Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*

Ralf Himmelreich, Helga Plagens, Helmut Hilbert+, Berta Reiner and Richard Herrmann*

Zentrum für Molekulare Biologie Heidelberg, Mikrobiologie, Universität Heidelberg, 69120 Heidelberg, Germany

Received December 9, 1996; Revised and Accepted January 8, 1997

ABSTRACT

The sequenced genomes of the two closely related bacteria *Mycoplasma genitalium* and *Mycoplasma pneumoniae* were compared with emphasis on genome organization and coding capacity. All the 470 proposed open reading frames (ORFs) of the smaller *M.genitalium* genome (580 kb) were contained in the larger genome (816 kb) of *M.pneumoniae*. There were some discrepancies in annotation, but inspection of the DNA sequences showed that the corresponding DNA was always present in *M.pneumoniae*. The two genomes could be subdivided into six segments. The order of orthologous genes was well conserved within individual segments but the order of these segments in both bacteria was different. We explain the different organization of the segments by translocation via homologous recombination. The translocations did not disturb the continuous bidirectional course of transcription in both genomes, starting at the proposed origin of replication. The additional 236 kb in *M.pneumoniae,* compared with the *M.genitalium* genome, were coding for 209 proposed ORFs not identified in *M.genitalium*. Of these ORFs, 110 were specific to *M.pneumoniae* exhibiting no significant similarity to *M.genitalium* ORFs, while 76 ORFs were amplifications of ORFs existing mainly as single copies in *M.genitalium*. In addition, 23 ORFs containing a copy of either one of the three repetitive DNA sequences RepMP2/3, RepMP4 and RepMP5 were annotated in *M.pneumoniae* but not in *M.genitalium,* although similar DNA sequences were present. The *M.pneumoniae*-specific genes included a restriction-modification system, two transport systems for carbohydrates, the complete set of three genes coding for the arginine dihydrolase pathway and 14 copies of the repetitive DNA sequence RepMP1 which were part of several different translated genes with unknown function.

INTRODUCTION

Since the first publication of the complete nucleotide sequence of the genome of the bacterium *Haemophilus influenzae* (1), four more sequences of bacterial genomes have been published, namely *Mycoplasma genitalium* (2), *Methanococcus jannaschii* (3), *Synechocystis sp.* (4) and *Mycoplasma pneumoniae* (5), but many more are expected to appear within the next 1–2 years. The large amount of data produced has already initiated several studies on whole genome comparisons, between the genomes from *Escherichia coli*, *Haemophilus influenzae* and *M.genitalium*; bacteria which are only distantly related (6,7).

The complete sequencing of the *M.pneumoniae* genome enabled us to accomplish a comparative analysis of two genomes of closely related organisms: *M.genitalium* and *M.pneumoniae.* Both *M.genitalium* and *M.pneumoniae* are human surface parasites and in nature depend on the host which supplies them with essential nutrients, but both organisms can be propagated *in vitro* in a serum-enriched cell-free medium. Yet, it is much more difficult to cultivate *in vitro M.genitalium*; in fact only a few isolates of *M.genitalium* have been made so far (8).

Mycoplasma pneumoniae is preferentially found in the respiratory tract (9) and *M.genitalium* in the urogenital tract (10), although exceptions are possible. The isolation of *M.genitalium* from the respiratory tract of *M.pneumoniae*-infected patients has been reported (11) and *M.pneumoniae* has recently been isolated from urogenital clinical specimens (12). This shows that both bacteria can exist in the same environment.

Mycoplasma pneumoniae is an established human pathogen causing atypical pneumonia, mostly in children and young adults (13,14). There is accumulating evidence that *M.genitalium* is one of the agents of nongonococcal urethritis in man (14).

Both bacteria share a similar flask-like morphology and show serological cross reactions (14), but they differ in several important features, including a difference in G+C content (8 mol%) and genome size (236 kb), different tissue specificity and pathogenic effects for humans (13,14), and genomic DNA:DNA hybridizations show low values (15). Since the complete nucleotide sequences of both genomes have been established, it has become feasible to compare both bacteria at the nucleotide and protein level.

The genome of *M.genitalium* consists of only 580 070 base pairs (bp) representing the smallest bacterial genome presently known (2). The additional genetic information contained in the larger genome (816 394 bp) of *M.pneumoniae* (5) is probably the key for explaining and understanding the observed biological differences between both species. The comparison of these two closely related bacteria might also provide information for defining essential functions of a self-replicating minimal cell as well as dispensable functions on the way to smaller genomes.

*To whom correspondence should be addressed. Tel: +49 6221 54 68 27; Fax: +49 6221 54 58 93; Email: r.herrmann@mail.zmbh.uni-heidelberg.de

+Present address: Qiagen Gmbh, Postfach 1064, 40719 Hilden, Germany

Table 1. Classification of *M.pneumoniae*-specific ORFs without orthologs in *M.genitalium*

I. Number of *M.pneumoniae* ORFs without significant similarities to those of *M.genitalium*:	**110 (≈120 kb)**
1) *M.pneumoniae*-specific ORFs with similarities to functional assigned ORFs:	25
2) *M.pneumoniae*-specific ORFs with similarities to hypothetical and/or in *M.p.* expressed ORFs:	7
3) *M.pneumoniae*-specific ORFs with pattern based similarities and/or gene amplifications:	25
4) *M.pneumoniae*-specific ORFs containing repetitive DNA element RepMP1:	14
5) *M.pneumoniae*-specific ORFs without significant similarities to ORFs in databases:	37
II. Number of *M.pneumoniae* ORFs (gene amplifications) with similarities to ORFs from *M.genitalium*:	**99 (≈105 kb)**
1) *M.pneumoniae*-specific gene amplifications of ORFs which are only present as single copies in *M.genitalium*:	76
2) *M.pneumoniae*-specific gene amplifications which contain sequences from RepMP2/3, RepMP4 andRepMP5 which were not annotated in *M.genitalium*:	23

Table 2. *Mycoplasma pneumoniae*-specific ORFs with significant similarities to proteins in databases

No	ORF-Name	Annotation / *Family, Function*
***M.pneumoniae*-specific ORFs with similarity to functional assigned ORFs: 26**		
532	H10_orf198	arginine deiminase (arcA); MYCCA
533	H10_orf238	arginine deiminase (arcA); MYCCA
282	H03_orf438	arginine deiminase (arcA); PSEPU
531	H10_orf273o	ornithine carbamoyl transferase (otc1); ECOLI
530	F10_orf309	carbamate kinase (EC 2.7.2.2) (arcC); PSEAE
278	H03_orf351	NADP-dependent alcohol dehydrogenase (adh); THEBR
344	P02_orf242	L-ribulose-5-phosphate 4-epimerase (araD); ECOLI
041	C09_orf600	carnitine palmitoyltransferase II precursor(cpt2); HUMAN
189	E09_orf143V	PTS system mannitol-specific component IIA (EIIA-MTL)(mtlF); STRMU
190	E09_orf364	mannitol-1-phosphate 5-dehydrogenase (EC 1.1.1.17)(mtlD); STRMU
191	E09_orf379	PTS system mannitol-specific component IIA (EIIA-MTL)(mtlA); ECOLI
347	P02_orf159	hypothetical phosphotransferase protein (yjfU) homolog; ECOLI
503	F10_orf326	protein (bcrA) homolog; BACLI *ABC transport*
551	A65_orf306	protein (prrB) homolog, ECOLI *Restriction, modification*
227	C12_orf249	restriction-modification enzyme subunit S1B (hsdS); MYCPU
066	R02_orf335	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
472	H91_orf268	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
490	H91_orf376	type 1 restriction enzyme (hsdR) homolog; ECOLI
491	H91_orf115	(type 1 restriction enzyme (hsdR) homolog; ECOLI)
492	H91_orf206	type 1 restriction enzyme (hsdR) homolog; ECOLI
494	H91_orf330	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
495	H91_orf543	type I restriction enzyme (hsdM); ECOLI
631	GT9_orf238	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
335	P02_orf363V	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
546	H10_orf145L	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
547	H10_orf187V	HsdS1B protein homolog; MYCPU
***M.pneumoniae*-specific ORFs with similarity to hypothetical and/or expressed ORFs: 7**		
187	E09_orf204o	protein P30, MYCPN *expressed, RepMP1 derived* (30)
346	P02_orf660	hypothetical protein (yjfS) homolog; ECOLI
348	P02_orf218	hypothetical protein (yjfV) homolog; ECOLI
349	P02_orf305	hypothetical protein (yjfW) homolog; ECOLI
358	P01_orf341	hypothetical protein (yibD) homolog; ECOLI
416	A05_orf102	hypothetical 13.2 KD protein homolog (ylxM); BACSU
010	E07_orf179	*expressed* (30)

This publication describes the results of our comparative analysis of these two mycoplasma genomes with emphasis on genome organization, coding capacity and gene to gene comparison.

METHODS

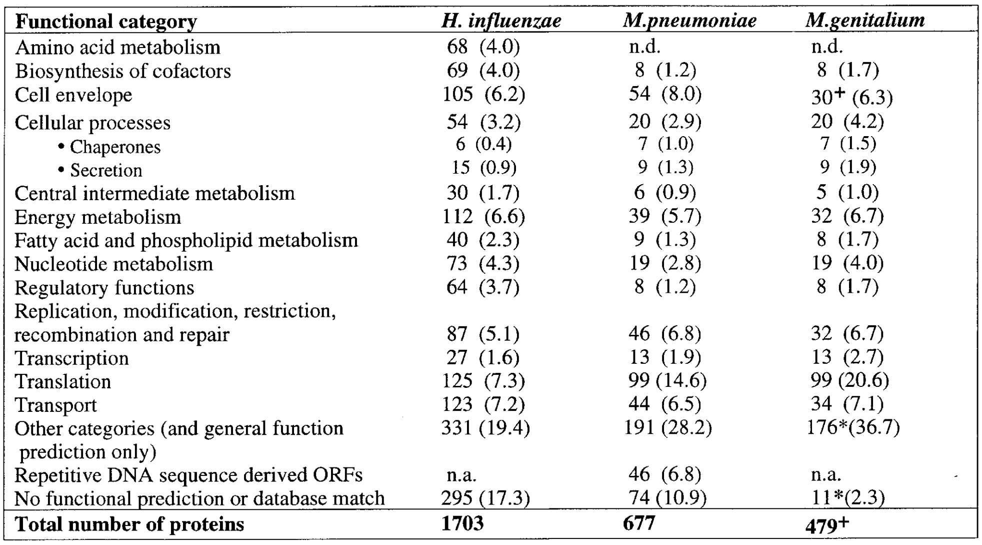
Computer assisted analysis

Analyses were performed with the *HUSAR* (Heidelberg Unix Sequence Analysis Resources) program package release 4.0 at the German Cancer Research Center, Heidelberg, Germany. This package is based on the *GCG* program package version Unix-8.1 of the Genetics Computer Group, Wisconsin.

For the DNA and protein comparisons, the *FASTA* (16) and *BLAST* (17) programs (*BLASTX, BLASTN* and *BLASTP*) were used. Protein sequences were aligned by using either the program GAP (pairwise alignment) based on the algorithm of Needleman and Wunsch (18) or CLUSTAL (19) for multiple alignments. The G+C content was calculated by the program *WINDOW.* Codon usage was assessed with the program *CODONFREQUENCY.*

The annotated sequence data from *M.genitalium* (2) and *M.pneumoniae* (5) serve as the basis for the comparative analyses.

Table 3. Comparison of functional classification of proteins based on sequence similarity



Functional category	***H. influenzae***	***M.pneumoniae***	***M.genitalium***
Amino acid metabolism	68 (4.0)	n.d.	n.d.
Biosynthesis of cofactors	69 (4.0)	8 (1.2)	8 (1.7)
Cell envelope	105 (6.2)	54 (8.0)	30^{+} (6.3)
Cellular processes	54 (3.2)	20 (2.9)	20 (4.2)
• Chaperones	6 (0.4)	7 (1.0)	7 (1.5)
• Secretion	15 (0.9)	9 (1.3)	9 (1.9)
Central intermediate metabolism	30 (1.7)	6 (0.9)	5 (1.0)
Energy metabolism	112 (6.6)	39 (5.7)	32 (6.7)
Fatty acid and phospholipid metabolism	40 (2.3)	9 (1.3)	8 (1.7)
Nucleotide metabolism	73 (4.3)	19 (2.8)	19 (4.0)
Regulatory functions	64 (3.7)	8 (1.2)	8 (1.7)
Replication, modification, restriction, recombination and repair	87 (5.1)	46 (6.8)	32 (6.7)
Transcription	27 (1.6)	13 (1.9)	13 (2.7)
Translation	125 (7.3)	99 (14.6)	99 (20.6)
Transport	123 (7.2)	44 (6.5)	34 (7.1)
Other categories (and general function prediction only)	331 (19.4)	191 (28.2)	176*(36.7)
Repetitive DNA sequence derived ORFs	n.a.	46 (6.8)	n.a.
No functional prediction or database match	295 (17.3)	74 (10.9)	11*(2.3)
Total number of proteins	**1703**	**677**	**479^{+}**

Abbreviations: n.d., not detected; n.a., not annotated; +, according to our calculations; *, see refs. 5, 20, 21 and 50, numbers are different from those published by Fraser *et al.* (2); numbers in brackets, percentage of total ORFs.
For comparison between *H.influenzae* and *M.genitalium* see http://www.ncbi.nlm.nih.gov./cgi-bin/complete_genomes

Corrections to the original paper on *M.genitalium* were only considered if they were published in a scientific journal.

Our published data (5) can also be accessed at the world wide web (www) page (http://www.zmbh.uni-heidelberg.de/M_pneumoniae). The www-pages contain the following additional information: differences to annotations of *M.genitalium* ORFs, missing ORFs in *M.genitalium*, lists of direct length-comparisons of the orthologous *M.genitalium* and *M.pneumoniae* ORFs, degree of identity between orthologous genes/proteins, and theoretical two-dimensional protein maps of both bacteria.

RESULTS AND DISCUSSION

Coding capacity

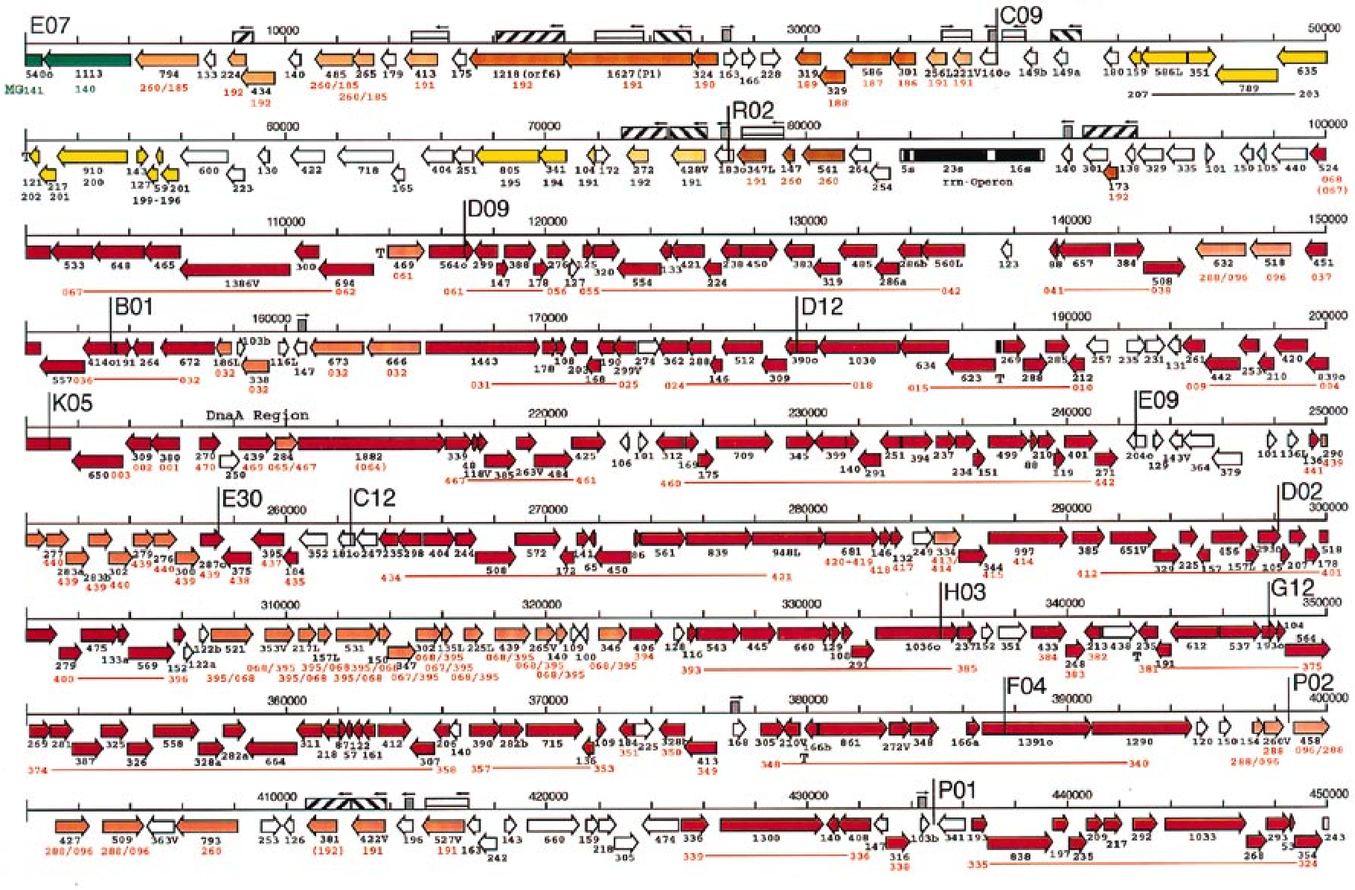
Originally, 470 ORFs were proposed for *M.genitalium* (2) and 677 for *M.pneumoniae* (5). After the publication of the *M.genitalium* sequence a few changes were introduced (20) and a number of functional corrections and new assignments were added [(21,22) see www]. But, independent of these ambiguities a comparison at the ORF and nucleotide sequence level shows clearly that all of the proposed *M.genitalium* ORFs are completely contained in *M.pneumoniae*. There are some discrepancies in annotation and functional prediction, but in all instances inspection of the *M.genitalium* DNA sequences showed that the corresponding DNA sequence was present in *M.pneumoniae*. An obvious example is the ORF MG468 in *M.genitalium* which was assigned in the gene map (2) but originally misnamed in the table on the www pages (http://www.tigr.org/tdb/mdb/mgdb/mgdb.html) or the ortholog to the proposed ORF F10_orf357 of the p65 operon of *M.pneumoniae* was not annotated in *M.genitalium*, but the DNA sequence coding for a protein with a significant similarity is present between the ORFs MG218 and MG219 in *M.genitalium* (23). Further, we proposed for *M.pneumoniae* ORFs containing the repetitive DNA sequences RepMP2/3, RepMP4 and RepMP5 [Table 1, (24)], this was not done for *M.genitalium*, but again at the DNA level the sequences were present. For more detailed information see also our www pages.

The difference in size between the two mycoplasma genomes amounts to 236 kb coding for 209 proposed ORFs [Table 1, www pages, (5)]. Among these ORFs (Table 1) there are two prominent groups: (i) ORFs showing similarities to functionally assigned proteins only present in *M.pneumoniae* and gene amplifications thereof (Tables 1 and 2); these ORFs may help to explain the biological differences between *M.pneumoniae* and *M.genitalium.* (ii) ORFs which are amplified and are also present in *M.genitalium* but with a smaller copy number; these ORFs contribute to the difference in genome size but not as much to the repertoire of new functions. The relatively small number of *M.pneumoniae*-specific proposed ORFs with significant similarities to proteins with known functions is summarized in Table 2.

The most important functions are: (i) an hsd-type restriction-modification (R-M) system; (ii) two phosphoenolpyruvate:carbohydrate phosphotransferase systems (PTS), one with a predicted transport specificity for mannitol and the other with an unknown specificity; (iii) an NADP-dependent alcohol dehydrogenase; (iv) the complete set of the enzymes involved in the arginine dihydrolase pathway consisting of arginine deiminase, ornithine carbamoyltransferase and carbamate kinase.

Also included here are the predicted ORFs which contain sequences of the repetitive DNA sequence RepMP1, first described by Wenzel and Herrmann (25).

The proposed restriction-modification (R-M) system shares the highest similarity with the type I restriction-modification system from *E.coli* (26). This system consists of an enzyme complex with three different subunits: R (1033 amino acids) for restriction, M (520 amino acids) for modification and S (410 amino acids) for sequence specificity. This type of R-M system has been already identified in *Mycoplasma pulmonis* (27). The comparison of the



R subunits size from *M.pulmonis* and *M.pneumoniae* and also of the orthologs of *E.coli* and *H.influenzae*, strongly suggests that the three ORFs H91_orf376, H91_orf115 and H91_orf206, corresponding to the N-terminal, the middle, and the C-terminal part of a complete R subunit are the result of frameshifts. Since the repeated sequence analysis with PCR-amplified genomic *M.pneumoniae* confirmed our original DNA sequence, we assume that our *M.pneumoniae* strain carries frameshift mutations in the hsdR gene. We also observed a number of gene amplifications of the hsdS gene coding for S subunits varying in length between 363 and 145 amino acids. Since the orthologs in *E.coli*, *H.influenzae* or *M.pulmonis* are ~400 amino acids long, we assume that the shorter ORFs in *M.pneumoniae* are truncated, inactive forms.

The two additional PTS systems found in *M.pneumoniae* should be advantageous for a cell which depends on the import of many precursors (28). The mannitol specificity needs to be taken with some reservation, since the highest score in database similarity searches does not give definite results concerning specificity of the substrate to be transported. The advantage of a mannitol transport system for *M.pneumoniae* can only be evaluated if we know more about the prevalence of mannitol in the human respiratory tract.

Mycoplasma pneumoniae codes for an alcohol dehydrogenase (adh). This enzyme reduces acetaldehyde to ethanol in an NADPH-dependent reaction. Acetaldehyde and glyceraldehyde-3-phosphate are produced by the deoxyribose-phosphate-aldolase from 2-deoxy-ribose-5-phosphate. Besides removing the toxic acetyldehyde from the cell the alcohol dehydrogenase activity is one way to regulate the NADPH:NADP$^+$ equilibrium in the bacterium (29).

Mycoplasma pneumoniae could use, in principle, the arginine dihydrolase pathway (ADI) to generate ATP from arginine. The pathway requires the three enzymes arginine deiminase, ornithine carbamoyltransferase and carbamate kinase. Citrulline, ornithine and carbamyl phosphate are the intermediates in the ADI pathway (29). The importance of this pathway for *M.pneumoniae* and other arginine-fermenting mycoplasmas as an energy source has not really been evaluated. Based on the now available detailed genetic information, it will be possible to construct and select mutants defective in one of the enzymes of the ADI pathway and analyze the consequences of these defects. This approach can be applied to all the *M.pneumoniae* genes which are absent in *M.genitalium*, since these genes might be advantageous but may not be essential for growth of *M.pneumoniae*; at least under laboratory conditions.

The group of functionally assigned gene products included proposed ORFs containing RepMP1 sequences in their coding region. We do not know the function of these proteins, but it could be shown that antibodies to fusion proteins containing RepMP1 derived protein sequences reacted positively in Western blots of electrophoresed *M.pneumoniae* protein extracts. Several proteins of different size were identified (30) indicating that more than one of the proposed ORFs was indeed expressed. In the case of these RepMP1-containing genes, the mutagenesis approach is not useful, since the complete genome sequence revealed 14 copies of these

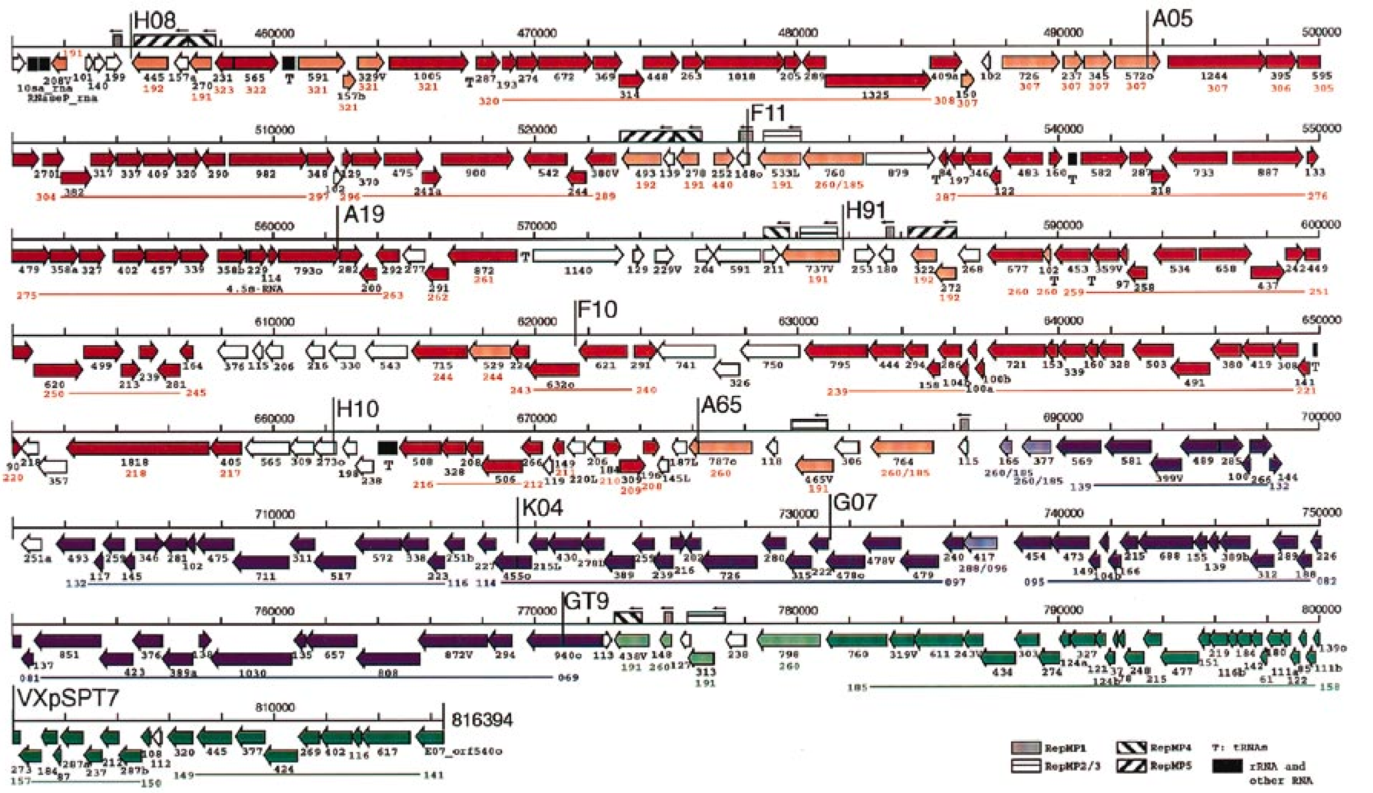


Figure 1. The complete gene map of the *M.pneumoniae* genome and the alignment of the corresponding *M.genitalium* ORFs. The thick, coloured arrows represent the position, size and direction of transcription of the proposed ORFs of *M.pneumoniae*. The official name of an ORF can be deduced from the cosmid name above the horizontal scale-line and the black number directly below the thick arrows, e.g. the correct ORF name represented by the first complete green arrow is E07_orf1113. For more details see ref. 5. The second number in colour below the black number indicates the corresponding ORF number (e.g. MG141) of *M.genitalium*, except for the yellow segment, where the *M.genitalium* ORFs were also drawn in black for better visibility. Throughout the genome map the ORFs of *M.genitalium* are oriented in the same direction as the ORFs of *M.pneumoniae*. The genes of the *M.genitalium* map are named in numerical order, starting with MG001 localized in this comparative map around nucleotide position 206 000 (dnaA region). Any divergence in the order of orthologous genes can be recognized by the discontinuity of the numbers of the *M.genitalium* ORFs. In all cases where the consecutive order of *M.genitalium* orthologs is conserved, coloured thin horizontal lines connect the first and last gene in the particular genome region. The order of genes is conserved in both bacteria within six DNA segments each marked by the distinct strong color. Interspaced *M.pneumoniae* ORFs without significant similarities to *M.genitalium* are shown by thick arrows in white (Fig. 1; Tables 1 and 2). *Mycoplasma pneumoniae*-specific ORFs representing amplification of *M.genitalium* homologs (Table 1) were indicated by thick arrows in light colors. Each of these six conserved DNA segments of *M.pneumoniae* (strong colors) is bordered by repetitive DNA sequences. Rectangles with various patterns above the scale-line indicate size, position and direction of transcription for repetitive DNA sequence-derived ORFs. The corresponding repetitive sequences of *M.genitalium* (MgPa repeats) are shown in Figure 2.

genes, some are truncated, so that it will be difficult to find a mutant phenotype.

The proposed lipoproteins are the best examples for the amplification of genes which occur frequently in single copies in *M.genitalium* but in multiple copies in *M.pneumoniae*. The characteristic features of these lipoproteins have been described (5,31). *Mycoplasma genitalium* codes for only 21 lipoproteins while *M.pneumoniae* codes for 46 lipoproteins. These lipoproteins are products of several gene families which also include 20 genes with sequence similarities to lipoproteins but without the functional lipoprotein signal peptides. The difference in the number of lipoproteins is mainly caused by gene amplifications, for instance see the eight proposed ORFs (Fig. 1) located between nucleotide positions 249 627 and 256 463 (cosmid pcosMPE09, light red arrows) which share significant similarities with the two *M.genitalium* ORFs MG439 and MG440. Other examples are the 13 proposed ORFs located between *M.pneumoniae* nucleotide positions 306 862 and 320 524 (pcosMPD02) which share similarities with the *M.genitalium* ORFs MG067, MG068 and MG395.

Apart from the described differences in coding capacity, a remarkable conformity exists in both bacteria concerning the composition, number and similarities of the individual components of the systems involved in basic processes like DNA replication, transcription, translation, regulation of gene expression, protein secretion and energy conservation (Table 3). The observation that the number of genes for certain functional categories is the same in both bacteria, indicates that here the minimal set of genes is already assembled and a further reduction would probably be very disadvantageous to the cells. The similarity between both bacteria is also shown by several genes or functions which are present or absent in both bacteria. Among the unidentified products are some which have to be present because they catalyze singular essential reactions in otherwise complete pathways (Table 4), for instance the nucleoside diphosphate kinase (ndk), the key enzyme for the conversion from NDP to NTP, could not be identified in

Table 4. Common features of *M.pneumoniae* and *M.genitalium*

gene name, catagory of function	proposed ORF		comments/references
	MP	MG	
DNA replication/repair:			
DNA polymerase III, dnaE, α subunit	A19_orf 872	MG261	highest similarity to Gram-negative bacteria, no 3'-5' exonuclease (5, 44)
α subunit, dnaE	B01_orf1443	MG031	highest similarity to Gram-posititve bacteria, 3'-5' exonuclease
DNA polymerase, polA,	A19_orf291	MG262	no polymerase specific domain, 5'-3' exonuclease domain
RNaseH, rnhA	-	-	not identified (45)*
Mismatch-repair system, mutS, mutL, mutH	-	-	not identified
Transcription			
RNA polymerase, sigma factor, sigA	H91_orf499	MG249	one sigma factor only
Transcription termination factor Rho	-	-	not identified
Translation:			
Ribosomal protein, rpl, rps	50 ribosomal proteins		rpS1 not identified
tRNA synthetases	19 different synthetases		glutaminyl-tRNA synthetase not identified, both proteins are missing in other Gram-positive bacteria
transfer RNAs	33 tRNAs		identical set of tRNAs (2, 46), UGA codon is read as tryptophan
10Sa small stable RNA	+	+	proposed function in transtranslation (47)
Peptide chain release factor RF-2, prfB	-	-	RF-2 not identified, RF-2 recognizes UGA, RF-2 has to be deleted to avoid premature translation stop
Protein secretion:			
the machinery for protein secretion is rather complex and consists in *E.coli* of 12 components, of which only a fraction is found in mycoplasmas	A05_orf348 (ftsY) D09_orf450 (ffh) + 4.5SRNA G07_orf808 (secA) GT9_orf477 (secY) F10_orf444 (tig) A05_orf595 (dnaK)	MG297 MG048 + 4.5S RNA MG072 MG170 MG238 MG305	Simplified version, some of the channel-forming or associated proteins (secD,E,F,G) and secB were not identified (48) Remarkable is the failure to detect signal peptidaseI (SPaseI)
Transport systems:			
ABC transporter for oligopeptides 1 substrate binding domain oppA	-	-	The substrate binding protein was not identified by similariy search (49)*, this protein seems to be essential unless the bacterium receives the substrate by close contact from host directly
Cell surface, lipoproteins:			
after cleavage of the signal-peptide a lipid modified Cys is the first amino acid. For modification of this Cys in *E.coli* a transacylase is required.	-	-	A transacylase linking a third fatty acid to the N terminal Cys was not identified; this type of acylation has not yet been shown to occur in mycoplasmas
Fatty acid, phospho- and glycolipid metabolism:			
About 10 genes are required for synthesis of the experimentally identified phospho- and glyco-lipids of *M.pneumoniae*	A65 orf272 (pgsA) E30 orf395 (cdsA) H10 orf266 (plsB)	MG114 MG437 MG212	Only three genes were identified by DNA sequence analysis in both bacteria, e.g. glucosyl-transferases are missing
Nucleotide synthesis:			
Purine and pyrimidine salvage pathway	+/-	+ /-	Salvage pathway complete but nucleoside diphosphate kinase (ndk) not identified (41)*, (45) *, the reaction is essential
Energy metabolism:			
Glycolysis,	+	+	Glycolysis is complete in both mycoplasmas, pentose phosphate pathway only truncated, no cytochromes. See the difference between *M.pneumoniae* and *M.genitalium*: arginine dihydrolase pathway for generation of ATP only in *M.pneumoniae*.
pentose phosphate pathway,	+/-	+/-	
tricarboxylic acid cycle,	-	-	
cytochromes	-	-	

References marked with asterisks: the authors proposed ORFs in *M.genitalium* which should code for the 'not identified' proteins. Since in our opinion the evidence is not convincing enough we consider these functions as still not identified.

either bacteria, but the enzymes which convert bases or nucleosides into NDPs were present. In other instances, some genes could not be identified which should be there based on physiological requirements or experimental evidence, for instance the enzymes catalase, superoxide peroxidase, dismutase or genes involved in motility (32). The most prominent examples for this type of common feature between *M.pneumoniae* and *M.genitalium* are summarized in Table 4. A detailed comparison of all the orthologs identified in *M.pneumoniae* and *M.genitalium* has been already published (5). Finally, the high degree of agreement of the sequence analysis and annotations for both genomes confirms the results of each analysis.

Degree of similarity between orthologs

The G+C contents of *M.pneumoniae* (40 mol%) and *M.genitalium* (32 mol%) vary by 8 mol%. Plotting the G+C content of the first, second and third position of all the codons used for the proposed ORFs against the G+C content of the genomes of *M.pneumoniae* and *M.genitalium* reveals that the third codon position is, with almost 19 mol% difference, the most variable (*M.pneumoniae*, 41.9 mol%; *M.genitalium*, 23.1 mol%), the first position is the next variable (*M.pneumoniae*, 46.9 mol%; *M.genitalium*, 41.6 mol%), and the second position is the most constrained (*M.pneumoniae*, 33.4 mol%; *M.genitalium*, 30.0 mol%). Changing the third positions in ~32 000 codons (corresponding to the 19 mol% difference) of a total of 170 400 codons from A or T to G or C would already cause a 5.6 mol% increase in the G+C content of the *M.genitalium* genome without affecting the amino acid composition. This indicates that the difference in G+C content is mainly caused by the difference in the third position of the codons (Table 5).

We still cannot explain why only *M.pneumoniae* has this relatively high G+C content among the *Mollicutes* (33). A nucleotide bias in the mutational mechanism (34) might be the cause. An

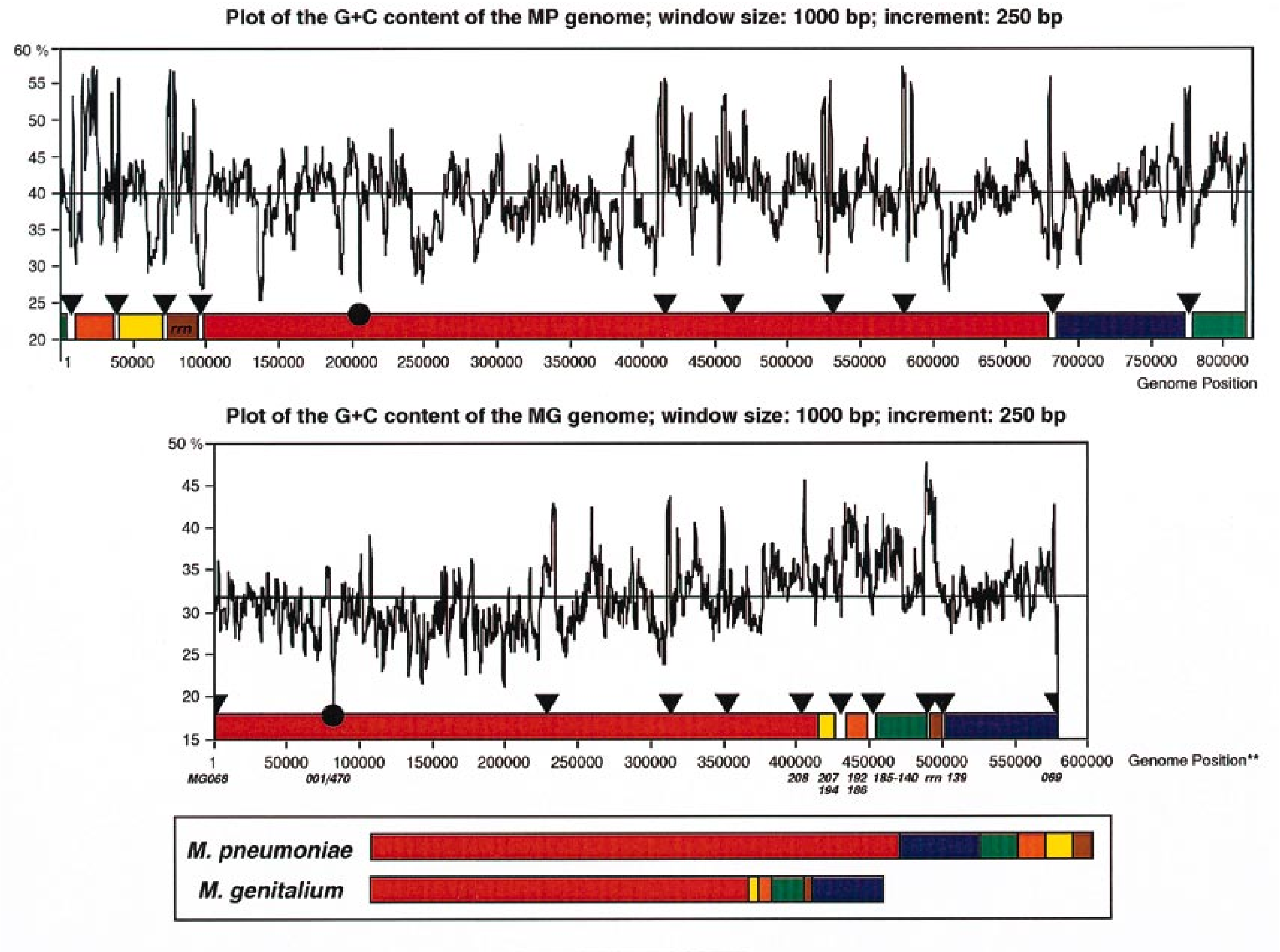
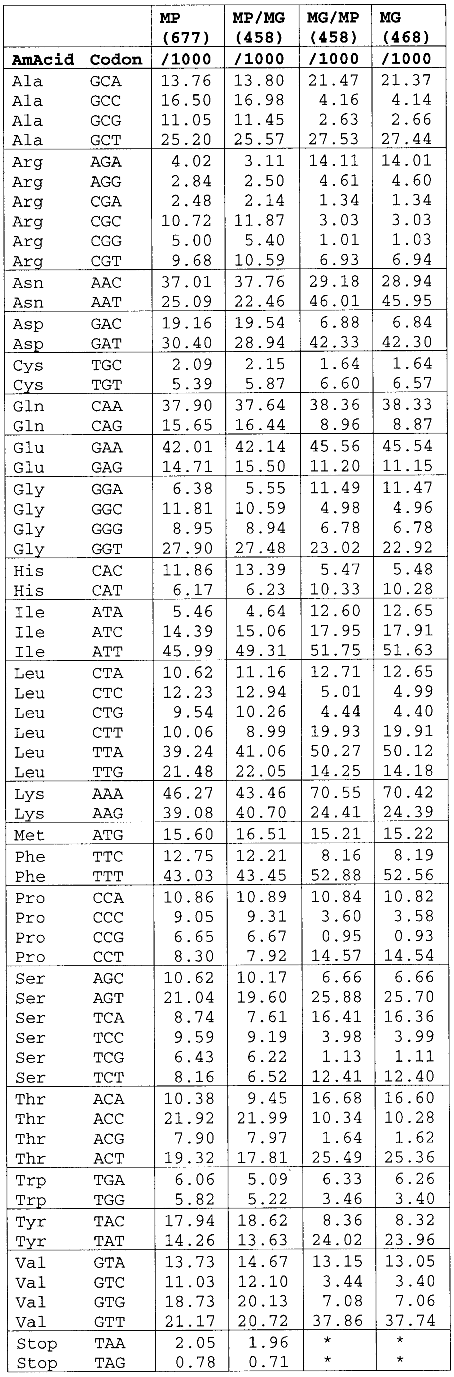


Figure 2. Low resolution G+C-plot of the complete genomes of *M.pneumoniae* and *M.genitalium*. The G+C-content was calculated within a windowsize of 1000 bp and an increment of each 250 bp. The positions of the segments of conserved gene order are indicated as coloured bars. The black dots mark the position of the putative origins of replication and rrn marks the position of the rRNA operon. The black triangles indicate repetitive DNA sequences. The gaps between the segments with black inverted triangles above mark a proposed site of translocation. The start point of the *M.genitalium* genome sequence was changed to MG068 and the sequence was reverse complemented. The *M.genitalium* ORF names (e.g. MG068) below the *M.genitalium* G+C plot indicate the new orientation of the genomic sequence and the first and last ORF of each segment e.g. the green segment reaches from MG140 to MG185 (see also Fig. 1). Below the plots is a schematic illustration of the different order of the genomic segments in these two bacteria.

unbalanced supply of the essential external precursors for nucleic acid synthesis and repair seems unlikely since both bacteria can grow in the same enviroment. On comparing the similarity of orthologs in *M.genitalium* and *M.pneumoniae* measured as amino acid identities, we found a large spectrum reaching from 95% to only 20% identity (www pages). The average was ~67% identity. The highest scores were found for housekeeping proteins like ribosomal proteins, elongation factors or subunits of the F_0F_1 ATPase. Interesting among these is G12_orf109 (MG353) with a high identity score but no functional assignment (Table 6). The high score suggests that this is a protein with a function also well conserved among other bacteria. Proteins with low identity scores were the components of the cytoskeleton and the lipoproteins which are mostly surface-exposed and play a role in antigenic variation in other mycoplasmas. In analogy to other organisms, antigenic variation could be accomplished by differential expression of members of lipoprotein gene families (35). In general, functionally assigned proteins, which occur in many bacteria, showed high identities and most of the functionally unassigned proteins had the lower identity values. About 90% of orthologous genes were similar in size in both mycoplasmas. Only 52 proposed ORFs exhibited significant size deviations. The differences were frequently due to different localization of the start codon (ATG, TTG, GTG) causing an extension of the N-terminus mainly in conserved proteins (Table 6, atpD gene) and gaps in genes with lower identity scores such as the cytadherence accessory genes hmw1 and hmw3 (Table 6).

The G+C content of the two genomes also influences the total content of amino acids assigned by codons with only A and/or T (Phe, Ile, Met, Tyr, Asn, Lys) or G and/or C (Pro, Ala, Arg, Gly) in the first and second position. We calculated for the 458 orthologous proteins (Table 5) the following values for the sum of AT codons: 31.78 mol% for *M.pneumoniae* and 36.1 mol% for *M.genitalium*, and for the GC codons, 18.52 mol% for *M.pneumoniae* and 14.43 mol% for *M.genitalium*. If we took for these calculations *M.pneumoniae* ORFs with a G+C content <35 mol% (5) a significant

Table 5. Codon usage by *M.pneumoniae* (MP) and *M.genitalium* (MG)



		MP (677)	MP/MG (458)	MG/MP (458)	MG (468)
AmAcid	**Codon**	**/1000**	**/1000**	**/1000**	**/1000**
Ala	GCA	13.76	13.80	21.47	21.37
Ala	GCC	16.50	16.98	4.16	4.14
Ala	GCG	11.05	11.45	2.63	2.66
Ala	GCT	25.20	25.57	27.53	27.44
Arg	AGA	4.02	3.11	14.11	14.01
Arg	AGG	2.84	2.50	4.61	4.60
Arg	CGA	2.48	2.14	1.34	1.34
Arg	CGC	10.72	11.87	3.03	3.03
Arg	CGG	5.00	5.40	1.01	1.03
Arg	CGT	9.68	10.59	6.93	6.94
Asn	AAC	37.01	37.76	29.18	28.94
Asn	AAT	25.09	22.46	46.01	45.95
Asp	GAC	19.16	19.54	6.88	6.84
Asp	GAT	30.40	28.94	42.33	42.30
Cys	TGC	2.09	2.15	1.64	1.64
Cys	TGT	5.39	5.87	6.60	6.57
Gln	CAA	37.90	37.64	38.36	38.33
Gln	CAG	15.65	16.44	8.96	8.87
Glu	GAA	42.01	42.14	45.56	45.54
Glu	GAG	14.71	15.50	11.20	11.15
Gly	GGA	6.38	5.55	11.49	11.47
Gly	GGC	11.81	10.59	4.98	4.96
Gly	GGG	8.95	8.94	6.78	6.78
Gly	GGT	27.90	27.48	23.02	22.92
His	CAC	11.86	13.39	5.47	5.48
His	CAT	6.17	6.23	10.33	10.28
Ile	ATA	5.46	4.64	12.60	12.65
Ile	ATC	14.39	15.06	17.95	17.91
Ile	ATT	45.99	49.31	51.75	51.63
Leu	CTA	10.62	11.16	12.71	12.65
Leu	CTC	12.23	12.94	5.01	4.99
Leu	CTG	9.54	10.26	4.44	4.40
Leu	CTT	10.06	8.99	19.93	19.91
Leu	TTA	39.24	41.06	50.27	50.12
Leu	TTG	21.48	22.05	14.25	14.18
Lys	AAA	46.27	43.46	70.55	70.42
Lys	AAG	39.08	40.70	24.41	24.39
Met	ATG	15.60	16.51	15.21	15.22
Phe	TTC	12.75	12.21	8.16	8.19
Phe	TTT	43.03	43.45	52.88	52.56
Pro	CCA	10.86	10.89	10.84	10.82
Pro	CCC	9.05	9.31	3.60	3.58
Pro	CCG	6.65	6.67	0.95	0.93
Pro	CCT	8.30	7.92	14.57	14.54
Ser	AGC	10.62	10.17	6.66	6.66
Ser	AGT	21.04	19.60	25.88	25.70
Ser	TCA	8.74	7.61	16.41	16.36
Ser	TCC	9.59	9.19	3.98	3.99
Ser	TCG	6.43	6.22	1.13	1.11
Ser	TCT	8.16	6.52	12.41	12.40
Thr	ACA	10.38	9.45	16.68	16.60
Thr	ACC	21.92	21.99	10.34	10.28
Thr	ACG	7.90	7.97	1.64	1.62
Thr	ACT	19.32	17.81	25.49	25.36
Trp	TGA	6.06	5.09	6.33	6.26
Trp	TGG	5.82	5.22	3.46	3.40
Tyr	TAC	17.94	18.62	8.36	8.32
Tyr	TAT	14.26	13.63	24.02	23.96
Val	GTA	13.73	14.67	13.15	13.05
Val	GTC	11.03	12.10	3.44	3.40
Val	GTG	18.73	20.13	7.08	7.06
Val	GTT	21.17	20.72	37.86	37.74
Stop	TAA	2.05	1.96	*	*
Stop	TAG	0.78	0.71	*	*

All values are calculated in thousands. The ‘MP/MG’ column contains the 458 *M.pneumoniae* ORFs with similarity to *M.genitalium* and the ‘MG/MG’ column represents the codon usage of the *M.genitalium* ORFs with similarity to *M.pneumoniae*. We compare here only ORFs with identical annotation.

*The stop codons are not included in the calculation.

increase from 31.44 to 36.9 mol% was observed for AT codons and a decrease from 18.71 to 13.3 mol% for GC codons. These results support the findings of Sueoka (37) and many others (for review see ref. 38) that a relationship exists between the G+C content of a DNA and the amino acid composition.

Genome organization

One of the main conclusions derived from comparative analyses of bacterial genome organizations was that gene order is not conserved. Even the proposed origins of replication, normally located around the dnaA gene (39) are not uniform. In contrast the genomes of *M.pneumoniae* and *M.genitalium* represent an example of two different species with a very conserved gene order (Fig. 1) and dnaA regions (36), revealing the same arrangement of genes and 69.4% identity at the nucleotide level from nucleotide position 196 519 to 217 156 (Fig. 1; 36).

The *M.pneumoniae* and *M.genitalium* genomes can be subdivided into six genomic segments (Figs 1 and 2). Within these segments, the order of genes was conserved with the only exception that additional genes were interspaced in the larger genome of *M.pneumoniae* (indicated by the white and light-coloured arrows in Fig. 1), but the order of the six fragments is different in both genomes. A closer inspection of the regions bordering these segments showed that in each case, one or more of the repetitive sequences RepMP1, RepMP2/3, RepMP4 and RepMP5 were present in *M.pneumoniae* (Fig. 3) and that relics of these sequences were still visible in *M.genitalium*. They were named MgPa repeats (2,40) and revealed strong sequence similarities to the above mentioned repetitive DNA sequences from *M.pneumoniae*, except for RepMP1, which could not be identified in *M.genitalium*. We concluded therefore that the reorganization of the *M.genitalium* genome took place by translocation of entire segments via homologous recombination between the repetitive DNA sequences. This conclusion is supported by the presence of the recA gene in both mycoplasmas. The proposed sites of translocation in *M.genitalium* were between the ORFs MG068/069, MG139/140, MG185/186, MG192/193 and MG207/208. Only between MG207 and MG208 there is no MgPa repeat (Fig. 2; 2).

Except for RepMP1, the repetitive DNA sequences of both genomes are characterized by their high G+C content, ~55% in *M.pneumoniae* and 43% in *M.genitalium*. Most of the peaks reaching above 50 mol% for *M.pneumoniae* and 40 mol% for *M.genitalium* in the plot of the G+C content represent the repetitive DNA sequences or the P1 gene and the ORF6 gene of the P1 operon (Figs 2 and 3). They contribute also to the uneven G+C distribution on the genome (Fig. 2).

It has been pointed out that in both the *M.genitalium* (2) and *M.pneumoniae* genomes (5) a remarkable uniformity of the direction of transcription is conserved (Fig. 4). We see a frequent switching of transcription only between nucleotide positions 520 000 and 608 000 on the *M.pneumoniae* map and between ORF MG291 and MG247 on the *M.genitalium* map (Fig. 1). In all other genome regions only ~15% of the proposed ORFs are transcribed against the general direction of transcription. The observed translocation of DNA segments which took place in *M.genitalium* did not change this uniform transcription pattern. One can see that in both genomes, the red-coloured region between nucleotide positions 100 000 and 675 000 of the *M.pneumoniae* genome and between ORFs MG068 and MG208 of the *M.genitalium* genome (Figs 1 and 2) has not been rearranged by translocations although the repetitive DNA sequences, the hypothetical sites for homologous recombination, were present as indicated for *M.pneumoniae* in Figures 1 and

Table 6. Sequence identities of selected genes/proteins

Functional category / gene name	*Mycoplasma pneumoniae*	*Mycoplasma genitalium*	Identity %	
	ORF name / length bp	**ORF name / length bp**	**nucleotides**	**amino acids**
Cytadherence accessory proteins				
hmw1	H08_orf1018 / 3054	MG312 / 3417	55.3	37.0
hmw2	F10_orf1818 / 5454	MG218 / 5415	62.7	58.0
hmw3	H08_orf672 / 2016	MG317 / 1797	48.3	37.1
Lipoproteins				
	E07_orf301 / 903	MG186 / 750	52.4	45.2
	P02_orf1300 / 3900	MG338 / 3813	59.0	52.5
	GT9_orf760 / 2280	MG185 / 2103	60.3	55.4
	D02_orf521 / 1563	MG395 / 1572	57.8	48.1
	G07_orf454 / 1362	MG095 / 1194	58.8	55.3
glpD	D09_orf384 / 1152	MG039 / 1152	64.0	62.5
Glycolysis				
pgk	A05_orf409 / 1227	MG300 / 1248	66.2	69.9
pyk	H10_orf508 / 1524	MG216 / 1524	71.2	78.2
gap	A05_orf337 / 1011	MG301 / 1011	72.2	82.2
tsr	B01_orf288 / 864	MG023 / 864	72.8	77.8
cno	C12_orf456 / 1368	MG407 / 1374	73.7	77.9
pgiB	K04_orf430 / 1290	MG111 / 1299	68.1	69.3
pgm	C12_orf508 / 1524	MG430 / 1521	67.2	70.8
F_0F_1 ATPase				
atpI	C12_orf157L / 471	MG406 / 339	62.8	56.6
atpB	C12_orf2930 / 879	Mg405 / 876	70.5	71.2
atpE	E02_orf105 / 315	MG404 / 306	71.2	76.5
atpF	D02_orf207 / 621	MG403 / 624	68.1	70.5
atpH	D02_orf178 / 534	MG402 / 528	70.5	72.2
atpA	D02_orf518 / 1554	Mg401 / 1554	74.3	86.3
atpG	D02_orf279 / 837	MG400 / 837	72.0	73.5
atpD	D02_orf475 / 1425	MG399 / 1146	77.2*	91.9*
atpC	D02_orf133 / 399	MG398 / 399	70.7	77.4
Selected examples with more than 90% identity (amino acids)				
tuf	K05_orf394 / 1182	MG451 / 1182	84.0	96.7
rpS12	G07_orf139 / 417	MG087 / 417	77.7	95.7
rpL14	GT9_orf122 / 366	MG161 / 366	77.9	95.1
rpoC	F04_orf1290 / 3870	MG340 / 3876	76.8	92.4
dnaK	A05_orf595 / 1785	MG305 / 1785	79.7	92.4
gyrB	K05_orf650 / 1950	MG003 / 1950	77.3	90.6
	G12_orf109 / 327	MG353 / 327	79.2	91.7
Ribosomal RNA				
rrs	16S rRNA / 1522	16S rRNA / 1522	98.1	-

The comparative analysis was done with the program GAP.
*These values consider only the overlapping regions.

2. The corresponding MgPa repetitive DNA sequences are located between the following pairs of ORFs: MG 226/227, MG260/261, MG287/288 and MG339/340 (Figs 1 and 2; 2). We explain this observed genomic stability by a selection pressure which tolerates translocations only when the transcription of the genes on the translocated segment does not interfere with the general direction of transcription of the genomic environment. Most translocations of DNA segments from outside of the red-coloured area into this region, by homologous recombination between the repetitive DNA sequences bordering the DNA segments, would interrupt the direction of transcription. The proposed mechanism of translocation permits only insertion in such an orientation, that the genes on this segment are transcribed in the opposite direction with respect to the general orientation of transcription of the cell. Such a conserved continuous bidirectional direction of transcription has not been seen in *H.influenzae* (1), *Methanococcus jannaschii* (3) or in *Synechocystis sp.* (4); either the genes were preferentially transcribed from one DNA strand (*H.influenzae*) or frequent switching of strands occurred (*M.jannaschii*). Assuming a bidirectional modus of DNA replication, it might be possible that transcription and DNA replication are coupled. This could be a way of regulating gene expression. Any reversion of this directionality might be disadvantageous to these bacterial cells.

When we analyzed the sites where, compared with *M.genitalium*, the specific additional proposed ORFs of *M.pneumoniae* were located, we found a remarkable conformity because they were mapped in regions of low G+C content. Examples are given in Figure 3, roughly from nucleotide positions 60 000 to 67 000 and from 94 000 to 99 000. All the proposed ORFs represented by white arrows are *M.pneumoniae*-specific. This phenomenon, the correlation between additional *M.pneumoniae*-specific genes and a relative low G+C content, appears throughout the entire genome (see www pages). In some instances, the segments with a low G+C content also have a lower coding density (Figs 1 and 3); such as the proposed origin of replication (Fig. 3; 36). Presently, we have no explanation for this observation. To understand these findings one has to wait until the direct ancestor of *M.pneumoniae* has been identified.

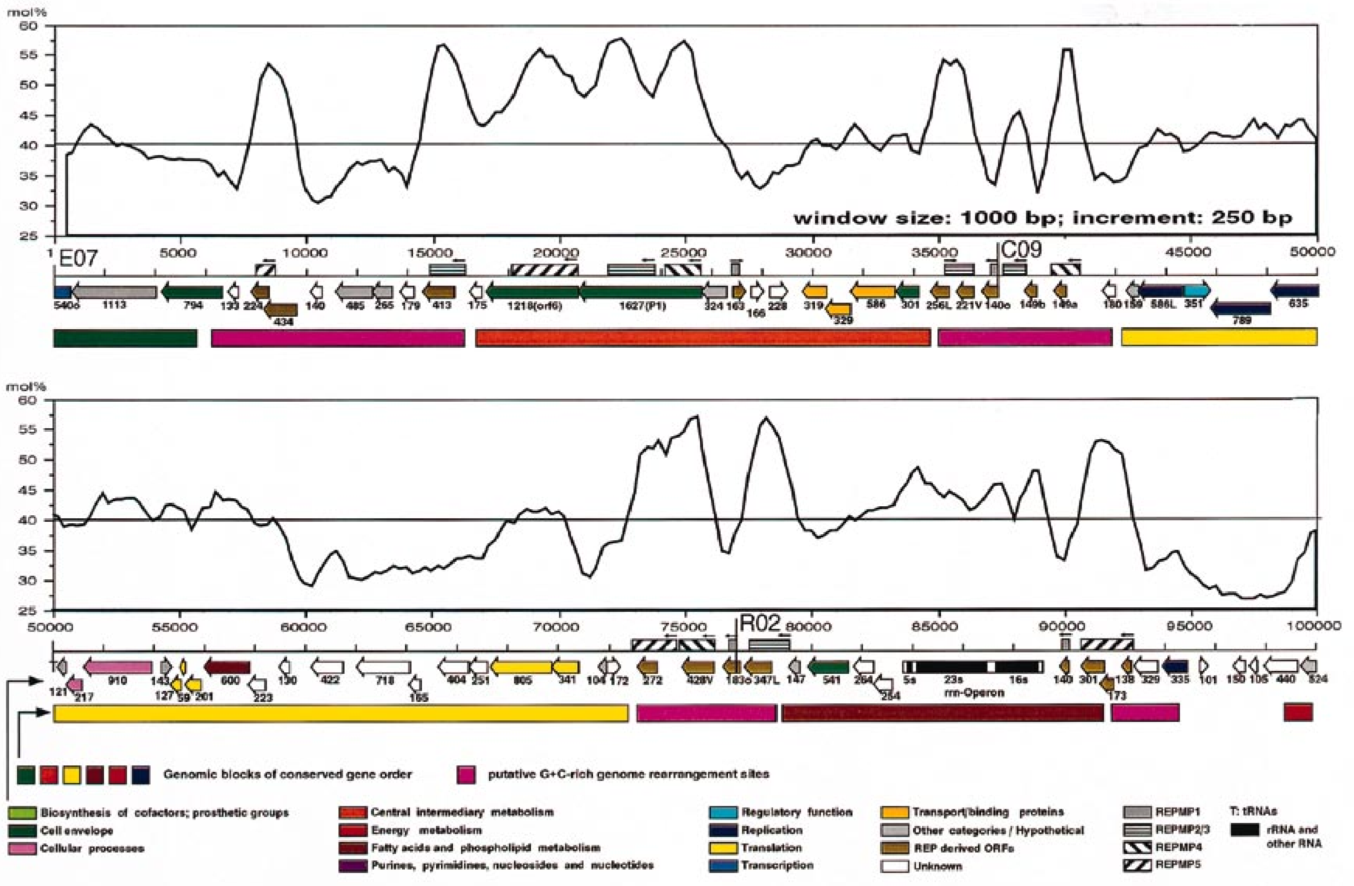


Figure 3. High resolution G+C plot combined with the gene map of the first 100 kb of the *M.pneumoniae* genome (5). This figure illustrates the proposed sites of genomic recombination and their correlation to the repetitive DNA sequences. The position of the repetitive DNA sequences can easily be recognized by an increase in the G+C-content. The segments of conserved gene order are displayed as coloured bars. The magenta coloured bars indicate the areas of putative genomic recombinations. The thick coloured arrows indicate the functional categories of the proposed ORFs from *M.pneumoniae* (5).

The difference in genome size between *M.pneumoniae* and *M.genitalium* can be explained by two processes: Size reduction of the *M.genitalium* genome by deleting genes and by an increase in size of the *M.pneumoniae* genome by amplification of existing genes. The best example for the latter event is the significantly higher number of lipoproteins in *M.pneumoniae*, many of which probably arose via gene amplification. If we subtract the length of the genes amplified from the genome of *M.pneumoniae* (Table 1), we end up with a *M.pneumoniae* genome of ~710 kb. Unless the gene amplifications turn out to code for important genetic information for *M.pneumoniae,* it appears that the larger genome does not code for as many more functions as would be anticipated from its higher DNA content. Following the same argument the genome size of *M.genitalium* could also be further reduced, e.g. to ~560 000 bp by deleting MgPa repeats.

The minimal cell

Mycoplasma genitalium has the smallest presently known genome, it is therefore the most promising candidate for defining and constructing a minimal cell by genetic manipulations e.g. inactivating or deleting genes. More importantly, the proposed or constructed minimal cell can be experimentally tested for its ability to survive and reduplicate under defined conditions. It is apparent that the minimal set of essential genes for an *M.genitalium*-derived minimal cell has to be different depending on growth conditions, e.g. whether the minimal cell is growing *in vitro* in a serum-enriched medium or in the respiratory or urogenital tracts of the host. It might grow well without adhesin proteins under laboratory conditions, but it would probably be unable to survive in the respiratory or urogenital tract without the ability to colonize following its adhesion to the epithelial surfaces. Therefore, when defining a minimal cell, one has also to define the environmental conditions for growth of this cell.

An obvious approach for defining a minimal cell is to start with *M.genitalium*, the smallest known existing cell, and gradually reduce its genetic complexity. The comparison with the larger *M.pneumoniae* genome provides hints for genetic information that may be deleted (see below). An alternative approach for defining the minimal cell was applied by Mushegian and Koonin (41). They identified all pairs of orthologous genes in the distantly related bacteria *H.influenzae* and *M.genitalium* and, on this basis, constructed a minimal gene set of 240 members complemented by a small number of non-orthologous genes, ending up with a final set of 256 genes. This approach has two disadvantages: essential functions could be missed and the difficulty in experimental verification of the minimal cell [see also the commentary of Maniloff (42) on the topic 'the minimal cell genome']. For instance a conventional bacterium possesses a cell wall which provides structural stability and protects it against osmotic stress.

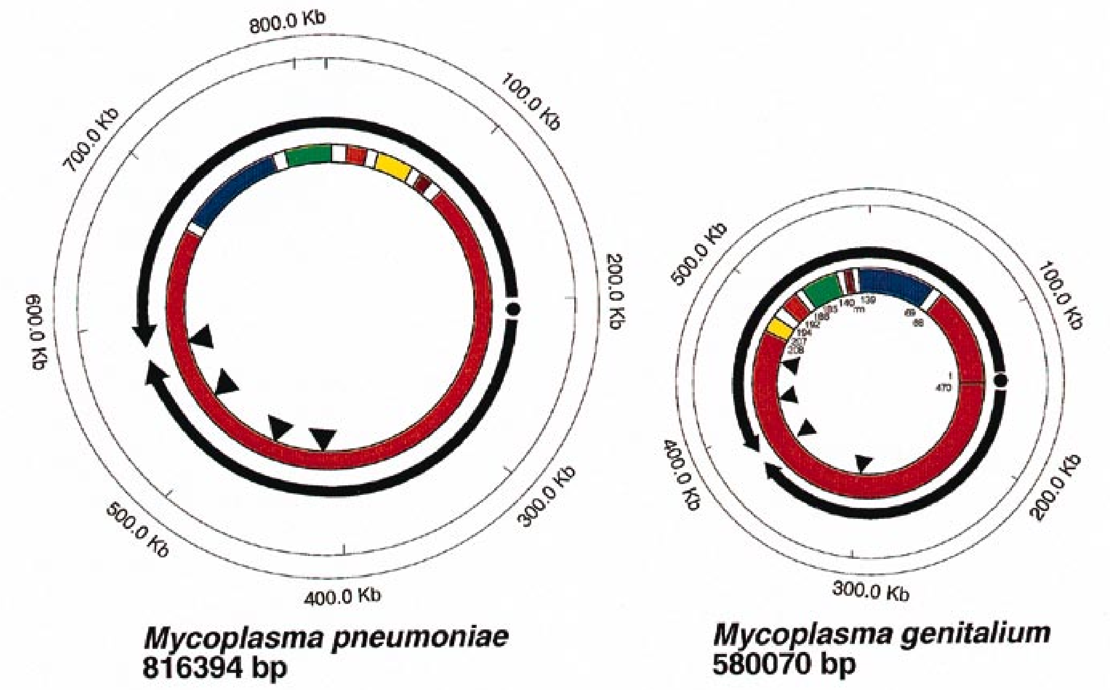


Figure 4. Comparative presentation of the organization of DNA segments with conserved gene order and general direction of transcription in *M.pneumoniae* and *M.genitalium.* Segments with conserved gene order in both bacteria are shown in the same colour. Gaps within the coloured block and triangles in the red segments represent repetitive DNA sequences. The black arrows starting in two opposite directions from the proposed origin of replication (black dot) indicate the general direction of transcription, the two arrowheads mark the region on the genome where the direction of transcription is frequently changing (see also Fig. 1). The numbers inside the coloured circle indicate the first and last *M.genitalium* ORF (MG is omitted) of each of the six DNA segments with conserved gene order. For this illustration the *M.genitalium* sequence was reverse complemented and the putative origin of replication was oriented in the 3 'o' clock position. This figure shows that the general transcription orientation is very conserved in these two bacteria and that the proposed genomic recombination took place only in one half of the genome. For clarity, the ORFs which are transcribed contrary to the general direction of transcription (~15%, Fig. 1) have not been indicated.

The wall-less *M.genitalium* and *M.pneumoniae* do not code for a single gene involved in cell wall formation but as a substitute they possess a cytoskeleton (43). The proteins which were proposed to participate in cytoskeleton formation do not share significant sequence similarities with proteins from other bacteria, therefore proteins which have the same function in both, *H.influenzae* and *M.genitalium*, do not share sequence similarities and might be eliminated as non-essential. On the contrary, comparison of the two mycoplasma genomes permits several conclusion as to the probable number of genes involved in macromolecule synthesis, metabolic and anabolic pathways, transport and formation of structural elements. The same or a similar number of genes involved in both mycoplasmas in DNA replication, transcription and translation suggests that in these functional categories the minimal gene set has already been established. In other functional categories, like cell envelope and cytoskeletal proteins, energy metabolism or transport, more flexibility seems possible, since environmental conditions might strongly influence the number of genes/functions required to be supplied by the minimal cell itself. One has also to consider that quite a proportion of gene functions in *M.genitalium* are either still unknown or insufficiently defined, and it might well be that among the hitherto unclassified genes, essential genes are hidden. In addition, it remains to be seen how much intergenic regions contribute to a functional chromosome, e.g. by influencing the chromosomal DNA topology.

The definition of the minimal cell might be considered as a pure academic problem and cannot be answered satisfactorily, but it may be possible to define a minimal set, a core of genes, which has to be present in every self-replicating cell and which has to be complemented by additional genetic information depending on the growth conditions provided by the specific environment. Mycoplasmas in general, and *M.genitalium* and *M.pneumoniae* in particular, can serve as excellent model organisms to tackle experimentally the question of the essential functions for small self-replicating cells.

ACKNOWLEDGEMENTS

We thank E. Pirkl for excellent technical assistance, R. Mosbach for technical assistance with hardware and software problems, I. Schmid for preparing the manuscript, H. Neimark for discussion, and S. Razin and R. Walker for reading the paper and giving valuable suggestions for its improvement. This research was supported by a grant from the Deutsche Forschungsgemeinschaft (He 780/5-1-He 780/5-4) and by the Fonds der Chemischen Industrie.

REFERENCES

1 Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M. *et al.* (1995) *Science,* **269,** 496–512.

2 Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M. *et al.* (1995) *Science,* **270,** 397–403.

3 Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D. *et al.* (1996) *Science*, **273,** 1058–1073.

4 Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S. *et al.* (1996) *DNA Res.*, **3,** 109–136.

5 Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C. and Herrmann, R. (1996) *Nucleic Acids Res*., **24,** 4420–4449.

6 Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayes, W. S., Borodovsky, M., Rudd, K. E. and Koonin, E. V. (1996) *Current Biol.*, **6,** 279–291.

7 Koonin, E. V., Mushegian, A. R. and Rudd, K. E. (1996) *Current Biol.*, **6,** 404–416.

8 Jensen, J. S., Hansen, H. T. and Lind, K. (1996) *J. Clin. Mircobiol.*, **34,** 286–291

9 Hu, P. C., Collier, A. M. and Baseman, J. B. (1977) *J .Exp. Med.*, **145,** 1328–1343.

10 Tully, J. G., Taylor Robinson, D., Cole, R. M. and Rose, D. L. (1981) *Lancet*, **1,** 1288–1291.

11 Baseman, J. B., Dallo, S. F., Tully, J. G. and Rose, D. L. (1988) *J. Clin. Microbiol.*, **26,** 2266–2269.

12 Goulet, M., Dular, R. , Tully, J. G., Billowes, G. and Kasatiya, S. (1995) *J. Clin. Microbiol.*, **33,** 2823–2825.

13 Jacobs, E. (1991) *Rev. Med. Microbiol*., **2,** 83–90.

14 Taylor-Robinson, D. (1996) *Clin. Infect. Dis.*, **23,** 671–684.

15 Yogev, D. and Razin, S. (1986) *Int. J. Syst. Bacteriol*., **36,** 426–430.

16 Pearson, W. R. and Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA*, **85,** 2444–2448.

17 Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) *J. Mol. Biol*., **215,** 403–410.

18 Needleman, S. B. and Wunsch, C. D. (1970) *J. Mol. Biol*., **48,** 443–453.

19 Higgins, D. G. and Sharp, P. M. (1988) *Gene*, **73,** 237–244.

20 Brosius, J. (1996) *Science*, **271,** 1302.

21 Venter, J. C. (1996) *Science*, **271,** 1303–1304.

22 Ouzounis, C., Casari, G., Valencia, A. and Sander, C. (1996) *Mol. Mircobiol*., **20,** 898–900.

23 Krause, D. C., Proft, T., Hedreyda, C. T., Hilbert, H., Plagens, H. and Herrmann, R. (1997) *J. Bacteriol*., in press

24 Ruland, K., Wenzel, R. and Herrmann, R. (1990) *Nucleic Acids Res*., **18,** 6311–6317.

25 Wenzel, R. and Herrmann, R. (1988) *Nucleic Acids Res*., **16,** 8337–8350.

26 Bickle, T. A. and Kruger, D. H. (1993) *Microbiol. Rev.*, **57,** 434–450.

27 Dybvig, K. and Yu, H. (1994) *Mol. Microbiol*., **12,** 547–560.

28 Postma, P. W., Lengeler, J. W. and Jacobson, G. R. (1993) *Microbiol. Rev.*, **57,** 543–594.

29 Pollack, J. D. (1992) In Maniloff, J., McElhaney, R. N., Finch, L. R., and Baseman, J. B. (eds), *Mycoplasmas - Molecular Biology and Pathogenesis.* American Society for Microbiology, Washington, DC, pp. 181–200.

30 Proft, T. (1995), PhD thesis, Ruprecht-Karls-Universität Heidelberg.

31 Braun, V. and Wu, H. C. (1994) In Ghuysen, J.-M., and Hakenbeck, R. (eds), *Bacterial Cell Wall.* Elsevier Science B.V., Vol. 27., Chapter 14, pp. 319–341.

32 Radestock, U. and Bredt, W. (1977) *J. Bacteriol*., **129,** 1495–1501.

33 Herrmann, R. (1992) In Maniloff, J., McElhaney, R. N., Finch, L. R., and Baseman, J. B. (eds), *Mycoplasmas - Molecular Biology and Pathogenesis.* American Society for Microbiology, Washington, DC, pp. 157–168.

34 Cox, E. C. and Yanofsky, C. (1967) *Proc. Natl. Acad .Sci. USA*, **58,** 1895–1902.

35 Citti, C. and Wise, K. S. (1995) *Mol. Mircobiol*., **18,** 649–660.

36 Hilbert, H., Himmelreich, R., Plagens, H. and Herrmann, R. (1996) *Nucleic Acids Res*., **24,** 628–639.

37 Sueoka, N. (1961) *Proc. Natl. Acad. Sci. USA*, **47,** 1141–1149

38 Osawa, S., Jukes, T. H., Watanabe, K. and Muto, A. (1992) *Microbiol. Rev.*, **56,** 229–264

39 Ogasawara, N. and Yoshikawa, H. (1992) *Mol. Microbiol*., **6,** 629–634.

40 Peterson, S. N., Bailey, C. C., Jensen, J. S., Borre, M. B., King, E. S., Bott, K. F. and Hutchison, C. A. (1995) *Proc. Natl. Acad. Sci. USA*, **92,** 11829–11833.

41 Mushegian, A. R. and Koonin, E. V. (1996) *Proc. Natl. Acad. Sci. USA*, **93,** 10268–10273.

42 Maniloff, J. (1996) *Proc. Natl. Acad. Sci. USA*, **93,** 10004–10006.

43 Krause, D. C. (1996) *Mol. Mircobiol***., 20,** 247–253.

44 Koonin, E. V. and Bork, P. (1996) *Trends Biochem. Sci*., **21,** 128–129.

45 Koonin, E. V., Mushegian, A. R. and Bork, P. (1996) *Trends Genet.*, **12,** 334–336.

46 Simoneau, P., Li, C. M., Loechel, S., Wenzel, R., Herrmann, R. and Hu, P. C. (1993) *Nucleic Acids Res*, **21,** 4967–4974.

47 Atkins, J. F. and Gesteland, R. F. (1996) *Nature*, **379,** 769–771.

48 Schatz, G. and Dobberstein, B. (1996) *Science*, **271,** 1519–1526.

49 Saurin, W. and Dassa, E. (1996) *Mol. Mircobiol*., **22,** 389–391.

50 Robinson, K., Gilbert, W. and Church, G. M. (1996) *Science*, **271** 1302–1313.