Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*

Roman L.Tatusov*§, Arcady R. Mushegian*§, Peer Bork†, Nigel P. Brown†, William S. Hayes‡, Mark Borodovsky‡, Kenneth E. Rudd* and Eugene V. Koonin*

Background: The 1.83 Megabase (Mb) sequence of the *Haemophilus influenzae* chromosome, the first completed genome sequence of a cellular life form, has been recently reported. Approximately 75 % of the 4.7 Mb genome sequence of *Escherichia coli* is also available. The life styles of the two bacteria are very different – *H. influenzae* is an obligate parasite that lives in human upper respiratory mucosa and can be cultivated only on rich media, whereas *E. coli* is a saprophyte that can grow on minimal media. A detailed comparison of the protein products encoded by these two genomes is expected to provide valuable insights into bacterial cell physiology and genome evolution.

Results: We describe the results of computer analysis of the amino-acid sequences of 1703 putative proteins encoded by the complete genome of *H. influenzae*. We detected sequence similarity to proteins in current databases for 92 % of the *H. influenzae* protein sequences, and at least a general functional prediction was possible for 83 %. A comparison of the *H. influenzae* protein sequences with those of 3010 proteins encoded by the sequenced 75 % of the *E. coli* genome revealed 1128 pairs of apparent orthologs, with an average of 59 % identity. In contrast to the high similarity between orthologs, the genome organization and the functional repertoire of genes in the two bacteria were remarkably different. The smaller genome size of *H. influenzae* is explained, to a large extent, by a reduction in the number of paralogous genes. There was no long range colinearity between the *E. coli* and *H. influenzae* gene orders, but over 70 % of the orthologous genes were found in short conserved strings, only about half of which were operons in *E. coli*. Superposition of the *H. influenzae* enzyme repertoire upon the known *E. coli* metabolic pathways allowed us to reconstruct similar and alternative pathways in *H. influenzae* and provides an explanation for the known nutritional requirements.

Conclusions: By comparing proteins encoded by the two bacterial genomes, we have shown that extensive gene shuffling and variation in the extent of gene paralogy are major trends in bacterial evolution; this comparison has also allowed us to deduce crucial aspects of the largely uncharacterized metabolism of *H. influenzae*.

Addresses: *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA. †European Molecular Biology Laboratory, Meyerhofstrasse 1, D-6900, Heidelberg, Germany. ‡School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332-0230, USA.

§The first two authors contributed equally to this work.

Correspondence to: Eugene V. Koonin
E-mail address: koonin@ncbi.nlm.nih.gov

Received: 13 November 1995
Revised: 2 January 1996
Accepted: 15 January 1996

Current Biology 1996, Vol 6 No 3:279–291

 ISSN 0960-9822

Background

The 1.83 Megabase (Mb) sequence of the *Haemophilus influenzae* chromosome, the first complete genome sequence of a cellular life form, has been recently reported by Fleischmann *et al.* [1]. Approximately 75 % of the 4.7 Mb genome sequence of *Escherichia coli* is also available [2]. The life styles of the two bacteria are quite different. *H. influenzae* is an obligate parasite that is adapted to living in human upper respiratory mucosa, occasionally invading the bloodstream and cerebrospinal fluid; it can be cultivated only on rich media [3,4]. *E. coli* is a saprophyte that can grow on minimal media [5]. Phylogenetically, *E. coli* and *H. influenzae* are relatively close to one another, both belonging to the same branch within the gamma subdivision of purple bacteria [6]. *E. coli* is arguably the best studied of all organisms [7]. In contrast, the information that is currently available on the biochemistry and physiology of *H. influenzae* is limited ([3,8–11] and references therein). Thus, a comparison of the *E. coli* and *H. influenzae* genomes is expected to be instrumental in deducing the metabolism of a poorly characterized bacterium from that of a well understood one. It may also provide insights into the nature of changes in gene repertoire and genome organization that parallel the 2.5-fold difference in genome size and the physiological dissimilarity between the two bacteria. Such information will be important for studying the molecular basis of *H. influenzae* pathogenicity.

Here we present the results of our computer analysis of the amino-acid sequences of 1703 putative proteins encoded by the complete genome of *Haemophilus influenzae*, and of our comparison of these sequences with those of proteins known to be encoded by the *E. coli* genome.

Results and discussion

Re-evaluation of the set of *H. influenzae* genes

H. influenzae is the first organism for which protein sequence conservation could be evaluated in the context of the complete genome. A prerequisite for such an analysis is a set of genes that can confidently be predicted to encode real proteins. The *H. influenzae* genome sequence submitted to GenBank includes 1747 predicted protein-encoding genes; 50 of these coding regions contained frameshifts and were not translated — these have been referred to as ''miscellaneous features'' ([1], and subsequent Genbank updates). We compared the intergenic regions from the *H. influenzae* GenBank entry to the protein sequence databases using the BLASTX program. The results revealed a number of highly significant sequence similarities, indicating that these regions may contain additional genes. Furthermore, among the putative *H. influenzae* proteins referred to in [1], there are a number of very short ones, and it is unclear whether the respective genes have been correctly identified.

Given these uncertainties, we considered it necessary to revise the set of proteins encoded by the *H. influenzae* genome. Using an approach that combines sequence similarity searches with statistical analysis of the DNA sequence, analysed using the GeneMark program (see Materials and methods), we produced a new set of 1703 putative *H. influenzae* protein-encoding genes. In addition to 1572 open reading frames (ORFs) that remained the same as in [1] (the updates taken into account), it contains 23 new ORFs, and 107 ORFs that are modifications of the original ones (including reconstructed translations of ''miscellaneous features''). These modifications included reassigning an initiation codon, and merging two adjacent ORFs that were separated by a frameshift, as deduced from their similarity to different parts of the same protein in the database. We eliminated 47 short putative genes from the original set, because their existence could not be corroborated by any of the applied methods. Frameshifts or in-frame stop codons were found in 62 *H. influenzae* genes (3.6 % of the total). Several genes also contained large deletions and/or rearrangements, as compared to the *E. coli* orthologs. Many of these mutations might have accumulated during the numerous laboratory passages of the Rd strain of *H. influenzae* that was used for sequencing, and the respective genes may not express functional proteins. An additional unusual feature was a direct repeat of 5432 nucleotides at positions 1 410 877–1 416 308 and 1 544 123–1 549 554 of the *H. influenzae* genome, which contained 5 complete genes. The origin of this repeat remains unclear. We reoriented the sequence of the *H. influenzae* chromosome to start from the replication origin (*oriC*), and new names, HIN0001–HIN1703, with pointers to *E. coli* orthologs and the original numbers from [1], were assigned to all of the putative proteins.

Conservation of protein sequences between *H. influenzae* and *E. coli*, and prediction of *H. influenzae* protein functions

The revised set of 1703 protein sequences encoded by the *H. influenzae* genome was compared to the NR database, as described in Materials and methods. A detailed analysis of relatively weak sequence similarities, as well as the rapid growth of databases, allowed us to detect statistically significant sequence similarities for 92 % of the *H. influenzae* proteins as compared to 78 % in [1]. About two-thirds of the proteins contained regions that are conserved at least at the level of distantly related bacteria (defined as those outside the *Proteobacteria* domain in the bacterial phylogenetic tree [6]). The fraction of proteins containing ancient conserved regions (ACRs) — that is, sequences shared with eukaryotic or archaeal proteins [12] was only slightly higher in *H. influenzae* than in *E. coli* [13], in spite of the 2.5-fold difference in the total number of proteins (Fig. 1). In other words, the much smaller set of *H. influenzae* gene products was not significantly enriched in highly conserved proteins. Conceivably, the presence of ACRs in about 50 % of the proteins may be typical of bacteria in general.

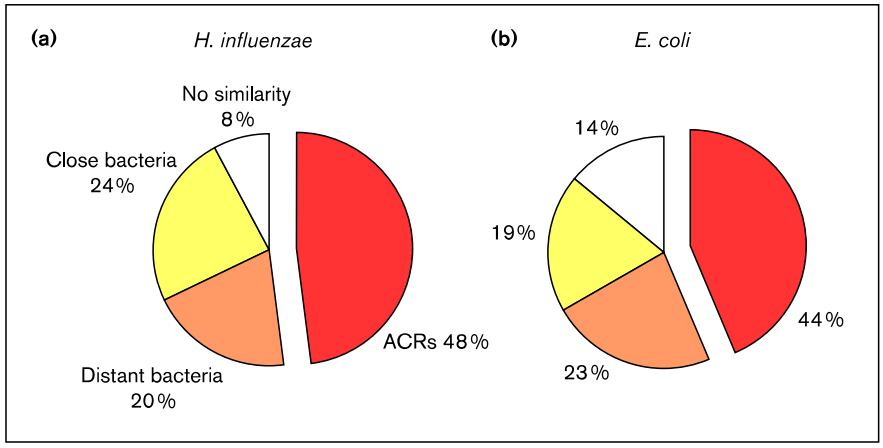
Originally, functional predictions were reported for 58 % of the putative *H. influenzae* proteins [1]. In a subsequent study, functions of another 8 % were tentatively identified [14]. The study by Fleischmann *et al.* [1] does not describe any attempt to predict the functions of *H. influenzae* gene products homologous to uncharacterized 'hypothetical proteins' in the database. In order to identify additional reliable similarities to functionally characterized proteins, we examined each of these cases using methods for motif analysis and multiple alignment. As a result, functional predictions were made for 83 % of the *H. influenzae* gene products. We were able to assign 63 % of the putative proteins to one of 12 broad categories that are a modification of the classification introduced by Riley [15]; for the remaining proteins only a more general prediction — such as ATP-utilizing enzyme, permease, or transcriptional regulator containing a helix-turn-helix domain — was feasible (Table 1). Notably, among the 295 putative *H. influenzae* gene products whose functions could not be predicted, 27 contained highly conserved regions that may represent novel families of essential proteins.

The vast majority of *H. influenzae* genes have orthologs in *E. coli*

In order to compare genomic organization or metabolic pathways in *E. coli* and *H. influenzae*, it is first necessary to clearly distinguish between orthologs and paralogs among

Figure 1

Amino-acid sequence conservation in *E. coli* and *H. influenzae*. The graph is organized using an hierarchical rule: 'distant bacteria' indicates that the proteins in the given set have detectable homologs in distantly related bacteria but not in eukaryotes or Archaea; 'closely related bacteria' indicates that the proteins in this set do not have detectable homologs in distantly related bacteria.



the genes of these bacteria. Orthologs have been defined as homologous genes in different organisms that encode proteins with the same function, and that have evolved by direct, vertical descent; paralogs are homologous genes within an organism encoding proteins with related but non-identical functions [16]. Because a specific function can usually only be inferred for orthologs, a reliable list of such genes is critical for functional and phylogenetic reconstructions. Similarly, comparative analysis of genome organization can be based only on a comparison of the order of orthologous genes.

Table 1

Functional classification of *H. influenzae* proteins based on sequence similarity.

Functional category*	Number of *H. influenzae* proteins	Percentage of total *H. influenzae* proteins
Amino-acid metabolism and transport	163	9.6
Genome replication, transcription, recombination and repair	141	8.3
Energy conversion	141	8.2
mRNA translation and ribosome biogenesis	125	7.3
Outer membrane and cell wall	105	6.2
Carbohydrate metabolism and transport	80	8.3
Nucleotide metabolism and transport	73	4.3
Cofactor metabolism	69	4.0
Chaperones	53	3.1
Inorganic ion transport	52	3.0
Lipid metabolism	40	2.3
Secretion	35	2.1
General functional prediction only	331	19.4
Total predicted function	1408	82.7

*Each category also included proteins that are involved in the regulation of expression of the respective genes.

Even though a number of sequence similarities between *E. coli* and *H. influenzae* proteins have been detected [1], the issue of orthology versus paralogy has not been specifically addressed. We therefore performed a new comparison of the 1703 putative *H. influenzae* proteins with the 3010 proteins encoded by the sequenced 75 % of the *E. coli* genome ([17]; K.E.R., data not shown). An *E. coli* protein was considered to be the ortholog of a *H. influenzae* protein if: firstly, it showed at least several percent higher similarity to a given *H. influenzae* protein than to any other *E. coli* protein; secondly, the two proteins showed a higher similarity to each other than to homologs from phylogenetically more distant organisms; and thirdly, the sequences of the potential orthologs aligned throughout most of their lengths (with a few exceptions for domain deletions). In a few cases, when comparable similarity was observed between the given *H. influenzae* protein sequence and several *E. coli* proteins, a decision on orthology could be reached on the basis of gene position analysis; thus, if a pair of proteins from *H. influenzae* and *E. coli* belonged to apparently orthologous operons (see below), they were likely to be orthologs. A similar approach to ortholog identification has been recently used in the analysis of a partial sequence of the *Mycoplasma capricolum* genome [18].

Altogether, significant sequence similarity to *E. coli* proteins was detected for 1307 putative *H. influenzae* proteins; of these, 1128 appeared to have an *E. coli* ortholog. As 66 % of the *H. influenzae* genes have an ortholog among the 75 % of the *E. coli* genes whose sequences are currently available, it may be inferred that, when the sequencing of the *E. coli* genome is complete, orthologs will be detected for nearly 90 % of the *H. influenzae* genes.

The percentage identity between orthologs spanned a wide range (between 18 % and 98 %), but the majority of pairs (875 out of 1128) contained between 40 % and 80 % identical amino-acid residues, with the median of the distribution at 60 % identity. On average, the *H. influenzae*/*E. coli* orthologs contained 59 % identical and 75 % similar amino-acid residues. The typically high sequence conservation between orthologs contrasts with the 2.5-fold difference in the total number of genes in the two genomes. The list of orthologs provides a basis for attempts to understand the nature and origin of this difference and to reconstruct at least some of the unknown biochemical pathways of *H. influenzae*.

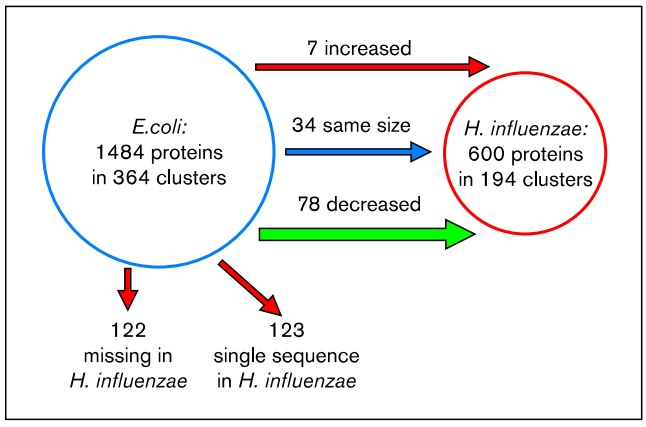
Low level of gene paralogy in a small bacterial genome

About 50 % of *E. coli* proteins belong to clusters of paralogs [13]. Most of these groups are small, although there are several large clusters, which consist of proteins involved in metabolite transport and regulation of gene expression. We performed an analogous clustering of the *H. influenzae* proteins. The level of paralogy, defined as the ratio of proteins that have paralogs to the total number of proteins, was lower in *H. influenzae* than in *E. coli* — only about one third of the *H. influenzae* proteins had paralogs. Very recently, a closely similar estimate of the number of paralogs in *H. influenzae*, based on a different method for comparing protein sequences, has been published [19].

A comparison of individual clusters (Fig. 2) revealed that the majority of them were either missing in *H. influenzae* or were represented by a single protein, orthologous to one of the proteins in the respective *E. coli* cluster. Most of the other clusters contained a smaller number of proteins in *H. influenzae* than in *E. coli*, and we only observed the opposite situation for seven clusters. Specifically, several families of permeases and regulators of transcription in *E. coli* were reduced to a single protein in *H. influenzae*, and the total number of proteins in each of these classes was much smaller than in *E. coli* (Table 2). Of the 110 pairs of paralogous enzymes (typically differentially regulated) that have so far been detected in *E. coli*, 60 were represented by only one ortholog in *H. influenzae*, whereas 26 were absent. The low level of paralogy in *H. influenzae*, compared to *E. coli*, may be explained by a reduction in the specificity of transport and regulatory proteins, simplified regulation of many metabolic pathways and an absence of entire functional systems that in *E. coli* are mirrored in other, paralogous systems (such as those utilizing different sugars).

Large clusters of paralogs are particularly unusual in *H. influenzae*, with only 2 containing 10 or more proteins, in contrast to 18 in *E. coli*. Over 20 % of the *E. coli* proteins belong to the four largest superclusters that, in addition to permeases and helix-turn-helix regulatory proteins, include ATPases and GTPases with the conserved 'Walker-type' motif, and dinucleotide-binding proteins [13]. We found that, compared to *E. coli*, the number of transcriptional regulators containing the helix-turn-helix domain and transporters (permeases) in *H. influenzae* was even smaller than expected, based on the difference in the total number of genes (Table 2). In contrast, the number of ATPases and GTPases is greater than expected, and in *H. influenzae* these proteins made up an even larger fraction of gene products than in *E. coli* (Table 2). Conceivably, the fraction of ATP- and GTP-binding proteins that are essential for *H. influenzae* cell function is higher than it is in the other large protein superclusters.

Figure 2



Clusters of paralogous proteins in *E. coli* and *H. influenzae*.

Table 2

The four largest superclusters of paralogs among *E. coli* and *H. influenzae* proteins*.

	Number of proteins (% of total)	
	E. coli	*H. influenzae*
Permeases†	210 (7.0 %)	82 (4.8 %)
ATPases and GTPases, with the 'Walker type' NTP-binding motif	175 (5.8 %)	130 (7.6 %)
DNA-binding proteins containing helix-turn-helix domains (mostly regulators of transcription)	145 (4.8 %)	49 (2.9 %)
NAD- and FAD-binding proteins (mostly oxidoreductases)	90 (3.0 %)	42 (2.5 %)
4 superclusters combined	620 (20.6 %)	303 (17.8 %)

*In addition to the proteins initially clustered on the basis of pairwise sequence comparisons, each of the superclusters also includes a number of proteins that contain only the respective conserved motif. The alignment blocks used for motif searches have been delineated in the course of our previous analysis of *E. coli* proteins [13]. †The indicated number of permeases is based on the results of BLASTP searches and searches with several distinct motifs and should be considered to be an estimate.

Table 3

Some protein families in *H. influenzae* without orthologs in *E. coli*.

Protein function/activity	Protein names	Homologs in other organisms	Implications
Immunoglobulin A1 proteases	HIN0408, HIN1382	Orthologs in *Neisseria* (such as GenBank (GB) X82724), paralogs in pathogenic *E.coli* strains (in K-12, *yejO* is interrupted by an insertion sequence), and in *Shigella* (GB Z48219)	Inactivation of host immunoglobulins by cleavage into Fab and Fc fragments. Variation of antigenically different IgA1 proteases is frequent upon infection and is thought to be important for pathogenesis [53]
Glycosyltransferases	HIN0181, HIN1672, distantly related HIN1090	Apparent orthologs in *Neisseria* (GB U14554) and in *Pasteurella haemolytica* (GB U15958)	HIN1097 (Lex2B) is involved in the biosynthesis of the lipopolysaccharide outer core [54]; see also Table 5
Opacity proteins	HIN0586, HIN0861, HIN1514	Closely related to *Neisseria* opacity proteins (such as GB Z18941)	Opacity proteins are involved in adherence. Subject to phase variation [55,56]
Uncharacterized proteins	HIN0531, HIN1587	*Shigella* VirK (GB D11025) and a protein in *Pasteurella haemolytica* (GB M59210)	The *Pasteurella* protein is encoded within a locus that inhibits the expression of a leukotoxin [57]
Virulence-associated proteins	HIN1452, HIN0364	Virulence-associated proteins of *Dichelobacter* (GB M74565) and *Shigella* (GB X66934)	
Adhesins	HIN1123, HIN1124	Weak similarities to several coiled-coil proteins	Coiled-coil proteins involved in surface interactions
Transferrin-binding proteins	HIN0389, HIN0412	Transferrin-binding proteins of *Neisseria* (GB U05205) and *Actinobacillus* (GB M85275)	The system of enterochelin biosynthesis is incomplete in *H. influenzae* (*entB*,*D*, *fec B*,*I* orthologs are missing). This gene family might be a part of an alternative pathway of iron uptake
Sodium-dependent neurotransmitter-type amino-acid transporters	HIN0151, HIN1083	Numerous amino-acid transporters in eukaryotes, none in bacteria	Horizontal transfer of a eukaryotic gene?
Zn finger proteins	HIN0698, HIN0744	Weak similarities to transcription factors from *Xenopus* (GB L13702, L81986)	Putative transcription regulators

We found 18 clusters of paralogs that were present in *H. influenzae*, but not in *E. coli* (counterparts to some of these may be eventually identified among the remaining 25 % of the *E. coli* genes). Notably, proteins in some of these unique groups were similar to pathogenicity factors of other bacteria, such as *Neisseria* and *Shigella*, and included cell-surface proteins, enzymes involved in the biosynthesis of surface lipopolysaccharides, and immunoglobulin proteases. These proteins may be directly relevant to *H. influenzae* pathogenicity (Table 3). Another unique cluster included two putative permeases that showed highly significant similarity to eukaryotic amino-acid transporters, but had no bacterial homologs, suggesting that these *H. influenzae* genes may have been acquired by horizontal transfer. A comparison of the organization of these clusters in Rd strain and pathogenic isolates of *H. influenzae* may be important for understanding the mechanisms of the *H. influenzae* pathogenicity.

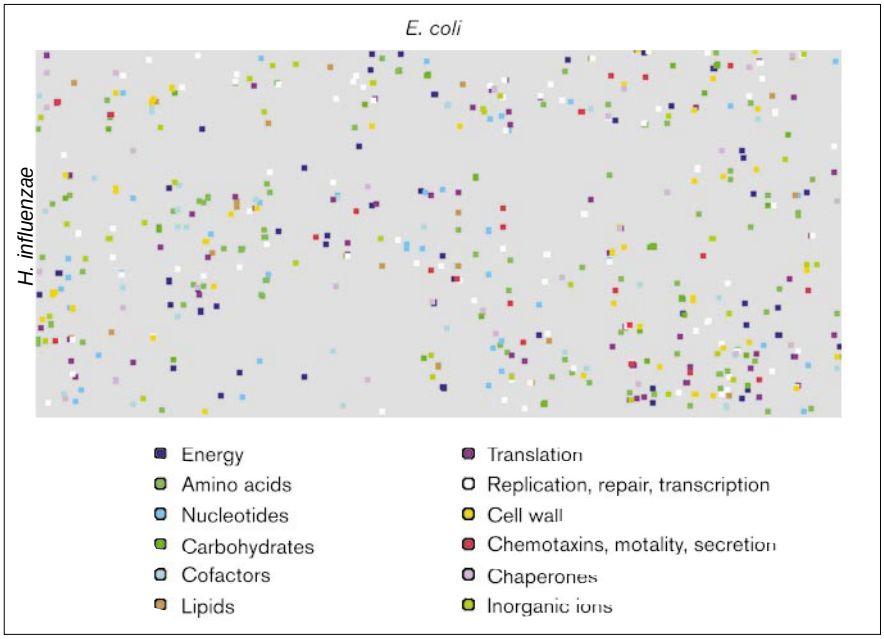
Extensive gene shuffling in bacterial evolution

Detailed comparisons of the genetic maps of *E. coli* and *Salmonella typhimurium* that are thought to be separated by 120–160 million years of evolution [20] have resulted in the paradigm of bacterial genome stability [21,22]. Indeed, in spite of the ample potential for different types of recombinational events, the two maps are very similar to each other [23], suggesting that strong selective constraints operate to preserve the gene order. Furthermore, some colinearity has been observed, even at much greater evolutionary distances, namely between the gene orders of *E. coli* and *Bacillus subtilis* [24,25]. On the other hand, significant differences between physical maps of *H. influenzae* type b strains have been described [26]. The availability of the complete *H. influenzae* genome sequence and most of the *E. coli* sequence allows one to address the genome rearrangement issue in a more definitive way.

A comparison of the positions of the orthologous genes in the *E. coli* and *H. influenzae* chromosomes clearly showed a lack of a long-range colinearity (Fig. 3). When the distribution of orthologs belonging to different functional categories (Table 1) was examined separately, no significant colinearity was detected either (Fig. 3). Locally, however, there was an easily discernible conservation of the gene order, with the majority of orthologs found in conserved gene strings. For the purpose of this comparison, we defined a conserved gene string as a group of two or more

Figure 3

Comparison of the order of orthologous genes in the *E. coli* and *H. influenzae* chromosomes. For each of the chromosomes, the replication origin, *oriC*, was chosen as the zero point. The axes represent the complete chromosomes in the clockwise direction. Each square is supposed to represent a pair of orthologous genes with the respective coordinates in the *E. coli* and *H. influenzae* chromosomes. However, because of the large scale of the plot, many squares actually correspond to several pairs. The functional categories of proteins are color-coded.



orthologous genes that are adjacent in both genomes, regardless of the direction of transcription. Most of the conserved strings contained only 2–4 genes, but there were also several long strings, with the largest one, the ribosome protein gene superoperon, containing 28 genes in a row (with an insertion of one unique gene in *H. influenzae*). Notably, only about one half of the conserved strings were operons in *E. coli* (Table 4). A large fraction of the genes in non-operon strings have no known functional relationship to one another and are frequently divergently transcribed. Even though some as yet unknown physiological links may exist, it seems unlikely that the conservation of many of the non-operon strings results from functional constraints. Rather, we believe that these conserved strings are vestiges of an ancestral gene order.

These findings show that there has been extensive gene shuffling along the evolutionary path, separating *E. coli* and *H. influenzae* from their common ancestor. In order to find out whether or not this shuffling was directly related to the difference in size between the *E. coli* and *H. influenzae* genomes, we evaluated the conservation of the gene order between *E. coli* and *Bacillus subtilis* (which has a 4.2 Mb genome [27]). In the two available long, contiguous *B. subtilis* sequences [28], we found 111 genes with readily detectable orthologs in *E. coli*; 21 of these genes are arranged in 6 strings conserved between *B. subtilis* and *E. coli*. In this set of 111 *E. coli* genes, 65 had orthologs in *H. influenzae*, with 46 genes belonging to 34 conserved strings (most of these conserved strings contained additional genes without counterparts in *B. subtilis*). From these observations, however limited, it seems that the extent of shuffling between bacterial genomes is more related to the evolutionary distance separating bacterial species than to the difference in the genome size.

Deducing *H. influenzae* metabolism from protein sequence comparisons

For 860 *H. influenzae* proteins, the *E. coli* orthologs could be assigned to one of the 12 broad functional categories. The representation of different categories of *E. coli* proteins by orthologs in *H. influenzae* was strongly non-uniform. Not unexpectedly, orthologs of numerous genes

Table 4

Conserved gene strings in *E. coli* and *H. influenzae*

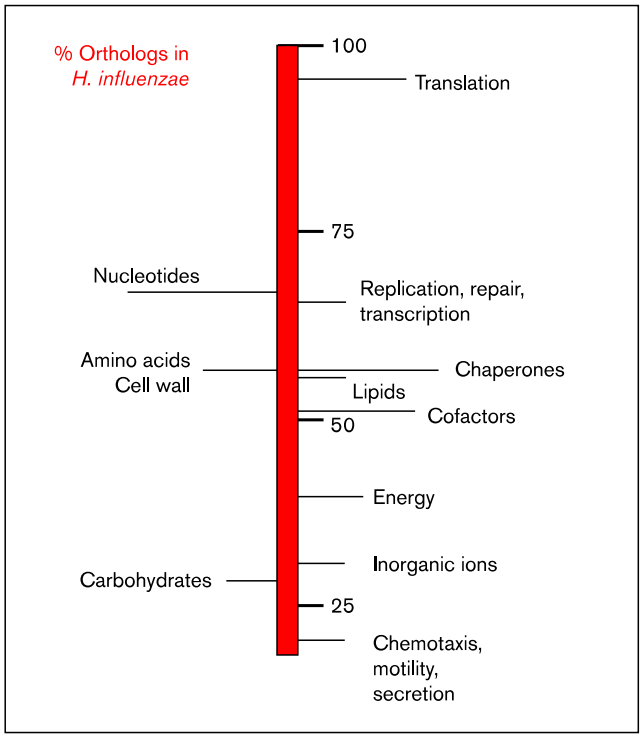
	Number of gene strings	Number of orthologous genes in conserved gene strings (% of total number of orthologs)
Total	226	825 (72.8 %)
Perfectly conserved strings	151	450 (39.7 %)
Conserved strings with 1–2 gene indels	75	375 (33.1 %)
Operons in *E. coli*	114	355 (31.3 %)
Not operons in *E. coli*	112	470 (41.5 %)

that are involved in cell motility and chemotaxis in *E.coli* (such as the *fim*, *flg*, *flh*, *fli*, *che* and *mot* genes) were missing from the genome of the non-motile *H. influenzae*. Nearly the full complement of proteins involved in translation was conserved, whereas the majority of the proteins involved in carbohydrate metabolism and transport were missing in *H. influenzae*; the representation of the other categories was intermediate between these two extremes (Fig. 4). The proteins involved in translation were also most highly conserved at the amino-acid sequence level, but otherwise, there was no obvious correlation between the fraction of proteins represented by orthologs within a functional category and sequence conservation (data not shown).

We attempted to systematically deduce the metabolism of *H. influenzae* from the results of protein sequence comparisons by superimposing the predicted enzymatic activities of orthologs on the known *E. coli* metabolic pathways. In addition, we examined the predicted functions of those *H. influenzae* proteins that did not have orthologs in *E. coli*, but had homologs in other organisms, because we reasoned that some of these genes might provide missing links in the reconstructed pathways. Table 5 summarizes the status of the principal metabolic pathways in *H. influenzae*, as deduced through the comparison with *E. coli*. The general trend seems to be that the repertoire of *H. influenzae* metabolic enzymes, as compared to that of *E. coli*, is tuned towards reducing conditions. Most of the enzymes of the respiratory chain that are known to function under aerobic conditions in *E. coli* are missing in *H. influenzae*. For example, 7 heme-dependent enzymes are encoded by the *H. influenzae* genome, compared to 25 in *E. coli*; in *H. influenzae*, none of these enzymes seem to be components of aerobic respiratory chains (Table 5). Furthermore, ubiquinone, the principal aerobic electron acceptor, has not been found in the bacteria of the genus *Haemophilus*, whereas menaquinone, the anaerobic acceptor, is present [29]. Accordingly, no orthologs of the genes encoding enzymes of ubiquinone biosynthesis were detected in *H. influenzae*, whereas the menaquinone biosynthesis system was intact.

The notion that *H. influenzae* has a largely anaerobic metabolism is compatible with the dramatic alteration of the tricarboxylic acid (TCA) cycle. It has been noticed that in *H. influenzae* the TCA cycle is incomplete, because of the absence of three enzymes, namely citrate synthase, isocitrate dehydrogenase and aconitase [1]. We found that another TCA cycle enzyme, succinate dehydrogenase, was also missing in *H. influenzae*. For yet another TCA cycle enzyme, fumarate hydratase, which in *E. coli* is represented by three paralogs, only one protein was found in *H. influenzae*, and the respective gene, *fumC*, contained a frameshift and may not express an active enzyme (it should be noted that fumarate hydratase activity has been

Figure 4



Representation of different functional categories of *E. coli* proteins by orthologs in *H. influenzae*. The scale shows the percentage of *E. coli* proteins in each functional category represented by orthologs in *H. influenzae*.

detected in clinical isolates of *H. influenzae* [30]). What remains of the TCA cycle in *H. influenzae* strain Rd is unlikely to form a cyclic, catabolic pathway, given also the absence of the enzymes of the glyoxylate cycle. Instead, a non-cyclic, branched, biosynthetic pathway can be envisaged (Fig. 5a). This pathway has been reported to function in *E. coli* under anaerobic conditions or under aerobic conditions in the presence of glucose [31]. Oxaloacetate entering this pathway is probably formed from pyruvate or phosphoenolpyruvate. Enzymes for both types of reactions were encoded by *H. influenzae* (Fig. 5a).

In contrast to the dramatic reduction in respiratory chains and carbohydrate metabolism, the 'damage' to the central reactions of nucleotide and amino-acid metabolism in *H. influenzae* was limited. The initial three steps in the biosynthesis of pyrimidines and five steps in the biosynthesis of arginine are lacking; in both cases, the missing steps can be circumvented by the presence of the same intermediate, citrulline; citrulline (or arginine plus uracil) is required for *H. influenzae* growth [32].

The pathway of cysteine biosynthesis and anabolic sulfate reduction in *H. influenzae* remains unclear. There were no

Table 5

Biochemical pathways in *H. influenzae* as deduced by protein sequence comparisons with *E. coli.*

Pathways	Genes present in both *E. coli* and *H. influenzae*	Enzymatic activities and genes missing in *H. influenzae*	Functional implications for *H. influenzae*
Energy conversion			
Glycolysis	*ptsG, crr, pgi, pfk, fda, tpiA, gap, pgk gpm,eno, pyk*	Glucokinase (*glk*)	Glucose-6-phosphate is probably synthesized by a glucokinase that is orthologous to the enzyme of *Streptomyces*, and to a novel putative glucokinase of *E. coli* (*yhcI* gene product), but not to *glk*
Pyruvate dissimilation	*pfl, pta, fld, fdx, ackA, pykA, ppc.*	Ferredoxin–NADP reductase (*fpr*).	Functional in *H. influenzae*
TCA cycle	*sucA,B,C,D, lpd, mdh, aceE,F, frdA,B,C,D*	Citrate synthase(*gltA*), aconitase(*acnA*), isocitrate dehydrogenase (*icd*), succinate dehydrogenase (*sdhA,B,C,D*), fumarases (*fumA,B*), frameshift in *fumC*	Anabolic rather than catabolic pathway (See Fig. 5a)
Respiratory chains	More than 40 genes, including oxidoreductases, (de)hydrogenases, cytochromes, enzymes of menaquinone biosynthesis (*men*)	Ubiquinone biosynthesis (*ubiA–E,H,X*); cytochrome O (*cyo*), NADH–ubiquinone oxidoreductase (*nuo*), aerobic α-glycerophosphate dehydrogenase (*glpD*) nitrate reductase (*nar*), hydrogenases I and II (*hya, hyb, hyp*), formate-hydrogen lyase (*hyc*); peroxidases (*katG, ykjA*)	Lack of ubiquinone, presence of menaquinone, repertoire of electron transfer enzymes, and a limited number of peroxidase genes suggest that *H. influenzae* metabolism is largely anaerobic
Pentose phosphate and Entner–Duodoroff pathways	*zwf, pgl, gnd, rpi, tkt*, *edd* (a paralog), *eda*		Both pathways are likely to function in *H. influenzae*
Glyoxylate bypass	None	Isocitrate lyases (*aceA,B*), malate synthase (*glcB*)	This pathway is missing in *H. influenzae*
Nucleotide metabolism			
Purines	*purA,B,D,E,F,K,L,M,N, guaA,B,C, adk, gmk, ndk.* One gene, *purC*, is represented by a distant paralog in *H. influenzae.*		SAICAR synthetase of *H. influenzae* is only distantly related to the known PurC enzyme of *E. coli*, but closely related to streptococcal/eukaryotic PUR7/ADE1.
Pyrimidines	*pyrD,E,F,G*	Carbamoyl-phosphate synthase subunits(*carA,B*), aspartate carbamoyl transferase (*pyrB,I*), dihydroorotase (*pyrC*)	Three initial steps (from the transfer of glutamine amido group onto bicarbonate to the formation of dihydroorotate) are missing in *H. influenzae*. This explains the requirement for uracil* in the growth medium. Uracil can be substituted by citrulline, which is linked to the pyrimidine pathway via the *argI* gene product (see also Arg biosynthesis, this table)
Amino-acid metabolism			
Arg and polyamines	*argI,G,H* represented orthologs; there are only paralogs in *H. influenzae* for *argE*, *speA* and *speC* genes	*argA,B,C,D,F,J; speB,D,E,G.*	Enzymes for five initial steps of arginine biosynthesis (leading from glutamine to citrulline) are missing in *H. influenzae*. Exogenous citrulline can salvage both arginine and pyrimidine biosynthesis; otherwise, arginine and uracil have to be provided. Absence of *spe* genes suggests that exogeneous polyamines are also needed
Cys	*cys E,K,Z*	*cysA,C,D,G,H,I,J,M.*	Serine acetylation (*cysE*), sulfide conjugation onto O-acetylserine (*cysK*) and sulfate permease (*cysZ*) are present. The steps from sulfate to sulfide (*cysC,D,G,H,I,J*) are missing in *H. influenza*e, as are the orthologs for another sulfite reduction system, *asnA,B,C* of *S. typhimurium*. Cysteine requirement is not absolute in *H. influenzae*, suggesting that sulfate reduction is in fact performed, presumably by a putative anaerobic oxidoreductase

Table 5 continued

Pathways	Genes present in both *E. coli* and *H. influenzae*	Enzymatic activities and genes missing in *H. influenzae*	Functional implications for *H. influenzae*
Amino-acid metabolism			
Glu, Gln, Asp,. and Asn	*glnA,B,D,E, gdhA, aspA, asnA*	*gltBD, asnB, rpoN ntrBC*	Glutamate is a major biosynthetic precursor in a pathway that involves some enzymes of the TCA cycle (Fig. 5a), and a crucial component of the nitrogen assimilation pathway, which is modified in *H. influenzae* (Fig. 5b)
Lys, Thr and Met	*thrA,B,C; metB,C,E,H; dap A,B,C,D,E. lysC* has a long deletion in *H. influenzae*	*metA, metL.*	*thrA, metL, lysC* are three similar (but differentially regulated) aspartate kinases/homoserine dehydrogenases in a common pathway; in *H. influenzae*, *metL* is missing, and *lysC* has a large internal deletion; *metA* in *H. influenzae* is substituted by an ortholog of the yeast homoserine acetyl-transferase (*met2*)
Leu, Ile, and Val	*ilvA,C,D,E,G,H,I leuA,B,C,D*	*ilvB,M,N, avtA*	Elimination of the differentially regulated pathways
His, Gly, Ser, Ala, Pro, Phe, Tyr and Trp	All essential *E. coli* genes are also found in *H. influenzae*		
Carbohydrate metabolism			
		Systems of utilization of cellobiose, galacticol, arbutin, glucitol/sorbitol, maltose, mannose, mellibiose, tre-halose, including PTS and permeases	Inability to utilize these sugars (also see text)
Coenzymes			
NADH	*aspC, pnt, pncA, C, pnuA,B*	*nadA,B,C,D,E, pncB, cob, pnuC*	Genes for biosynthesis of NMN from aspartate (*nadA,B,C*) or nicotinate (*pncB*), and conversion of NMN into NAD (*nadD,E*), are missing. Repertoire of NAD(H)-dependent enzymes is significantly reduced (no aldehyde dehydrogenases or iron-dependent alcohol dehydrogenases)
Hemin	*hemM,Y*	*hemA,B,C,D,E,F,G*	The heme biosynthesis system is destroyed in *H. influenzae*. Out of the 25 heme-containing enzymes of *E. coli* , 18 do not have orthologs in *H. influenzae*. Of the 7 orthologs, 1 is a catalase, 2 are involved in nitrate/nitrite reduction (*nrfB* and *fdnI*), and 4 are uncharacterized gene products
Pantothenate and coenzyme A	*panE,F, coaA*	*panB,C,D*	Initial steps in pantothenate biosynthesis are missing.
Biotin	*bioA,B,D,E,F; birA; bisC*	*btuB,E,R; bioC,H*	Early steps of pimeloyl biosynthesis (*bioC,H*) are missing; in *E. coli,* alternative pathways exist but are poorly studied. The pathway from pimeloyl-CoA to biotin is intact; biotin-dependent enzymes are present in *H. influenzae* (such as the *accB* product)
Lipoate	*lipA,B*	Lipoate–protein ligase (*lplA*)	The product of *lipB* gene is the second lipoate-protein ligase in *E.coli* [58] and is apparently the only such enzyme in *H. influenzae*. Some lipoyl-linked enzymes are missing (*gcvH*), while others are present (*aceF, sucB*)
Pyridoxal	*pdxB,H, serC*	Pathway for synthesis of pyridoxine from hydroxythreonine (*pdxA,J*)	Alternative pathways of pyridoxal synthesis in *E. coli* exist but are poorly studied. Pyridoxal-dependent enzymes (aminotransferases) are present

Table 5 continued

Pathways	Genes present in both *E. coli* and *H. influenzae*	Enzymatic activities and genes missing in *H. influenzae*	Functional implications for *H. influenzae*
Glutathione	*gst, sspA, gor*	Gamma-glutamyl transpeptidase(*ggt*), glutaredoxin(*grxA*), glutamate-cysteine ligase(*gshA*), glutathione synthase (*gshB*), glutathione-S-transferases (*yibF, yqjG*); glutathione peroxidase (*btuE*) may be synthesized by a novel system	Two glutathione S-transferases and glutathione reductase are present; the enzymes for glutathione synthesis are missing; no orthologs of glutamate–cysteine ligases from eukaryotes. Some glutathione-dependent enzyme genes are missing in *H. influenzae* (*argE*). Glutathione unrelated to the one known in *E. coli,* or supplied exogeneously (via an oligopeptide transport system?)
Outer membrane			
Lipopoly-saccharides	*envA, fir, lpxA,B; kdsA,B, kdtA, rfaC,D,E,F,Q,G, P,B,J,K, rfaL.*	Inner core modification and outer core biosynthesis (*rfH,L,I,P,S,Y,Z, tol C*) O-antigen biosynthesis (*rfbC,D, X, rfc, rffD,L,M,T*)	Enzymes for biosynthesis of lipid A (*envA, fir, lpxA,B*), non-branched inner core (*kdsA,B, kdtA, rfaC,D,E,F*), and the first hexose residue of the outer core (*rfaG*) are orthologous in Enterobacteriaceae and in *H. influenzae*; further steps in lipopolysaccharide biosynthesis are quite different (also see Table 3)

*Nutrients required for *H. influenzae* growth are underlined.

orthologs or readily detectable paralogs of the *E. coli* enzymes that catalyze the reduction of sulfate to sulfide (products of the *cysC*, *D*, *G*, *H*, *I*, *J* genes), even though the sulfate permease was present. Conceivably, these steps may be carried out by an as yet unidentified oxidoreductase system (perhaps anaerobic); candidates for such a system exist among the putative *H. influenzae* proteins that have no *E. coli* orthologs.

Notably, among the two groups of aminotransferases that catalyze the interconversion of dicarboxylic amino acids and their amides in *E. coli*, only the ammonia-dependent enzymes were present in *H. influenzae*, whereas the glutamine-dependent enzymes were missing. Together with the observation that the σ^{54} RNA polymerase subunit and the σ^{54}-dependent NtrC-related transcription regulators are missing in *H. influenzae* [1], this illustrates the simplification of regulatory pathways in *H. influenzae*, as compared to *E. coli*, and suggests that *H. influenzae* is adapted to growth in a nitrogen-rich environment (Fig. 5b). Generally, the significant decrease in the number of proteins involved in nucleotide and amino-acid metabolism in *H. influenzae* as compared to *E. coli* (Fig. 4) results from the reduction in the multiplicity of transport and regulatory systems, rather than to elimination of central pathways.

It is known that several metabolites and coenzymes are either required for *H. influenzae* growth or strongly stimulate it [4]. Based on the results of protein sequence comparisons, each of these requirements can be explained in terms of the absence of specific enzymes. As a rule, the extent of damage in the biosynthetic pathway for a given coenzyme correlates with the extent of reduction in the number of enzymes that require this coenzyme for activity (Table 5).

The lipopolysaccharide component of the outer membrane in *H. influenzae* lacks a long O-antigen chain [11]. In accord with this, *H. influenzae* did not have orthologs of the *rfb*, *rfc*, and *rff* genes, which encode enzymes involved in the biosynthesis of O-antigen and the distal part of the lipopolysaccharide outer core. In contrast, orthologs of the *E. coli* enzymes that mediate the biosynthesis of lipid A, the inner core of the lipopolysaccharide and the first hexose residue of the outer core were encoded by *H. influenzae* (Table 5).

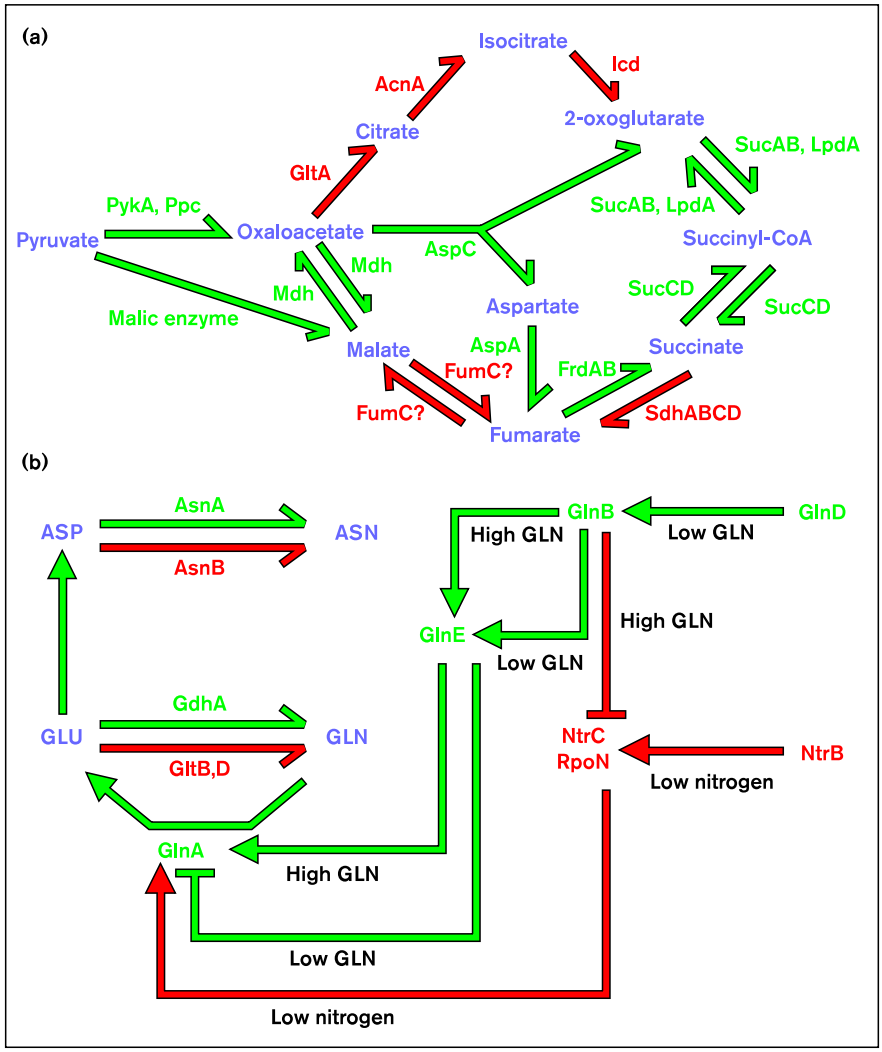
In general, it should be noted that Rd is a non-pathogenic laboratory strain whose gene repertoire might not be fully representative of all highly variable isolates of *H. influenzae* [4,11,30]; hence, caution should be exercised in generalizing these reconstructions to other *H. influenzae* strains.

Conclusions

The comparison of the *H. influenzae* and *E. coli* genomes reveals, at a new level, the dichotomy between ''tempo and mode'' that is typical of bacterial evolution [33]. In contrast to the generally high sequence conservation between orthologous proteins, drastic differences are observed both in the repertoires of genes and in their arrangement in the chromosomes of the two bacteria.

The 2.5-fold reduction in the number of genes in *H. influenzae* compared with *E. coli* is partly a result of the absence of many functional systems, notably respiratory chains and pathways of carbohydrate utilization. The

Figure 5



H. influenzae metabolic pathways deduced from sequence comparison. Green, enzymes and reactions conserved in *E. coli* and *H. influenzae*; red, enzymes and reactions found only in *E. coli*. (**a**) The pathways replacing the TCA cycle. Reactions of the *E. coli* TCA cycle are shown by clockwise arrows. The *fumC* gene in *H. influenzae* contains a frameshift, and the enzyme may not be functional. The pathway that operates in *E. coli* under anaerobic conditions and is likely to be the principal one in *H. influenzae* is: oxaloacetate → aspartate → fumarate → succinate → succinyl-CoA [31]. AcnA, aconitase; AspA, aspartase; AspC, glutamate–oxalate aminotransferase; GltA, citrate synthase; FrdAB, fumarate dehydrogenase; FumC, fumarate hydratase; Icd, isocitrate dehydrogenase; Mdh, malate dehydrogenase; SdhABCD, succinate dehydrogenase; SucAB, LpdA, 2-oxoglutarate dehydrogenase complex; SucCD, succinyl-CoA synthase. The *E. coli* malic enzyme sequence is unavailable; the *H. influenzae* malic enzyme was identified by similarity to the homologous enzymes from Gram-positive bacteria. (**b**) Simplification of the nitrogen assimilation cycle in *H. influenzae*. Both metabolic reactions and regulatory interactions [50–52] are shown. Reaction conditions are indicated in black. Lines with short bars at the end show negative regulation. NtrB and NtrC comprise the two-component system of transcriptional regulation that in *E. coli* activates the RpoN(σ^{54})-dependent transcription of *glnA* under low nitrogen conditions. The other regulatory interactions are covalent modifications of the target proteins. ASN, asparagine; ASP, aspartic acid; GLN, glutamine; GLU, glutamic acid; AsnA, asparagine synthase; AsnB, asparagine synthase B (glutamine-hydrolyzing); AspA, aspartate–ammonia lyase; GdhA, glutamate dehydrogenase; GlnA, glutamine synthase; GlnB, nitrogen regulatory protein P_{II}; GlnD, uridylyltransferase/uridylyl-removing enzyme; GlnE, adenylyltransferase; GltBD glutamate-synthase; The other designations are as in (a).

other major contribution to the decrease in the genome size comes from the lower extent of gene paralogy. This may involve a lower specificity of transport and regulatory systems and a simplification of regulatory networks.

There is no detectable long-range colinearity between the orders of orthologous genes in the *E. coli* and *H. influenzae* chromosomes but the majority of orthologs belong to conserved gene strings. Apparently some of these gene strings are conserved because of functional constraints, and others just for historical reasons.

A large amount of information on the metabolism and physiology of a bacterium is encrypted in its protein sequences. A detailed analysis of these sequences has allowed us to reconstruct the central metabolic pathways of *H. influenzae*. Notably, all of the known nutritional requirements could be explained in terms of the absence of specific enzymes.

Is *H. influenzae* a product of degenerative evolution that could be expected in a parasite, or is it close to a primitive ancestral bacterium? We believe that these views are not necessarily mutually exclusive. The average size of a bacterial genome is intermediate between that of *E. coli* and *H. influenzae* — approximately 3 Mb [34,35]. Conceivably, the common ancestor of *E. coli* and *H. influenzae* could have a genome of about this size. The subsequent evolution may have proceeded in different directions — primarily towards the reduction of the genome size by deletion

of genes and entire gene blocks in the *Haemophilus* lineage, and towards the diversification of regulatory and transport functions via gene duplication in the *E. coli* lineage.

This work is an attempt to deduce important aspects of bacterial metabolism and evolution from a comparison of two genomes, one of which is not yet complete. There is little doubt that in the near future, when the complete genome sequences of several diverse bacteria become available, such comparisons will develop into a powerful approach for reconstructing the physiology and genome organization of ancestral forms and deducing the biochemical strategies of extant bacteria, with likely practical implications. These deductions will serve as starting points for experimental verification of at least the most important of the predicted pathways.

Materials and methods

Databases

The non-redundant database is constructed at the National Center for Biotechnology Information (NIH) by merging non-identical entries from the GenPept, EMBL, SWISS-PROT, and PIR databases. The set of *E. coli* protein sequences used in this study is a conceptual translation of EcoSeq8, a non-redundant *E. coli* DNA sequence collection that contains approximately 75 % of the proteins encoded in the *E. coli* K-12 genome ([17]; K.E.R., unpublished). The *E. coli* metabolic pathways were primarily from [7]. Additional information was extracted from the EcoCyc database ([36]: URL http://www.ai.sri.com/ecocyc) and the PUMA database (URL http://www.mcs.anl.gov/home/compbio/PUMA). Whilst this manuscript was under review, two World Wide Web sites containing information on the H. influenzae metabolic pathways have become available (http://www.mcs.anl.gov/home/gaaster/hi and http://www.genome.ad.jp/kegg/kegg2).

Identification of protein-coding regions

For identifying protein-coding regions in *H. influenzae* DNA, the sequence, conceptually translated in all six reading frames was compared to the protein non-redundant database using the BLASTX program [37]. The regions that showed significant similarity to proteins in the database (see above) were conceptually translated into putative *H. influenzae* proteins; when a frameshift was detected, an attempt was made to reconstruct the protein sequence so as to maximize the similarity with its homolog(s). At the next step of analysis, the regions of the *H. influenzae* DNA between the putative genes identified with BLASTX were translated, and all the resulting protein sequences containing more than 50 amino-acid residues were compared to nucleotide sequence databases using the TBLASTN program, in order to identify additional similarities. The putative genes identified by the similarity analysis were used to derive a fourth-order Markov model of *H. influenzae* coding regions. In order to predict additional genes that may exist in the regions between the genes detected by the similarity searches, the GeneMark program [38] was run with the new matrix. The set of putative coding regions delineated by this two-step procedure was analyzed for overlaps, and when these were detected, the position of the starting codon in the downstream gene was verified by inspection of the alignment with the most closely related protein from the database.

Database searches and protein sequence alignment

The initial database screening was performed with the BLASTP program [39]. The resulting 'hits' were classified according to their taxonomic origin with the BLATAX program [40]. The BLASTP alignments with scores above 90 in virtually all cases indicate biologically relevant relationships [13,40]. The alignments with lower scores were ''post-processed'' by conserved motif analysis with the CAP and MoST programs [41]. Those proteins for which significant sequence similarity could not be detected on the basis of the BLASTP search results and motif analysis were searched against the nucleotide database translated in six reading frames with TBLASTN. This step is important for detecting additional homologs because many bacterial genes remain unannotated in sequence databases [42–44]. The proteins for which no homologs were detected at this step were subjected to an additional database screening with the FASTA program (ktuple=1) [45]; this version of FASTA has been reported to have a very high sensitivity [46].

Pairwise alignments of protein sequences were constructed with the BESTFIT program [47]. Multiple alignments were constructed with the MACAW program [48].

Comparison of gene orders

The program GENESTRING (R.L.T., unpublished) automatically produces the complete list of conserved gene strings from a list of orthologs with positions in the two chromosomes.

Clustering of related proteins

The *H. influenzae* protein set was compared to itself with BLASTP, and the proteins were clustered with a single-linkage algorithm. Using this approach, a cluster is defined as a group of protein sequences connected by similarity scores above a chosen cut-off (BLASTP score of 70 or greater, for both *E. coli* and *H. influenzae*) but without the requirement that each pair of sequences within a cluster had such a score. The BLASTP outputs for clusters containing more than two members were analyzed with the CLUSDOM program [40], in order to resolve the complications encountered by the single-linkage clustering algorithm with multidomain proteins. Such proteins may artifactually lump together otherwise unrelated clusters. Individual domains delineated with CLUSDOM were included in different clusters.

Availability of the results

The set of *H. influenzae* protein sequences used in this study, a table containing information on sequence conservation and predicted function for all putative proteins of *H. influenzae*, a list of *E. coli* and *H. influenzae* orthologs with alignment information, lists of clusters of paralogous proteins for *E. coli* and *H. influenzae*, and a list of conserved gene strings are available via the World Wide Web (URL http://www.ncbi.nlm.nih.gov/Complete_Genomes/Hin/, and by anonymous ftp at ncbi.nlm.nih.gov, directory repository/HIN). The sequence of the *H. influenzae* Rd genome reannotated based on the results of this analysis is available through the Genome division of GenBank [49].

Acknowledgements

We are grateful to D. Lipman and A. Bairoch for numerous helpful discussions, to A. Bairoch, M. Boguski, G. Casari, D. Landsman, D. Lipman, W. Nichols, A. Preston, J. Reizer, M. Sunshine, R. Wade and G. Weinstock for critical reading of the manuscript, and to J. McInnch for valuable programming assistance. The work of M.B. and W.S.H. was supported by NIH grant HG00783.

References

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, *et al*: Whole-genome random sequencing and assembly of ***Haemophilus influenzae*** Rd. *Science* 1995, 269:496–512.
2. 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.
3. Kilian M, Biberstein EL: ***Haemophilus*** Winslow, Broadhurst, Buchanan, Krumwiede, Rogers, and Smith. In *Bergey's Manual of Systematic Bacteriology*. Edited by Krieg NR, Holt JG. Baltimore, Maryland: Williams and Wilkins; 1984:558–569.
4. Hoiseth SK: The genus ***Haemophilus***. In *The Prokaryotes: handbook on the biology of bacteria.* Edited by Balows A, 2nd edn. New York: Springer–Verlag; vol 4, 3304–3330.

5. Ørskov F: ***Escherichia*** Castellani and Chalmers. In *Bergey's Manual of Systematic Bacteriology.* Edited by Krieg NR, Holt JG. Baltimore, Maryland: Williams and Wilkins; 1984:420–423.
6. Olsen GJ, Woese CR, Overbeek R: The winds of (evolutionary) change: breathing new life into microbiology. 1994, *J Bacteriol* 176:1–6.
7. Neidhardt F, Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE, eds: Escherichia coli *and* Salmonella typhimurium. *Cellular and Molecular Biology.* Washington, DC: American Society for Microbiology; 1987.
8. Kilian M, Frederiksen W: Ecology of ***Haemophilus***, ***Pasteurella*** and ***Actinobacillus***. In *Haemophilus, Pasteurella, and Actinobacillus group.* Edited by Kilian M, Frederiksen W, Biberstein E. London: Academic Press; 1981:11–38.
9. Friesen CA, Cho CT: Characteristic features of neonatal sepsis due to ***Haemophilus influenzae***. *Rev Infect Dis* 1986, 8:777–780.
10. Redfield RJ: Evolution of natural transformation: testing the DNA repair hypothesis in ***Bacillus subtilis*** and ***Haemophilus influenzae***. *Genetics* 1993, 133:755–761.
11. Zamze SE, Moxon ER: Composition of the lipopolysaccharide from different capsular serotype strains of ***Haemophilus influenzae***. *J Gen Microbiol* 1987, 133:1443–1451.
12. Green P, Lipman D, Hillier L, Waterston R, States D, Claverie JM: Ancient conserved regions in new gene sequences and the protein databases. *Science* 1993, 259:1711–1716.
13. 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.
14. Casari G, Andrade MA, Bork P, Boyle J, Daruvar A, Ouzounis C, *et al.*: Challenging times for bioinformatics. *Nature* 1995, 376:647–648.
15. Riley M: Functions of the gene products of ***Escherichia coli***. *Microbiol Rev* 1993, 57:862–952.
16. Fitch WM: Distinguishing homologous from analogous proteins. *Systematic Zool* 1970, 19: 99–106.
17. Rudd KE: Maps, genes, sequences, and computers: An ***Escherichia coli*** case study. *ASM News* 1992, 59:335–341.
18. Bork P, Ouzounis C, Casari G, Schneider R, Sander C, Dolan M, *et al.*: Exploring the ***Mycoplasma capricolum*** genome: a small bacterium reveals its physiology. *Mol Microbiol* 1995, 16:955–963.
19. Brenner SE, Hubbard T, Murzin A, Chothia C: Gene duplication in ***H. influenzae.*** *Nature* 1995, 378:140.
20. Ochman H, Wilson AC: Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J Mol Evol* 1987, 26:74–86.
21. Krawiec S, Riley M: Organization of the bacterial chromosome. *Microbiol Rev* 1990, 54:502–539.
22. Charlebois RL, St Jean A: Supercoiling and map stability in the bacterial chromosome. *J Mol Evol* 1995, 41:15–23.
23. Riley M, Sanderson KE: Comparative genetics of ***Escherichia coli*** and ***Salmonella typhimurium***. In *The bacterial chromosome*. Edited by Drlica K, Riley M. Washington, DC: American Society for Microbiology; 1990:85–95.
24. Ogasawara N, Moriya S, von Meyenburg K, Hansen FG, Yoshikawa H: Conservation of genes and their organization in the chromosomal replication origin region of ***Bacillus subtilis*** and ***Escherichia coli***. *EMBO J* 1985, 4:3345–3350.
25. Kunisawa T: Identification and chromosomal distribution of DNA sequence segments conserved since divergence of ***Escherichia coli*** and ***Bacillus subtilis***. *J Mol Evol* 1995, 40:585–593.
26. Butler PD, Moxon ER: A physical map of the genome of ***Haemophilus influenzae*** type b. *J Gen Microbiol* 1990, 136:2333–2342.
27. Devine KM: The Bacillus subtilis genome project: aims and progress. *Trends Biotechnol* 1995, 13:210–216.
28. Moszer I, Glaser P, Danchin A: Subtilist: a relational database for the ***Bacillus subtilis*** genome. *Microbiology* 1995, 141:261–268.
29. Holländer R, Hess-Reihse A, Mannheim W: Respiratory quinones in ***Haemophilus***, ***Pasteurella*** and ***Actinobacillus***: pattern, functions and taxonomic evaluation. In *Haemophilus*, *Pasteurella*, and *Actinobacillus* group. Edited by Kilian M, Frederiksen W, Biberstein E. London: Academic Press; 1981:83–97.
30. Musser JM, Kroll JS, Granoff DM, Moxon ER, Brodeur BR, Campos J,*et al.*: Global genetic variation and molecular epidemiology of encapsulated ***Haemophilus influenzae***. *Rev Inf Dis* 1990, 12:75–111.
31. Nimmo HG: The tricarboxylic acid cycle and anaplerotic reactions. In Escherichia coli *and* Salmonella typhimurium. *Cellular and Molecular Biology.* Edited by Neidhardt F, Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE. Washington, DC: American Society for Microbiology; 1987:156–169.
32. Herriott RM, Meyer EY, Vogt M, Modan M: Defined medium for growth of ***Haemophilus influenzae***. *J Bacteriol* 1970, 101:513–516.
33. Woese CR, Stackebrandt E, Ludwig W: What are mycoplasmas: the relationship of tempo and mode in bacterial evolution. *J Mol Evol* 1984, 21:305–316.
34. Römling U, Grothues D, Heuer T,Tümmler B: Physical genome analysis of bacteria. *Electrophoresis* 1992, 13:626–631.
35. Cole ST, Saint Girons I: Bacterial genomics. 1994, *FEMS Microbiol Rev* 14:139–160.
36. Karp P, Paley S: Representations of metabolic knowledge: pathways. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. Edited by Altman R, Brutlag D, Karp P, Lathrop R, Searls D. Menlo Park, CA: AAAI Press; 1994.
37. Gish W, States DJ: Identification of protein-coding regions by sequence similarity searches. *Nature Genet* 1993, 7:205–214.
38. Borodovsky M, McIninch J: GenMark: Parallel gene recognition for both DNA strands. *Comput Chem* 17:123–133.
39. Altschul SF, Boguski MS, Gish W, Wootton JC: Issues in searching molecular sequence databases. *Nature Genet* 1994, 6:119–129.
40. Koonin EV, Tatusov RL, Rudd KE: Protein sequence comparison at a genome scale. *Methods Enzymol*, 1996, 266:295–322.
41. 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.
42. Robison K, Gilbert W, Church GM: Large scale bacterial gene discovery by similarity search. *Nature Genet* 1994, 7:205–214.
43. Borodovsky M, Rudd KE, Koonin EV: Extrinsic and intrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res* 1994b, 22:4756–4767.
44. Krogh A, Mian IS, Haussler D: A hidden Markov model that finds genes in ***E. coli*** DNA. 1994, *Nucleic Acids Res* 22:4768–4778.
45. Pearson W.R. and Lipman DJ: Improved tools for biological sequence comparisons. *Proc Natl Acad Sci USA* 1988, 85:2444–2448.
46. Pearson WR: Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* 1994, 11:635–650.
47. Devereux J, Haeberli P, Smithies O: A comprehensive set of sequence analysis programs for VAX. *Nucleic Acids Res* 1984, 12:387–395.
48. Schuler GD, Altschul SF, Lipman DJ: A workbench for multiple alignment construction and analysis. *Proteins* 1991, 9:180–190.
49. Benson DA, Boguski M, Lipman DJ, Ostell J: GenBank. *Nucleic Acids Res* 1996, 24:1–6.
50. Reitzer LJ, Magasanik B: Ammonia assimilation and the biosynthesis of glutamine, glutamate, aspartate, asparagine, L-alanine, and D-alanine. 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, DC: American Society for Microbiology; 1987:156–169.
51. Liu J, Magasanik B: Activation of the dephosphorylation of nitrogen regulator I-phosphate of ***Escherichia coli***. *J Bacteriol* 1995, 177:926–931.
52. Merrick MJ, Edwards RA: Nitrogen control in bacteria. *Microbiol Rev* 1995, 59:604–622.
53. Lomholt H, van Alphen L, Kilian M: Antigenic variation of immunoglobulin A1 proteases among sequential isolates of ***Haemophilus influenzae*** from healthy children and patients with chronic obstructive pulmonary disease. *Infect Immun* 1993, 61:4575–4581.
54. Jaroschik GP, Hansen EJ: Identification of a new locus involved in expression of ***Haemophilus influenzae*** type b lipo-oligosaccharide. 1994, *Infect Immun* 62:4861–4867.
55. Kupsch EM, Knepper B, Kuroki T, Heyer I, Meyer TF: Variable opacity (opa) outer membrane proteins account for the cell tropisms displayed by ***Neisseria gonorrhoeae*** for human leukocytes and epithelial cells. *EMBO J* 1993, 12:641–650.
56. Weiser JN, Chong STH, Greenberg D, Fong W: Identification and characterization of a cell envelope protein of ***Haemophilus influenzae*** contributing to phase variation in colony opacity and nasopharyngeal colonization. *Mol Microbiol* 1995, 17:555–563.
57. Highlander SK, Wickersham EA, Garza O, Weinstock GM: Expression of the ***Pasteurella haemolytica*** leukotoxin is inhibited by a locus that encodes an ATP-binding cassette homolog. *Infect Immun* 1993, 61:3942–3951.
58. Morris TW, Reed KE, Cronan JE, Jr: Lipoic acid metabolism in ***Escherichia coli***: the ***lplA*** and ***lipB*** genes define redundant pathways for ligation of lipoyl groups to apoprotein. *J Bacteriol* 1995, 177:1–10.