction' indicates proteins for which only an enzymatic or other activity, but not the actual function, could be predicted; 'genome expression' indicates proteins predicted to be involved in translation, transcription, DNA replication and DNA repair.

Whatever these activities might be, they do seem to be a bargain for the bacterium, as having them allows it to shed almost everything else, as we discuss below.

Reduction of the gene repertoire in parasitic bacteria

H. influenzae and *M. genitalium* are both parasitic bacteria that can be cultivated only on rich media [41,42]. It is common knowledge in biology that many parasites have a grossly simplified organization. The degree of morphological simplification in different parasites varies greatly and reaches extremes in some parasitic helminths, which are essentially bags with reproductive organs and very few other physiological systems [43]. Of course, such parasites are well equipped with devices allowing them to extract everything they need from the host. The parasitic bacteria turn out to be not much of an exception. As fittingly noted by the TIGR team [1], it may be no less revealing to know what these bacteria do not have than what they do have. The availability of both the 1.83 Mb *H. influenzae* and 0.58 Mb *M. genitalium* genome sequences is particularly valuable, as they represent different levels of parasitism, with *M. genitalium* apparently being the paradigm of a 'minimal' bacterial genome [2].

The functional category that is most highly conserved includes components of the translation machinery. The great majority of *E. coli* proteins involved in translation are represented by orthologs in both *H. influenzae* and *M. genitalium*. Two notable exceptions are glutamine aminoacyl-tRNA synthetase and the ribosomal protein S1, which are missing in *M. genitalium*. As discussed by Fraser *et al.* [2], it is most likely that glutaminyl-tRNA is formed after the aminoacylation step in *M. genitalium*, as it is in other Gram-positive bacteria [44]. The absence of S1, which contains a highly conserved RNA-binding domain found also in eukaryotes and archaea [45,46], is unexpected. This is an example of a domain that previously appeared to be ubiquitous, but turns out not to be essential for cell function.

Predictably, the principal components of the replication and transcription systems are also conserved in *H. influenzae* and *M. genitalium*. There are at least two conspicuous omissions in *M. genitalium*, however, namely the genes for RNase H and transcription-termination factor Rho [2]. The most likely candidate for being the enzyme that removes RNA primers during DNA replication in *M. genitalium* is the MG262 protein, a predicted 5′–3′ exonuclease that is homologous to the exonuclease domain of the *E. coli* and *H. influenzae* DNA polymerase I. The absence of Rho suggests that the transcription of all genes in *M. genitalium* terminates by Rho-independent mechanisms [2], even though other transcription factors — NusA [47], for example, which is encoded in the *M. genitalium* genome — may be involved in termination.

The third class of proteins that are typically conserved even in small genomes includes the molecular chaperones and chaperone-like proteins that are involved in the folding of other proteins and the assembly of macromolecular complexes. Representatives of all the families of molecular chaperones found in *E. coli* are present in *H. influenzae*, and most are also represented in the *M. genitalium* genome. Notable exceptions include heat-shock 90 family proteins and two of the three families of peptidyl-prolyl isomerases, which are missing in *M. genitalium*.

The representation of all other functional categories of proteins is dramatically reduced in *M. genitalium*. In *H. influenzae*, by contrast, proteins involved in such house-keeping functions as DNA repair and nucleotide biosynthesis are largely preserved. The DNA repair systems in *E. coli*, *H. influenzae* and *M. genitalium* are compared in Table 3. In spite of its smaller genome, *H. influenzae* retains all the major repair mechanisms identified in *E. coli*, with the exception of the UmuDC system, the very-short-patch repair (Vsr/Dcm) system, and some poorly characterized nucleases. *M. genitalium* has lost most of the repair systems, though, interestingly, repair genes occupy roughly the same fraction of the *M. genitalium* and *H. influenzae* genomes. The minimal repair capacity that *M. genitalium* has apparently retained is provided by the Uvr excinuclease complex, the oxidative-damage repair protein MutM (missed in [2]), the ortholog of the *E. coli* DinP protein, and a small repertoire of nucleases.

One aspect of the rudimentary DNA repair systems of *M. genitalium* is quite surprising. The *M. genitalium* genome encodes two DNA-dependent DNA polymerases, both of which belong to the DNA polymerase III family. One of these proteins, MG031, is the ortholog of Gram-positive bacterial DNA polymerase III, whereas the other, MG261, appears to be orthologous to Gram-negative bacterial DNA polymerase III [48]. The Gram-negative enzyme may be responsible for repair DNA synthesis, as it appears to form an operon with the two repair genes that encode the putative 5′–3′ exonuclease mentioned above, and MutM ([48] and Table 3). In *H. influenzae*, the most likely candidate for the repair polymerase is DNA polymerase I, whereas the counterpart of the other repair polymerase of *E. coli*, DNA polymerase II, is missing. The drastic reduction in DNA repair capability in *M. genitalium* is likely to result in a relatively high replication error rate, but the mutation rate per genome may still be similar to that of bacteria with larger genomes.

A dramatic aspect of the gene repertoire reduction in *M. genitalium* is the virtual absence of proteins that in other bacteria are involved in the regulation of gene expression. Specifically, the helix–turn–helix DNA-binding domain, one of the most widespread of all protein domains in both *E. coli* [28] and *H. influenzae* [17], was found in only one *M. genitalium* protein, namely the σ subunit of the RNA polymerase. The class of helix–turn–helix proteins includes both proteins, such as the classical Lac repressor, that regulate specific operons, and those, such as the LexA repressor and catabolite gene activator protein (CAP), that affect the expression of large gene classes. The absence of CAP correlates with the absence of adenylate cyclase, and suggests that cyclic AMP, a regulatory molecule previously thought to be ubiquitous, has no role in *M. genitalium*.

M. genitalium also lacks ‘two-component’ regulatory systems, which consist of histidine kinase sensor domains and response regulator domains and are widely represented in *E. coli* and *H. influenzae* [2]. The conspicuous absence of these regulatory systems suggests that the principles of gene-expression regulation in *M. genitalium* may be very different from those in bacteria with larger genomes. The regulatory circuits are expected to be much less differentiated and less responsive to environmental signals. Guanosine tetraphosphate (ppGpp), an alarmone synthesized with the participation of the SpoT protein [49,50], an ortholog of which is encoded in the *M. genitalium* genome (E.V.K., unpublished observations), may be important for global transcriptional regulation.

A predominant role in gene-expression regulation in *M. genitalium* is likely to be played by *cis*-acting signals, such as promoters, ribosomal-binding sites of different strength, and mRNA stability determinants. It can be imagined that, in *M. genitalium*, there are several classes of

Table 3

Genes involved in DNA repair in *E. coli*, *H. influenzae* and *M. genitalium.**

E.coli genes	Presence in		Enzymatic or other known activity
	H. influenzae	*M. genitalium*	
Photoreactivation			
phrA	–	–	Photolyase?
phrB	–	–	Photolyase
Removal or repair of modified nucleotides			
ada	–	–	O-6-methylguanine DNA methyltransferase
alkA	–	–	3-methyladenine DNA glycosidase
alkB	–	–	?
dut	+	–	dUTPase
mutM	+	+	Formamidopyrimidine DNA glycosylase
mutT	+	–	8-oxo-dGTPase
mutY	+	–	A•G-specific adenine glycosylase
nfo	–	+	Endonuclease IV
ogt	+	–	O-6-methylguanine DNA methyltransferase
tag	+	–	DNA-3-methyladenine glycosidase I
ung	+	+	Uracil-DNA glycosylase
uvrA	+	+	Excinuclease subunit, DNA-binding, ATPase
uvrB	+	+	Excinuclease subunit, helicase
uvrC	+	+	Excinuclease subunit, nuclease
Mismatch repair			
dam	+	–	A-specific DNA methylase
dcm	–	–	C-specific DNA methylase
mutH	+	–	Endonuclease
mutL	+	–	?
mutS	+	–	ATPase
uvrD	+	–	Helicase
vsr	–	–	Endonuclease
Recombinational and strand-specific repair			
dnaE †	+	+	DNA polymerase III
mfd	+	–	Helicase
polA (polymerase domain)	+	–	DNA polymerase I
polA (exonuclease domain)	+	+	5′–3′ exonuclease
recA	+	+	ATPase, DNA strand exchange
recB	+	–	Exonuclease V subunit, helicase
recC	+	–	Exonuclease V subunit
recD	+	–	Exonuclease V subunit, ATP-binding
recG	+	–	Helicase
ruvA	+	–	Helicase subunit
ruvB	+	–	Helicase subunit, ATPase
ruvC	+	–	Endonuclease
uvrA	+	+	Excinuclease subunit, ATPase
uvrB	+	+	Excinuclease subunit, helicase
uvrC	+	+	Excinuclease subunit, nuclease
uvrD	+	–	Helicase
SOS repair			
dinG	+	–	Helicase
dinP	–	+	?
exA	+	–	Transcriptional regulator, autoprotease
polB	–	–	DNA polymerase
recA	+	+	ATPase, DNA strand exchange
recF	+	–	ATPase
recN	+	–	ATPase
recO	+	–	?
recQ	+	–	Helicase
recR	+	–	?
ruvA	+	–	Helicase subunit
ruvB	+	–	Helicase subunit, ATPase
ruvC	+	–	Endonuclease
umuC	–	–	?
umuD	+	–	Autoprotease
uvrA	+	+	Excinuclease subunit, DNA-binding, ATPase
uvrB	+	+	Excinuclease subunit, helicase
uvrC	+	+	Excinuclease subunit, nuclease
uvrD	+	–	Helicase

* The table includes all identifiable repair genes of *M. genitalium*; *E. coli* has 11 and *H. influenzae* 8 additional, poorly characterized genes.
† Primarily a replicative enzyme in *E.coli* and *H. influenzae*; probably involved in DNA repair in *M. genitalium* (see text).

differentially expressed genes, and that genes encoding proteins in different functional categories are expressed at different levels. This resembles genome-expression regulation in large DNA viruses, such as poxviruses and herpesviruses, rather than the classical bacterial regulation. These viruses have a small number of gene classes that are expressed at different times during infection, under the control of a small number of transcription factors that interact with distinct *cis*-elements [51,52]. Gene expression in *M. genitalium* may follow a similar pattern. Gene-expression regulation in *M. genitalium* may also involve the modulation of transcription initiation by differential superhelicity, and the modulation of translational elongation rates by codon usage.

H. influenzae has clearly preserved more conventional modes of gene-expression regulation, despite having markedly fewer genes than *E. coli*. Furthermore, analysis of the *M. capricolum* genome, the estimated size of which is only about 700 kb, has revealed genes encoding several helix–turn–helix proteins [27], suggesting that even this bacterium with a small genome is likely to have conventional regulatory systems. Sequencing additional small bacterial genomes, such as other *Mycoplasma* or *Chlamydia*, should show whether there is a complexity threshold, below which a genome is stripped of regulatory genes, or whether *M. genitalium* is an anomaly.

M. genitalium appears to have a minimal metabolism. Its intermediate metabolism *sensu strictu* is virtually limited to glycolysis. Also maintained are salvage pathways of nucleotide biosynthesis, and pathways of lipid biosynthesis using exogenous fatty acids. Other biosynthetic pathways, with a few exceptions, are missing, and, accordingly, all amino acids, sugars and coenzyme components have to be imported into the *M. genitalium* cell. To do so, *M. genitalium* uses seventeen predicted transport ATPases and about twenty permeases; it also has a phosphotransferase system for glucose import [2]. This limited repertoire of transport systems suggests that some of the *M. genitalium* transporters are likely to have a low specificity, and that new, unknown transport mechanisms may be involved.

H. influenzae, in contrast, retains the principal metabolic pathways known to exist in *E. coli*, even though their regulation seems in a number of cases to be simplified, and the elimination of a few biosynthetic enzymes renders *H. influenzae* dependent on a rich growth medium. The major exceptions are the missing tricarboxylic acid (TCA) cycle, which appears to be replaced by a simplified biosynthetic pathway, the missing glyoxylate cycle, and several missing respiratory chains [1,17].

An important aspect of the gene repertoire reduction in both *H. influenzae* and *M. genitalium* is a reduced extent of gene paralogy. We found that only 35 % and 25 %, respectively, of the *H. influenzae* and *M. genitalium* proteins belong to clusters of paralogs, compared to nearly half of the *E. coli* proteins. A similar fraction of paralogs has been reported for *H. influenzae* by other workers using different methods for sequence comparison and clustering [53]. In part, the relatively small level of paralogy in *H. influenzae* and *M. genitalium* may result from the elimination of entire functional systems (such as those for sugar utilization). But there are many cases where *E. coli* has two enzymes that catalyze the same metabolic reaction but operate under different conditions and/or are differently regulated and *H. influenzae* has only one [17]. Interestingly, the *M. genitalium* genome, which has the lowest level of gene paralogy, has a few gene duplications not found in *H. influenzae* or (so far) in *E. coli* — examples include two genes (MG010 and MG240) encoding putative DNA primases, and two genes (MG011 and MG012) encoding homologs of ribosomal protein S6 modification enzyme. Given the trend towards genome contraction, it seems likely that such duplicated proteins have indispensable functions in *M. genitalium.*

Extensive gene shuffling

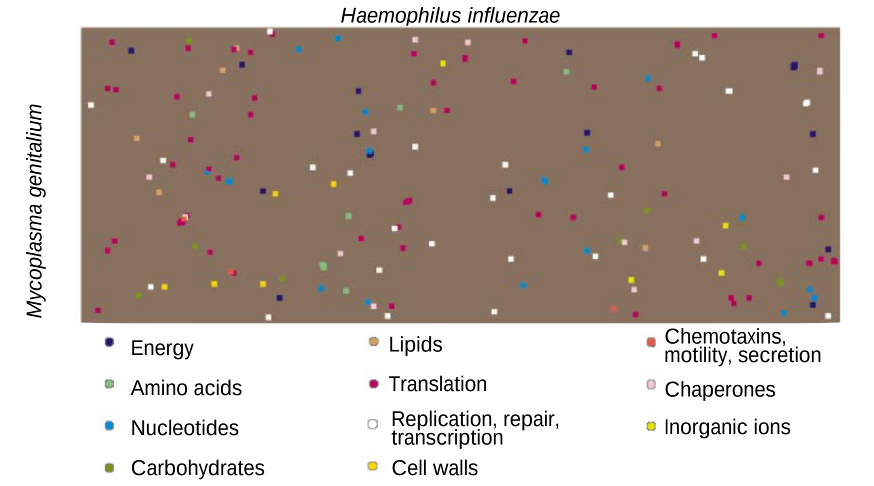
A comparison of the arrangement of orthologous genes in the *E. coli*, *H. influenzae* and *M. genitalium* chromosomes revealed no long-range colinearity (Fig. 4), suggesting that extensive gene shuffling has occurred during bacterial genome evolution. Closer examination, however, shows that, in *E. coli* and *H. influenzae*, about 70 % of orthologous genes belong to short conserved arrays, about half of which are known to be operons in *E. coli* [17]. In contrast, only a few essential operons remained intact throughout the enormous evolutionary span separating *M. genitalium* from *E. coli* and *H. influenzae*. The most prominent of these are the ribosomal protein superoperon, with twenty five genes in the same order, and the proton ATPase operon, with a conserved array of six genes. Also notable is the partial conservation of a gene array around the origin of replication, which has already been described in a wide variety of bacteria [54].

Predicting functions of eukaryotic proteins

As discussed above, about half of all bacterial proteins have eukaryotic or archaeal homologs. In many cases, the functions of these eukaryotic proteins are not known, but it may be possible to predict them if their bacterial homologs have been functionally characterized. In many cases, the predictions could probably be made by analysing sequences of eukaryotic proteins themselves, but bacterial genome analysis is a more systematic approach. Furthermore, many eukaryotic proteins contain regions with compositionally-biased sequences, and this may make it harder to detect functionally important motifs [37,38]. In the course of our analysis of *E. coli* protein sequences, we have made functional predictions for a number of eukaryotic proteins, some of which are

Figure 4

Lack of large-scale colinearity between the *H. influenzae* and *M. genitalium* genomes. For each genome, the replication origin was chosen as the zero point. The axes represent the complete chromosomes in the clockwise direction. Each point represents a pair of orthologous genes with the respective coordinates in the *H. influenzae* and *M. genitalium* chromosomes. The functional categories of proteins encoded by the respective genes are color-coded as indicated.



associated with human diseases ([28] and E.V.K., unpublished observations, available in part by anonymous FTP at ncbi.nlm.nih.gov/repository/Eco/EcoProt). Examples that have been explored in detail include the predictions that translation elongation factor EF-1γ has glutathione S-transferase activity [55], that translation elongation factor EF-2B has nucleotidyltransferase activity [56], and that human tumor marker P120 [57] and fibrillarins [28] have rRNA methyltransferase activity.

A new example of such a functional prediction involves cytoskeletal proteins known as adducins. Database searches using the sequence of the *E. coli* FucA protein (fuculose-5-phosphate aldolase) identified three proteins with moderately similar sequences: two were other *E. coli* proteins — AraD (l-ribulose-5-phosphate epimerase) and YiaS (an uncharacterized protein closely related to AraD) — and the third was adducin. Subsequent motif searches [58] showed that the most conserved part of the alignment corresponds to a motif that has been previously recognized in isopropylmalate and homocitrate synthases [59]. This motif is, in fact, conserved in a much wider range of lyases and epimerases (Fig. 5), in which it is likely to comprise part of the active center. We predict that adducins contain an active lyase domain that may be impaired in the *Drosophila* homolog (Fig. 5). Adducins are large, heterodimeric cytoskeletal proteins that promote organization of the spectrin–actin lattice in a calmodulin-dependent fashion [60]. Mutations in adducin genes have been implicated in hereditary hypertension in rats [61], and very recently also in humans [62]. The presence of the predicted lyase domain may suggest a new, uncharacterized function for these cytoskeletal proteins; identification of this function will be important for understanding the possible role of adducin defects in disease.

From computer analysis to experimentation

Computer comparisons of genome sequences produce conclusions that are important in their own right on, for example, genome organization conservation and probable evolutionary events. But the most important outcome of these analyses may be their utility for interpreting experimental results and directing new experimentation. Genome sequence analyses are likely to be used by experimenters in two conceptually different ways. The first way is the testing of computer predictions for specific gene products. There is certainly nothing ‘genome-specific’ about this strategy. There are numerous examples of such studies based on individual gene sequences. With complete genome sequences becoming available, computer predictions are important for prioritizing experiments. For example, for researchers working on *M. genitalium*, the predicted roles of one of its two DNA polymerases III in DNA repair, and of its DNA-polymerase-I-related exonuclease in primer removal, may have a high priority.

The second way that experimenters are likely to use genome sequence analyses is to guide more global studies. Typical examples of such studies are the global analysis of gene transcription in *E. coli* under various physiological conditions [63], and systematic expression analysis of *E. coli* genes using two-dimensional electrophoresis of the synthesized proteins [64,65]. Combined with sequence information, the latter approach is capable of producing a definitive expression map of a bacterial genome. As the *M. genitalium* genome includes only 470 protein-coding genes, the two experimental approaches may converge, as testing the functions of all gene products, for which predictions are available, seems feasible in this case.

One of the principal genome-oriented experimental approaches involves the inactivation of specific genes,

Figure 5

```
HTS_DROME  151   EYFLVNPYGLLYHEITASALNKVDM 16
ADDL_RAT   165   DHFLISPKGVSCSEVTASSLIKVNI 16
ADDB_HUMAN 165   DHFLISPKGVSCSEVTASSLIKVNI 16
ADDA_HUMAN 177   EHFLIVPFGLLYSEVTASSLVKINL 16
                    ** * *: *  :* * :* *:
FUCA_ECOLI  35   DGMLITPTGIPYEKLTESHIVFIDG 11
ARAD_ECOLI  36   GVFVIKPSGVDYSVMTADDMVVVSI 13
YIAS_ECOLI  36   QWMVIKPSGVEYDVMTADDMVVVEI 13
Consensus        ..UUU.P.GU....UT...UU.U..

HTS_DROME   SHFVLHSVVHAARPDIRCAIYIGCSPVVAISSLKTGLLPLTKD-ACVLGEITTHAYTGLF 10
ADDL_RAT    TGFSLHSAIYAARPDVRCAIHLHTPATAAVSAMKCGLLPVSHN-ALLVGDMAYYDFNGEM 11
ADDB_HUMAN  TGFCLHSAIYAARPDVRCIIHLHTPATAAVSAMKWGLLPVSHN-ALLVGDMAYYDFNGEM 11
ADDA_HUMAN  AGFTLHSAIYAARPDVKCVVHIHTPAGAAVSAMKCGLLPISPE-ALSLGEVAYHDYHGIL 11
            : :  * * *  ***   *** *     *** :  ::  *    * : *:
FUCA_ECOLI  SEWRFHMAAYQSRPDANAVVHNHAVHCTAVSILNRSIPAIHYMIAAAGGNSIPCAPYATF 10
ARAD_ECOLI  SDTPTHRLLYQAFPSIGGIVHTHSRHATIWAQAGQSIPATGTTHADYFYGTIPCTRKMTD 23
YIAS_ECOLI  SDTPTHLALYRRYAEIGGIVHTHSRHATIWSQAGLDLPAWGTTHADYFYGAIPCTRQMTA 23
Consensus   .....H..UY...P.....UH.H......US.....U.......A..U...........

     RHAD_ECOLI       137: DRVIMHCHATNLIALTYVLENDTAVFT
     NIFV_RHOCA       191: LPIEMHAHNDFGMATANTIMAAHAGAT
     NIFV_KLEPN       188: GEIEMHAHNDLGMATANTLAAVSAGAT
     LEU1_LACLA       196: IIFSPHCHDDLGMAVANSLAAIKAGAG
     NIFV_RHOSH       203: LPVEFHGHNDLGMATANSLAAARAGAS
     NIFV_AZOVI       188: MELEVHAHDDFGLATANTLAAVMGGAT

     LEU1_ECOLI       197: AIISVHTHDDLGLAVGNSLAAVHAGAR
     NIVA_CLOPA       187: IDIEIHVHNDFGMAISNSFAAFKAGAK
     PYC_MOUSE        766: LPLHIHTHDTSGAGVAAMLACAQAGAD
     LEU1_YEAST       270: VCISTHCHNDRGCGVAATELGMLAGAD
     DCOA_KLEPN       196: VTLHLHCHATTGMAEMALLKAIEAGVD
     HMGL_CHICK       201: GALAVHCHDTYGQALANILVALQMGVS
     HMGL_PSEMV       199: AALAGHFHDTWGMAIANVHAALAQGVR
                                ! !

HTS_DROME     SLGPNSKVILLTNHGALCCGETIEEAFFAACHIVQACETQLKLLPVGLDNLVL 842
ADDL_RAT      CLGPTCKILVLRNHGMVALGDTVEEAFYKVFHLQAACEVQVSALSSAGGTENL 233
ADDB_HUMAN    CLGPTCKILVLRNHGVVALGDTVEEAFYKIFHLQAACEIQVSALSSAGGVENL 397
ADDA_HUMAN    NLGPKSKVLILRNHGLVSVGESVEEAFYYIHNLVVACEIQVRTLASAGGPDNL 449
                    *  :*  ***::    :* * :  *:: *  :: : ***       *
FUCA_ECOLI    LALKNRKATLLQHHGLIACEVNLEKALWLAHEVEVLAQLYLTTLAITDPVPVL 21
ARAD_ECOLI    IDAAQMPGVLVHSHGPFAWGKNAEDAVHNAIVLEEVAYMGIFCRQLAPQLPDM 21
YIAS_ECOLI    RSPAQIPAVLVHSHGPFAWGKNAADAVHNAVVLEECAYMGLFSRQLAPQLPAM 21
Consensus     ........UUU..HG.U......E.AU....U.......U..........U
```

Adducins – eukaryotic cytoskeletal proteins– contain a lyase domain. The alignment between the amino-acid sequences of mammalian adducins and their *Drosophila* homolog (HTS) with the sequences of three *E. coli* proteins – FucA (L-fuculose phosphate aldolase), AraD (L-ribulose-5-phosphate 4-epimerase) and YiaS (uncharacterized) – was constructed using the MACAW program [80]. The asterisks show identities and the colons similar residues in FucA and human adducin α subunit (ADDA); the numbers show the distances from the protein termini and the distances between the aligned segments. The consensus line shows amino acid residues conserved in the aligned sequences, with one exception allowed; U indicates a bulky hydrophobic residue. An additional alignment block, which has been selected from the database by a motif search using the MoST program [58], includes sequences of *E. coli* rhamnulose-1-phosphate aldolase (RhaD) and various lyases. The two histidine residues that are conserved throughout the alignment, except for HTS, and that are predicted to be in the lyase active center, are indicated by exclamation marks. The sequences were from the SWISS-PROT database.

followed by evaluation of the effects of gene disruption [66,67]. An ingenious recent study [68] employed this approach to estimate the minimal size of a bacterial genome that is still compatible with reproduction. It has been shown that out of 79 randomly selected *Bacillus subtilis* genes, disruption of only six rendered the bacteria non-viable; from this the minimal genome size has been estimated to be 562 kb, remarkably close to the size of the *M. genitalium* genome [68]. The availability of complete genome sequences and functional predictions for most of the genes gives the researchers flexibility in choosing the gene-inactivation ('knockout') strategy — researchers may aim to disrupt all genes one by one, genes in a specific functional category, or individual genes of interest. Furthermore, only with the availability of complete genome sequences does it become possible to knockout all genes in a given cluster of paralogs, in order to assess the importance of their common function.

Now that complete genome sequences are available, it seems appropriate to consider making a comprehensive analysis of the chemical composition of cells growing on defined media, especially the repertoire of small molecules, in order to match it with the predicted gene functions. Such analysis should significantly facilitate the identification of specific metabolic pathways.

Concluding remarks

We believe that the comparative genome analysis that we have presented shows that complete genome sequencing has not merely increased the amount of sequence information available, but rather has led to a paradigm shift in genomics. For the first time, conclusions drawn from genome comparisons can be definitive. This is particularly important for negative statements, such as the absenceof helix–turn–helix proteins in *M. genitalium*, that only make sense when the genome sequence is complete. The results that we have described clearly represent only a preliminary analysis of the newly available complete genome sequences. A number of other important issues can and will be addressed: for example, the deduction of unknown metabolic pathways and regulatory circuits [17], the prediction of operon structure, and the identification of regulatory signals such as promoters, operators and terminators.

Ultimately, one would want to be able to deduce the entire biochemistry and physiology of a cell from its genome sequence alone. This goal may never be reached literally, but it is certainly conceivable that with the accumulation of complete genome sequences, and further development of methods for genome comparison, progressively more precise approximations will be attained. As it is obvious that complete genome sequencing, at least in the foreseeable future, will exceed the ability of researchers to study gene functions, the sequence-based reconstructions are important for focusing experiments on those genes and reactions that will fill the most important gaps in existing knowledge.

One of the greatest intellectual challenges in the area of genomics is to reconstruct, even if hypothetically, the genome organization, and by inference the biochemistry and physiology, of ancestral forms, including the last common ancestor of eukaryotes, archaea and bacteria [69,70]. A distinction should be made between a 'minimal' and an 'ancestral' genome. A 'minimal' genome can be defined as the minimal repertoire of genes compatible with cellular life. The *M. genitalium* genome itself is a big step toward the minimal genome [2], and a further theoretical reduction is possible through genome comparisons. Approximately half of the genes of *M. genitalium* appear to have orthologs in *H. influenzae* (E.V.K. and A.R.M., unpublished observations). Detailed analysis of the proteins encoded by this gene set may indicate how likely it is that an organism may exist with as few as 200 genes. It may be possible to design experiments specifically focused on the discovery of bacteria with such tiny genomes that they might have escaped detection because of their inability to grow outside their host organism.

The genomes of *H. influenzae* and, particularly, *M. genitalum* have been shaped to a large extent by the degenerative evolution that accompanied their adaptation to parasitism. It is uncertain whether the result of this degeneration resembles the hypothetical progenote [69]. It cannot be ruled out that it does, especially if the environment in which the progenote thrived was a 'soup', rich in diverse organic molecules [71]. Even though reconstructions of ancestral genomes will always remain speculative, there is a strong hope that with further accumulation of complete sequences of phylogenetically diverse genomes, we will be able to draw a realistic sketch of this elusive primordial entity.

Acknowledgements

We are grateful to Peer Bork and John Wootton for helpful discussions, to Roman Tatusov for valuable help with sequence comparisons and preparing Figure 4, and to Doug Smith and Satoshi Tabata for communicating their results prior to publication.

References

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Dougherty BA, Merrick JM *et al.*: Whole-genome random sequencing and assembly of ***Haemophilus influenzae*** Rd. *Science* 1995, 269:496–512.
2. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM *et al.*: The minimal gene complement of ***Mycoplasma genitalium***. 1995, *Science* 270:397–403.
3. Daniels D, Plunkett G, Burland V, Blattner FR: Analysis of the ***Escherichia coli*** genome: DNA sequence of the region from 84.5 to 86.5 minutes. *Science* 1992, 257:771–778.
4. Rudd KE: Maps, genes, sequences, and computers. An ***Escherichia coli*** case study. *ASM News* 1993, 59:335–341.
5. Kunisawa T, Nakamura M, Watanabe H, Otsuka J, Tsugita A, Yeh LS, George DG, Barker WC: ***Escherichia coli*** K12 genomic database. *Protein Seq Data Anal* 1990, 3:157–162.
6. Medigue C, Viari A, Henaut A, Danchin A: Colibri: a functional data base for the ***Escherichia coli*** genome. *Microbiol Rev* 1993, 57:623–654.
7. Wahl R, Rice P, Rice CM, Kröger M: ECD – a totally integrated database of ***Escherichia coli*** K-12. *Nucleic Acids Res* 1994, 22:3450–3455.
8. Moszer I, Glaser P, Danchin A: Subtilist: a relational database for the ***B. subtilis*** genome. *Microbiology* 1995, 141:261–268.
9. Perriere G, Gouy M, Gojobori T: NRSub: a non-redundant database for the ***Bacillus subtilis*** genome. *Nucleic Acids Res* 1994, 22:5525–5529.
10. Benson DA, Boguski M, Lipman DJ, Ostell J: GenBank. *Nucleic Acids Res* 1996, 24:1–6.
11. Casari G, Andrade MA, Bork P, Boyle J, Daruvar A, Ouzounis C, Schneider R, Tamames J, Valencia A, Sander C: Challenging times for bioinformatics. *Nature* 1995, 376:647–648.
12. Medigue C, Moszer I, Viari A, Danchin A: Analysis of a ***Bacillus subtilis*** genome fragment using a co-operative computer system prototype. *Gene-Combis* 1995, 165:37–51.
13. Burland V, Plunkett G, Sofia HJ, Daniels DL, Blattner FR: Analysis of the ***Escherichia coli*** genome VI: DNA sequence of the region from 92.8 through 100 minutes. *Nucleic Acids Res* 1995, 23:2105–2119.
14. Kunst F, Vassarotti A, Danchin A: Organization of the European ***Bacillus subtilis*** genome sequencing project. *Microbiology* 1995, 141:249–255.
15. Ogasawara N, Fujita Y, Kobayashi Y, Sadaie Y, Tanaka T, Takahashi H, Yamane K, Yoshikawa H: Systematic sequencing of the ***Bacillus subtilis*** genome: progress report of the Japanese group. *Microbiology* 1995, 141:257–259.
16. Devine KM, Wolfe K: Bacterial genomes: a TIGR in the tank. *Trends Biochem Sci* 1995, 11:429–431.
17. Tatusov RL, MushegianA R, Bork P, Brown NP, Borodovsky M, Hayes WS, Rudd KE, Koonin EV: Metabolism and evolution of ***Haemophilus influenzae*** deduced from a whole genome comparison to ***Escherichia coli.*** *Curr Biol*, 1996, 6:279–291.
18. Koonin EV, Tatusov RL, Rudd KE: Protein sequence comparison at a genome scale. *Meth Enzymol* 1996, 266:295–322.
19. Borodovsky M., Koonin EV, Rudd KE: New genes in old sequence: a strategy for finding genes in the bacterial genome. *Trends Biochem Sci* 1994, 19:309–313.
20. Gish W, States DJ: Identification of protein-coding regions by sequence similarity searches. *Nat Genet* 1993, 3:266–272.
21. Borodovsky M, Rudd KE, Koonin EV: Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res* 1994, 22:4756–4767.
22. Krogh A, Mian IS, Haussler D: A hidden Markov model that finds genes in ***E. coli*** DNA. *Nucleic Acids Res* 1994, 22:4768–4778.
23. Robison K, Gilbert W, Church GM: Large scale bacterial gene discovery by similarity search. *Nat Genet* 1994, 7:205–214.
24. Fickett JW, Tung CS: Assessment of protein-coding measures. *Nucleic Acids Res* 1992, 20:6441–6450.
25. Gelfand MS: Prediction of function in DNA sequence analysis. 1995. *J Comput Biol* 2: 87-115.
26. Borodovsky M., McIninch J, Koonin EV, Rudd KE, Medigue C, Danchin A: Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Res* 1995, 23:3554–3562.
27. Bork P, Ouzounis C, Casari G, Schneider R, Sander C, Dolan M, Gilbert W, Gillevet PM: Exploring the ***Mycoplasma capricolum***

genome: a small bacteruim reveals its physiology. *Mol Microbiol* 1995, 16:955–963.

28. Koonin EV, Tatusov RL, Rudd KE: Sequence similarity analysis of ***Escherichia coli*** proteins: functional and evolutionary implications. *Proc Natl Acad Sci USA* 1995, 92:11921–11925.
29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol.* 1990, 215:403–410.
30. Altschul SF, Boguski MS, Gish W, Wootton JC: Issues in searching molecular sequence databases. *Nat Genet* 1994, 6:119–129.
31. Riley M: Functions of the gene products of ***Escherichia coli.*** 1993, *Microbiol Rev* 57:862–952.
32. Riley M, Anilionis A: Evolution of the bacterial genome**.** 1978, *Annu Rev Microbiol* 32:519–560.
33. Labedan B, Riley M: Widespread protein sequence similarities: origins of ***Escherichia coli*** genes. *J Bacteriol* 1995, 177:1585–1588.
34. Labedan B, Riley M: Gene products of ***Escherichia coli***: sequence comparisons and common ancestries. *Mol Biol Evol* 1995, 12:980–987.
35. Fitch WM: Distinguishing homologous from analogous proteins. *Syst Zool* 1970, 19: 99–106.
36. Goffeau A: Life with 482 genes. *Science* 1995, 270:445–446.
37. Wootton JC, Federhen S: Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* 1993, 17:149-163.
38. Wootton JC: Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* 1994, 18:269–285.
39. Lupas A, Van Dyke M, Stock J: Predicting coiled coils from protein sequences. *Science* 1991, 252:1162–1164.
40. LaPolla RJ, Haron JA, Kelly CG, Taylor WR, Bohart C, Hendricks M, Pyati JP, Graff RT, Ma JK, Lehner T: Sequence and structural analysis of surface protein antigen I/II (SpaA) of ***Streptococcus sobrinus***. *Infect Immun* 1991, 59:2677–2685.
41. Hoiseth SK: The genus ***Haemophilus.*** In *The Prokaryotes: Handbook on the Biology of Bacteria*, 2nd edn. Edited by Balows A. New York: Springer-Verlag; 4:3304–3330.
42. Razin S: Peculiar properties of mycoplasmas: the smallest self-replicating prokaryotes. *FEMS Microbiol Lett* 1992, 79:423–431
43. Barnes RD: *Invertebrate Zoology.* Saunder College Publishers, Philadelphia; 1987.
44. Strauch MA, Zalkin H, Aronson AI: Characterization of the glutamyl-tRNA(Gln)-to-glutaminyl-tRNA(Gln) amidotransferase reaction. *J Bacteriol* 1988, 170:916–920.
45. Regnier P, Grunberg-Manago M, Portier C: Nucleotide sequence of the pnp gene of ***Escherichia coli*** encoding polynucleotide phosphorylase. Homology of the primary structure of the protein with RNA-binding domain of ribosomal protein S1. *J Biol Chem* 1987, 262:63–68.
46. Gribskov M: Translational initiation factors IF-1 and eIF-2 alpha share an RNA-binding motif with prokaryotic ribosomal protein S1 and polynucleotide phosphorylase. *Gene* 1992, 119:107–111.
47. Schmidt MC, Chamberlin MJ: nusA protein of Escherichia coli is an efficient transcription termination factor for certain terminator sites. *J Mol. Biol* 1987, 195:809–818.
48. Koonin EV, Bork P: Ancient duplication of DNA polymerase inferred from analysis of complete bacterial genomes. *Trend Biochem Sci,* 1996
49. Cashel M, Rudd KE: The stringent response. In *Escherichia coli and Salmonella typhimurium: cellular and molecular biology.* Edited by Neidhardt FC, Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE. Washington: American Society for Microbiology; 1987:1410–1438.
50. Condon C, Squires C, Squires CL: Control of rRNA transcription in ***Escherichia coli***. *Microbiol Rev* 1995, 59:623–645.
51. Moss B: Regulation of vaccinia virus transcription. *Annu Rev Biochem* 1990, 59:661–688.
52. Nevins JR: Transcriptional activation by viral regulatory proteins. *Trends Biochem Sci* 1991, 16:435–439.
53. Brenner SE, Hubbard T, Murzin A, Chothia C: Gene duplications in ***H. influenzae***. *Nature* 1995, 378:140.
54. Cole ST, Saint Girons I: Bacterial genomics. *FEMS Microbiol Rev* 1994, 14:139–160.
55. Koonin EV, Mushegian AR, Tatusov RL, Altschul SF, Bryant SH, Bork P, Valencia A: Eukaryotic translation elongation factor 1g contains a glutathione transferase domain — study of a diverse, ancient protein superfamily using motif search and structural modeling. *Protein Sci* 1994, 3:2045–2054.
56. Koonin EV: Multidomain organization of eukaryotic guanine nucleotide exchange translation initiation factor eIF-2B subunits revealed by analysis of conserved sequence motifs. *Protein Sci* 1995, 4:1608–1617.
57. Koonin EV: Prediction of an rRNA methyltransferase domain in human tumor-specific nucleolar protein P120. *Nucleic Acids Res* 1994, 22:2476–2478.
58. Tatusov RL, Altschul SF, Koonin EV: Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci USA* 1994, 91:12091–12095.
59. Wang S-Z, Dean DR, Che JS, Johnson JL: The N-terminal and C-terminal portions of NifV are encoded by two different genes in ***Clostridium pasteurianum***. *J Bacteriol* 1991, 173:3041–3046.
60. Gardner K, Bennett V: Modulation of spectrin-actin assembly by erythrocyte adducin. *Nature* 1987, 328:359–362.
61. Bianchi G, Tripodi G, Casari G, Salardi S, Barber BR, Garcia R, Leoni P, Torielli L, Cusi D, Ferrandi M *et al*: Two point mutations within the adducin genes are involved in blood pressure variation. *Proc Natl Acad Sci USA* 1994, 91:3999–4003.
62. Casari G, Barlassina C, Cusi D, Zagato L, Muirhead R, Righetti M, Nembri P, Amar K, Gatti M, Macciardi F *et al*: Association of the alpha-adducin locus with essential hypertension. *Hypertension* 1995, 25:320–326.
63. Chuang SE, Daniels DL, Blattner FR: Global regulation of gene expression in ***Escherichia coli***. *J Bacteriol* 1993, 175:2026–2036
64. Van Bogelen RA, Sankar P, Clark RL, Bogan JA, Neidhardt FC: The gene-product database of ***Escherichia coli***: edition 5. *Electrophoresis* 1992, 13:1014–1054.
65. Sankar P, Hutton ME, Van Bogelen RA, Clark RL, Neidhardt FC: Expression analysis of cloned chromosomal segments of ***Escherichia coli.*** *J Bacteriol* 1992, 175:5145–5152.
66. Itaya M, Tanaka T: Complete physical map of the ***Bacillus subtilis*** 168 chromosome constructed by a gene-directed mutagenesis method. *J Mol Biol* 1991, 220:631–648.
67. Takiff HE, Baker T, Copeland T, Chen SM, Court DL: Locating essential ***Escherichia coli*** genes by using mini-Tn10 transposons. *J Bacteriol* 1992, 174:1544–1553.
68. Itaya M: An estimation of minimal genome size required for life. *FEBS Lett* 1995, 362:257–260.
69. Doolittle WF, Brown JR: Tempo, mode, the progenote, and the universal root. *Proc Natl Acad Sci USA* 1994, 91:6721–6728.
70. Benner SA, Ellington AD, Tauer A: Modern metabolism as a palimpsest of the RNA world. *Proc Natl Acad Sci USA* 1989, 86:7054–7058.
71. De Duve C: *Blueprint for a Cell*., London: Portland Press; 1991.
72. Hilbert H, Himmelreich R, Pirkl E, Plagens H, Proft T, Herrmann R: Analysis of the ***Mycoplasma pneumoniae*** genome. *Abstr 95th Gen Meet Am Soc Microbiol*. 1995: 505.
73. Kaneko T, Tanaka A, Sato S, Kotani H, Sazuka T, Miyajima N, Sugiura M, Tabata S: Sequence analysis of the genome of the unicellular cyanobacterium ***Synechocystis*** sp. strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64 % to 92 % of the genome. *DNA Res* 1995, 2:153–166; 191–198 (suppl).
74. Kalman SS, Dietrich FS, Davis RW, Stephens SS: Partial sequence of the ***Chlamydia trachomatis*** genome. *Gen Sci Technol* 1995, 1:P-42.
75. Nowak R: Bacterial genome sequence bagged. *Science* 1995, 269:468–470
76. Smith DR, Lee HM, Dubois J, Qui D, Caubet W, Bashirzadeh R, Parenteau P, Wierzbowski J, Wang X, Shimer S, Nolling J, Reeve J: Microbial genome sequencing. *Gen Sci Technol* 1995, 1: P-48
77. Honore N, Bergh S, Chanteau S, Doucet-Populaire F, Eiglmeier K, Garnier T, Georges C, Launois P, Limpaiboon T, Newton S, *et al*: Nucleotide sequence of the first cosmid from the Mycobacterium leprae genome project: Structure and function of the Rif-Str region. *Mol Microbiol* 1993, 7:207–214.
78. Rabb FT, Brummet SR, Bogert A, Hujer K, Krall J, Domke S, Szasz J, Borges KM, Davis M, Fuller C, Chase JW: ‘Eukaryotic’ gene functions in the hyperthermophilic archaeon, ***Pyrococcus furiosus***. *Gen Sci Technol* 1995, 1:P-46
79. Olsen GJ, Woese CR, Overbeek R: The winds of (evolutionary) change: breathing new life into microbiology. *J Bacteriol* 1994, 176:1–6
80. Schuler GD, Altschul SF, Lipman DJ: A workbench for multiple alignment construction and analysis. *Prot Struct Funct Genet* 1991, 9:180–190.