Chapter 11

Comparative Genome Analysis of the Mollicutes

THOMAS DANDEKAR, BEREND SNEL, STEFFEN SCHMIDT, WARREN LATHE, MIKITA SUYAMA, MARTIJN HUYNEN and PEER BORK
EMBL, *Postfach 102209, D-69012 Heidelberg, Germany*

1. OVERVIEW

The Mollicutes are Eubacteria that have probably been derived from Lactobacilli, Bacilli, and Streptococci by regressive evolution and genome reduction to produce the smallest and simplest free-living and self-replicating cells. The life style is in general parasitic. Structurally, the Mollicutes are characterized by the complete lack of a cell wall, and the presence of an internal cytoskeleton[27,46].

As in other comparative studies, comparative genomics of Mollicutes requires the availability of a sufficient number of species and genomes to draw solid conclusions[42]. Several Mollicute genomes have been partially sequenced and many will be completely determined in the near future. This includes *Mycoplasma capricolum*, *Mycoplasma mycoides subsp. mycoides SC* (The Royal institute of Technology, Stockholm; National Veterinary Institute, Uppsala; still in progress) and *Mycoplasma hyopneumoniae* (in progress, University of Washington). However, for comparative genome analysis it is desirable to have these genome sequences complete, well annotated, and available in public databases. Until recently, this did only apply to *Mycoplasma genitalium*[12], *Mycoplasma pneumoniae*[19], and *Ureaplasma urealyticum*[15]. At the time of writing, the genomic analysis of *Mycoplasma pulmonis* became available too[7]. Moreover, the number of total genomes analyzed, and in particular those of prokaryotes has rapidly increased in the last years (see e.g. www.tigr.org/tdb/mdb/mdbinprogress.html for an overview). Table 1a

Molecular Biology and Pathogenicity of Mycoplasmas, Edited by Razin and Hermann, Kluwer Academic/Plenum Publishers, New York, 2002

summarizes some useful comparative genomics WEB pointers for Mollicutes.

Table 1a. Some useful Mollicute genome WEB pointers

Microbial genomes in progress	www.tigr/tdb/mdb/mdbinprogress.html
Annotation for *Escherichia coli*	www.genome.wisc.edu
Annotation for *Mycoplasma genitalium*	www.bork.embl-heidelberg.de/Annot/MG
Annotation for *Mycoplasma pneumoniae*	www.bork.embl-heidelberg.de/Annot/MP
Annotation for *Mycoplasma pulmonis*	Genolist.pasteur.fr/MypuList
Annotation for *Ureaplasma urealyticum*	Genome.microbio.uab.edu/uu/uugen.html
M. pneumoniae Genome and proteome project	www.zmbh.uni-heidelberg.de/M_pneumoniae/MP_HOME.html mail.zmbh.uni-heidelberg.de/M_pneumoniae/genome/Results.html
Mycoplasma genitalium genome page	www.tigr.org/tigr-scripts/CMR2/GenomePage3.spl?database=gmg
Mycoplasma metabolism	www.med.ohio-state.edu/medmicro/pages/jdpollack.htm
Mycoplasma gallisepticum (veterinary site)	members.aol.com/FinchMG/MGLinks.htm #Consolidated
Clusters of orthologous genes	www.ncbi.nlm.nih.gov/COG
gene context revealed by STRING	www.bork.emb1-heidelberg.de/STRING
Global transposon mutagenesis of *M.pneumoniae* and *M.genitalium*[21]	www.sciencemag.org/feature/data/1042937.shtml (supplementary material)

Table Ib. Basic genome parameters

	size[1]	genes[2]	density[3]	families[4]
Mycoplasma genitalium	580,074	480	0.83	393
Mycoplasma pneumoniae	816,394	687	0.84	462
Ureaplasma urealyticum	751,719	611	0.81	429
Mycoplasma pulmonis	963,879	782	0.81	519

[1] Given is the size in basepairs

Given is the number of identified protein reading frames

[3] Given is the ratio of genes per kilobase

[4] Given is the number of different protein families. Note that many of these families contain only one protein because no further related protein was found in the Mollicute genome. We use an E-value cut-off of 0.01. Our results are comparable to those of Teichmann *et al.* (1999) who refer to a similar procedure to assist structure predictions.

The position of the Mollicute genomes from M. *pneumoniae,* M. *genitalium, U. urealyticum,* and *M. pulmonis* in the phylogenetic tree shows the Mollicutes to be a well defined, monophyletic group with the *M. genitalium* and *M. pneumoniae* genomes being closely related sister species, the *U. urealyticum* genome *a* little bit more distant and *M. pulmonis* having diverged the earliest from the four (Figure la-c). Figure 1a shows a genome

phylogeny based on shared gene content between completely sequenced genomes. This measure is an alternative to measures that are based on levels of sequence identity and has the advantage that it better reflects the evolution of the whole genome[40]. An alternative to reconstructing genome phylogenies based on gene content, is to reconstruct them on the basis of gene order. As conservation of gene order decreases almost linearly with time over short evolutionary distances[42], this might be appropriate for comparisons within the Mollicutes. Figure 1b shows the genome phylogeny based on shared gene order. Genome phylogenies as the ones above can easily be derived by calculating pairwise distances between all pairs of genomes based on the number of shared genes (or based on the number of conserved gene pairs in the case of gene order). Subsequently, a simple, distance-based phylogenetic inference method is used (such as neighbor-joining) to construct the phylogeny.

Figure 1a shows that for the phylogeny based on gene content, the Mollicute genomes examined are monophyletic and positioned on a branch of the tree within the Gram-positive bacteria. The Gram-positive bacteria themselves are however non-monophyletic, falling apart in the low- and high-GC content branch. However, when using gene order as measure (Figure lb) all gram-positive bacteria do appear to be monophyletic. In Figure 1c we compare genome phylogenies with the results from a standard ribosomal RNA tree (based on 23 S rRNA). Despite the differences regarding early evolutionary events, all three independent measures confirm that *M.pulmonis* has diverged the earliest of the four Mollicute genomes compared. An overview of basic genome parameters such as size and number of genes is given in Table Ib. For example, the smallest genome, *Mgenitalium*, contains 480 genes that fall into 393 gene families. Although there is substantial variation in the number of genes and gene families per genome, gene density is similar, varying between 81 to 84%.

There is a considerable variation among the Mollicute genome sizes. Already early studies[3] with contour clamped homogeneous field (CHEF) agarose gel electrophoresis of DNA showed genome size variation for different serotype standard strains of *Ureaplasma urealyticum* isolates. Genome sizes (in kbp) were 760 for four biotype 1 strains (characterized by temporary inhibition of growth in broth by manganese) and 840-1140 for eleven biotype 2 (permanent growth inhibition by manganese) strains. Other estimates were: 720 for *Mycoplasma hominis*, 1070 for *Mycoplasma hyopneumoniae*, 890 for *Mycoplasma flocculare*, 1180 and 1350 for *Mycoplasma mycoides subsp. mycoides Y* and GC1176-2, respectively, and 1650 and 1580 for *Acholeplasma laidlawii* B and *PG* 8, respectively[32, 36].

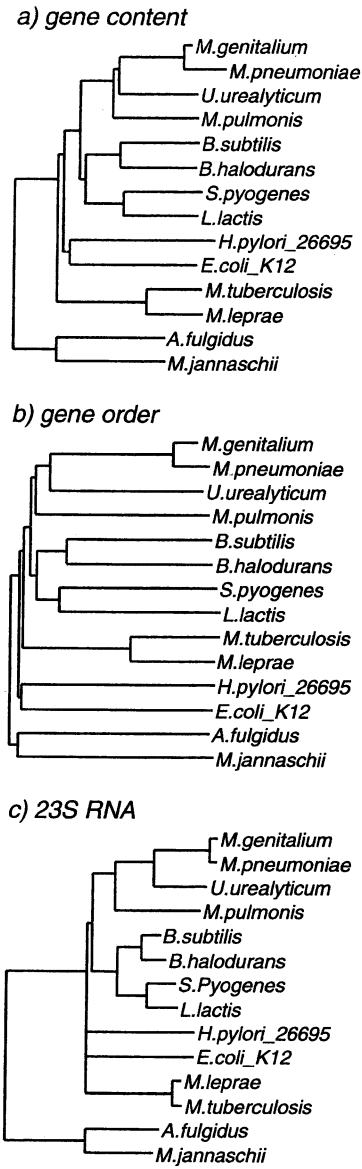

Figure 1. Phylogenetic trees for the four well annotated Mollicute genomes (a) drawn according to gene content. (b) drawn according to gene order. (c) based on ribosomal RNA content.

These data illustrate that Mollicute genomes are diverse in size with some larger than estimated from DNA renaturation kinetics. The genome inventory may vary quite significantly between different strains of the same species, as seems to be the case for many prokaryotic genomes in general. This clear genetic diversity in individual Mollicute species has also been confirmed in current studies. Thus *Mycoplasma capricolum subsp. capripneumoniae* strains were analyzed for amplified fragment length polyrnorphisms (AFLP[25]). AFLP fingerprints of 38 strains derived from different countries in Africa and the Middle East consisted of over 100 bands in the size range of 40-5OObp. The similarity between individual AFLP profiles, calculated by Jaccard's coefficient, ranged from 0.92 to 1.0. On the basis of the polymorphisms detected, the analysed strains can explicitly be grouped into two major clusters, equivalent to two evolutionary lines of the species found by 16s rDNA analysis.

Horizontal gene transfer has also to be considered in comparative genomics of prokaryotes[13, 43, 40]. Archaea and nonpathogenic bacteria seem to have the highest percentages of such genes, pathogenic bacteria, except for Mycoplasma genitalium, have the lowest. Furthermore, genes involved in transcription and translation are less likely to be transferred than metabolic household genes.

As illustrated in the next section, one has to bear in mind that any sequencing project and its genome annotation provide only a momentary picture of the genome. First, every species is evolving in time, caused by adaptation and other selective processes (e.g. rapid co-evolution of host and parasite) or by neutral evolution. Even if genome sequences would be static, our knowledge on those is evolving, as are the methods to analyse them.

2. EVOLVING MOLLICUTE ANNOTATION: THE *M. PNEUMONIAE* EXAMPLE

The Mollicute genomes contain all basic features of a living cell and many specific modifications, though most of these genomes are quite compact and small. The genome of *M.genitalium*[12] was the first one to be sequenced and published, followed a few months later by *M.pneumoniae*[19]. The second genome was also extremely valuable for the annotation of the first one[18] as the tools for identifying functional features are not perfect. Even when reassessing the functional information available for the very same genome, numerous differences become apparent. For example, the *Mycoplasma* pneumoniae genome was meticulously re-annotated four years after the original publication to incorporate novel data applying latest software[8]. The total number of identified ORFs increased from 677 to 688: Ten new

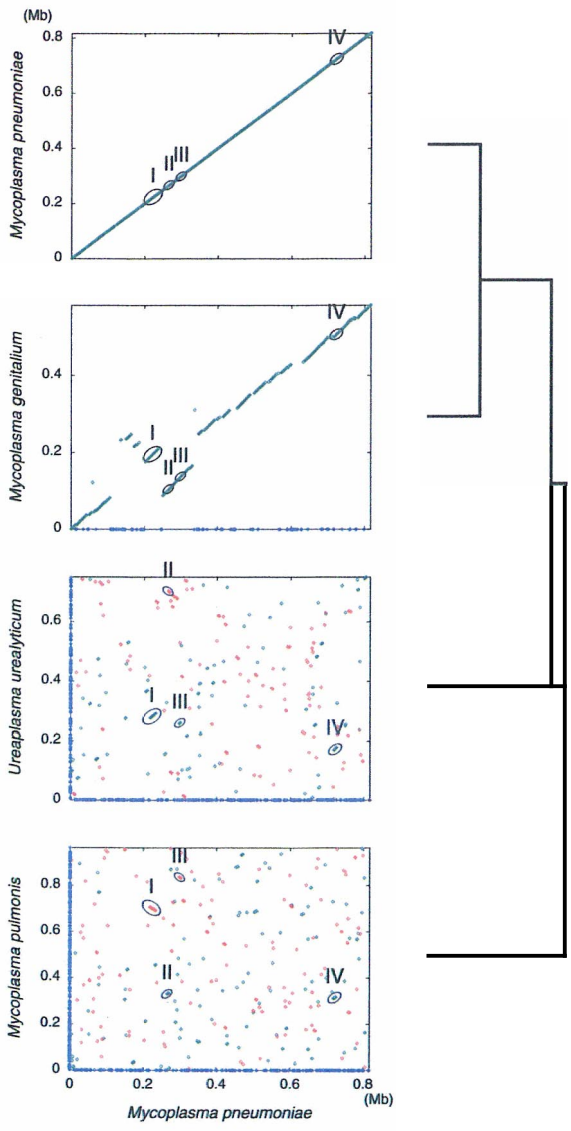

Figure 2. Dotplots of orthologous genes in four genome pairs. ***M.pneumoniae*** is always in the horizontal axis in the four panels. The top panel is the comparison with itself. Directional similarity is indicated by colors: green, pairs of genes with the same direction; red, those with opposite directions. The open reading frames (ORFs) without significant similarity to the other compared genome, even in local DNA sequence level, are defined as the species specific ORFs and indicated by blue dots on each axis. Four syntenic regions are marked with ellipses with Roman numerals; I, II, III, and IV (see Table **3).**

proteins were predicted in intergenic regions, and two were newly identified by mass-spectroscopy. One protein ORF was discarded. The predicted number of RNAs was increased from 39 to 42 genes. For 19 of the now 35 tRNAs and for six other functional RNAs the exact genome positions were re-annotated, two new Leu tRNAs and a small 200nt RNA were identified, 16 protein reading frames were extended and 8 shortened. A consistent annotation vocabulary was introduced and annotation reasoning, categories and comparisons to other published data on protein function assignments were given. Experimental evidence included 2D gel chromatography in combination with mass-spectroscopy and new gene expression data. Compared to the original annotation, this study increased the number of proteins with predicted functional features from 349 to 458. This involved 36 new predictions and 73 protein assignments confirmed by published literature. 23 previous annotations were to broad in their definitions and had to be reduced to better describe the function of the encoded protein. For 30 protein genes we found additions to their predicted function compared to the previous annotation. mRNA expression data support transcription of 184 of the functionally unassigned reading frames. Proteins missing in the first annotation were identified such as the subunit A of glutamine-tRNA amidotransferase.

Reannotation examples include:

- **Molecular functions** for several proteins were clarified (in the following MPN denotes the new numbering, MP the original numbering of the *M.pneumoniae* genome[19]. An example are MPN558(MP284) and MPN557(MP285), previously annotated as glucose-inhibited cell-division proteins B and A, which are a methyltransferase (MPN558(MP284)) and an NADH-oxidoreductase (MPN557(MP285)), and that have homologs with known structure (1BHJ and 1FEA, respectively).
- In 36 cases the functional assignment was completely new. An example is the **protein secretion system** in *M.pneumoniae* (Figure 3). The system has been well characterized in *Escherichia coli*. Cytosolic chaperons or regulators (trigger factor, SecB, DnaK, bacterial signal recognition particle and FtsY) deliver the protein to a membrane transporter (SecA). The receptor should also function as a motor to push the protein across the membrane via specific protein channels (SecY, SecG, SecE, SecD and SecF). Himmelreich *et al*. (1996) noted that they had identified trigger factor, DnaK, SRP and FtsY as well as SecA, whereas from the channel-forming proteins only SecY could be assigned, leaving the secretion pathway incomplete. Protein reading frames similar to SecD, SecE and SecG have now been identified, yielding a new and more complete picture of this secretory pathway in

M.pneumoniae. Since several pathogenicity factors are secreted, the respective protein channels are potential drug targets. No homologous sequence has been found for SPase I in the secretory pathway in *M.pneumoniae.* SPase I would cleave the signal peptide before secretion. Suitable cleavage sites have been identified for several *M.pneumoniae* proteins. One of the proteases identified may contain this function, e.g. the new annotated intracellular protease MPN386(MP542). The dihydroxyacetone kinase domain from MPN547(MP295) could yield ATP for *M.pneumoniae* by transforming dihydroxyacetone phosphate and ADP into dihydroxyacetone and ATP. The predicted activity can be metabolically connected to the phospholipid metabolism in *Mycoplasma pneumoniae* and the necessary supply of dihydroxyacetone phosphate via MPN051(MP103) (glycerol 3-phosphate dehydrogenase reading frame, confirmed in re-annotation).

- **Carbohydrate metabolizing operons** were known previously for fructose (MPN078(MP077); MPN079(MP076)) and mannitol (MPN651(MP191) to MPN653(MP189)). It is now apparent that Ribulose is transported (MPN496(MP346), MPN494(MP347)) and channeled via D-arabinose 6-hexulose 3-phosphate synthase (MPN493(MP348)) and D-arabinose 6-hexulose 3-phosphate isomerase MPN492(MP349) into fructose 6-phosphate and glycolysis.

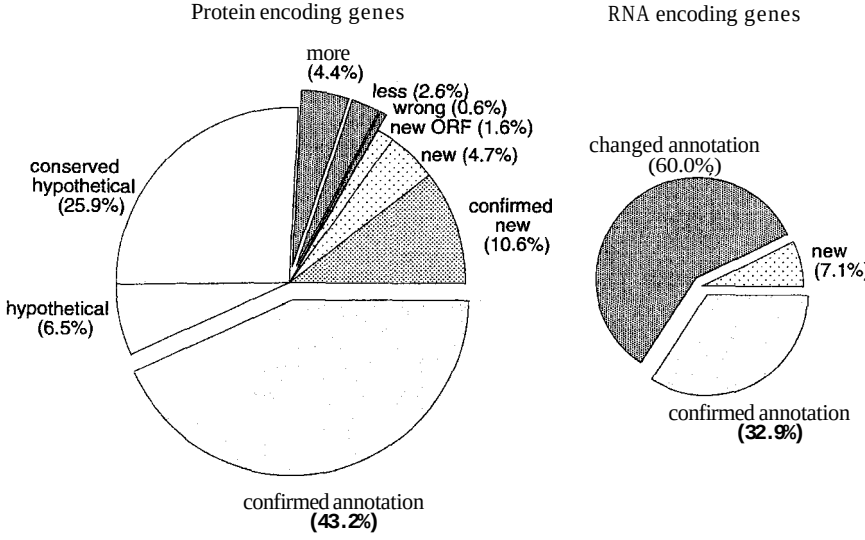

***Figure* 3.** Annotation gain after re-annotation in *M.pneumoniae.* Different re-annotation categories are given and represented according to size. Both the re-annotation of proteins and of RNAs are compared.

3. EVOLVING METHODS FOR MOLLICUTE GENOME ANALYSIS: THE *M. GENITALIUM* EXAMPLE

Mollicute genome annotation heavily relies on comparative sequence analysis methods. Due to their compactness Mollicute genomes have been frequently used as a benchmark for method development. This, in turn, adds further to our knowledge of the Mollicute genomes.

For example, the *Mgenitalium* genes have been used as a benchmark for three-dimensional proteins structure prediction[11]. A dramatic sensitivity and selectivity increase could be shown when iterative homology search techniques have been applied to the same genome[23,] . Other groups developed methods for this purpose further and also used *M.genitalium* as benchmark. This led to the estimate that currently for more than 50% of all the proteins encoded in *M.genitalium,* reliable assignments to domains with known three dimensional structure can be made (cf. reviews of Teichmann[45,44]).

Genome analysis of *M.pneumoniae* and *M.genitalium* is also a good test case for examining the predictive power of the so-called genome context methods to predict functional interactions of encoded proteins (Table 2). Various qualitatively new methods (independent of homology based assignments) have been recently proposed to predict functional interactions between proteins based on the genomic context of their genes. Important types of genomic context that one can examine are (i) the fusion of genes; (ii) the conservation of gene-order or co-occurrence of genes in potential operons and (iii) the co-occurrence of genes across genomes (phylogenetic profiles). These types were compared by Huynen *et al.* (2000). Their coverage, their correlations with various types of functional interaction, and their overlap with homology-based function assignment was analysed in *M.genitalium.* Quantitatively, conservation of gene order is the technique with the highest coverage, applying to 37% of the genes. By combining gene order conservation with gene fusion (6%), the co-occurrence of genes in operons in the absence of gene order conservation (8%), and the co-occurrence of genes across genomes (1 1%), significant context information can be obtained for 50% of the genes (the categories overlap). Qualitatively, it was observed in this computational analysis of *M.genitalium* that the functional interactions between genes became stronger as the requirements for physical neighbourhood were set more stringent, while the fraction of potential false positives decreased. In cases in which gene order was conserved in a substantial fraction of the genomes (equal or more than in six out of twenty-five genomes) a single type of functional interaction clearly dominated, namely physical interaction (>80%). In the other cases,

complementary function information from homology searches, available for most of the genes with significant genomic context, remained essential for prediction of the type of interaction. Using a combination of genome context and homology searches, new functional features could be predicted for 10% of the M. *genitalium* genes.

Table 2. M. genitalium genes for which genomic context adds functional information'

Protein	conserved with	proposed role
MG008	MG466 (rib. Protein L34)	translation
MG009	MG006 (thymidilate kinase)	nucleotide. metabolism (dCTPase?)
MG053	MG052 (cytidine deaminase)	phosphoribomutase
MG134	MG240 (dnaX), recR	physical interaction with DNA pol. gamma subunits
MG233	MG232, MG234 (rib. proteins)	ribosomal protein
MG464	MG465 (ribonul. P component)	ribonuclease

[1] In some cases some information about the molecular function can already be retrieved by homology searches: MG008 encodes the so-called thiophene and furan oxidation protein. Its genomic association with ribosomal protein L34 indicates a role in translation. The inability of a species to oxidize thiophene or furan in the absence of MG008 might be a secondary effect, caused by the inability to translate an mRNA into the protein required for the oxidation. MG009 is homologous to deaminases, dehydratases and phosphohydrolases that generally have pyrimidines as substrate. Its genomic link with thymidilate kinase indicates a role in nucleotide metabolism, possibly in the creation of a precursor of thymidilate. MG052 is homologous to phosphohexomutases, and is generally annotated as a phosphomannomutase. The location of MG053 in a nucleoside salvage pathway operon from which deoB, a phosphoribomutase, is missing, indicates that MG053 might have acquired the phosphoribose as substrate. Note that in the genome-based metabolic map of M. *pneumoniae*[19] a phosphohexomutase is however still required to fill the gap between glucose-1-phosphate and glucose-6-phosphate. There is no other candidate for this function than MG053, which leads us to propose that MG053 has two substrate specificities: glucose-1,6-phosphate and deoxy-ribose-l,5-phosphate. Also for proteins for which no information can be derived by homology searches, information can be gained by context searches. MG134 is a hypothetical protein that tends to occur with MG240 and recR, coding for DNA polymerase subunit gamma/tau and a protein that is involved in recombinational repair respectively. The high frequency of occurrence with dnaX and recR indicate a physical interaction between the proteins. MG464 encodes an inner membrane protein that has a strong genomic link with the ribonuclease P protein component, indicating a role as a ribonuclease, or possibly in translation.

As an example may serve the two *Mycoplasma genitalium* proteins MG246 and MG130 that do not have orthologs with experimentally determined functions. Information about their role in the cell can be gained by a combination of homology searches and genome context searches. Using sensitive, profile-based homology searches (PSI-BLAST[1]), MG246 can be shown to be homologous to the catalytic domain of 5-prime nucleotidase from *Escherichia coli* (UshA)[22]. It is a good candidate for the nucleotidase

activity that has been measured in *M. genitalium*[17] but for which no gene has yet been identified. MG130 contains a KH single-stranded ribonucleotide-binding domain (protein domain which is homologous to hnRNP-K)[30] and an phosphohydrolase domain (HD-domain) that hydrolyses phosphates from nucleotides[2]. The functions of these proteins can be linked to each other and to other proteins in the cell using the conservation of genomic context (Figure 5). MG246 and MG130 tend to occur in potential operons with each other, and with 1) MG244, a type II DNA helicase orthologous to the PcrA helicase of *Bacillus subtilis* that is involved in DNA repair and in rolling circle replication. 2) 5-formyl tetrahydrofolate cyclo-ligase, involved in the synthesis of tetrahydrofolate (a co-factor in nucleotide metabolism 3) recA, a single stranded DNA binding protein involved in DNA repair, 4) cinA, a competence-induced protein and 5) phosphatidylglycero-phosphate synthase. The homology information indicates that MG246 and MG130 play a role in nucleotide metabolism. The unifying theme in the context information is that of DNA repair. One possibility would be that MG246 and MG130 are involved in the degradation of nucleotides that are removed in DNA repair.

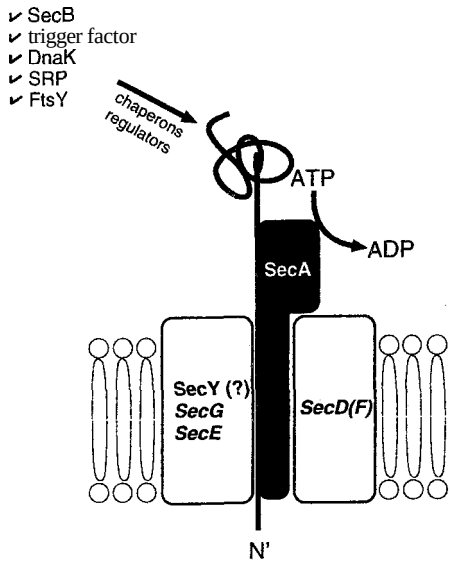

Figure 4. Identification of complete pathways after re-annotation. Several missing components of the secretion system in *M. pneumoniae* could be identified after re-annotation of the *M. pneumoniae* genome[8]. The original genome annotation[19] identified the components shown in black. The new genome analysis added the components shown in grey.

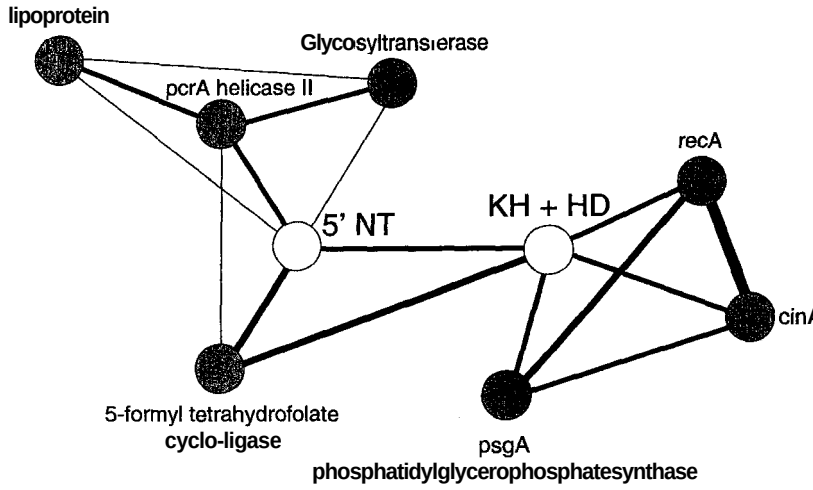

Figure 5. The functions of two hypothetical *Mycoplasma genitalium* proteins MG130 (containing a KH and a HD domain) and MG246 (homologous to 5' nucleotidase) can be related to each other and to a number of other Mycoplasma and Ureaplasma proteins[22]. Specifically to MG244 (a helicase II), MG245 (5-formyl tetrahydrofolate cyclo-ligase, MG339 (recA), MGI15 (cinA) and MGI14 (pgsA phosphatidylglycerophosphate synthase). The thickness of the lines indicates the number of times the genes neighbour each other on a genome. See the text for farther information.

4. COMPARATIVE GENOME ANALYSIS TO DETECT FEATURES OF SPECIFIC MOLLICUTES

The previous section showed that comparative genome analysis such as using the conservation of genome context is a powerful tool in the analysis of Mollicute reading frames and in predicting previously unknown structures or functions. This section will now focus on genome specific features apparent from comparative analysis of Mollicutes.

An initial genome analysis[6] of 214kb from *Mycoplasma capricolum* detected 287 putative proteins representing about a third of the estimated total. A large fraction of these (75%) could be assigned a likely function as a result of homology searches. There is a relatively large number of enzymes involved in metabolic transport and activation suitable for efficient use of host cell nutrients. Bork *et al.* (1995) found in addition an unexpected diversity of enzymes involved in DNA replication and repair and several

anabolic enzymes. A sizeable number of orthologs in E. *coli* was identified (82, corresponding to 28.6% of the proteins analysed).

The subsequent sequencing and analysis of the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae* made it possible to define the essential functions of a self-replicating minimal cell and the specific features of a Mollicute. Glass *et al*. (2000) subsequently reported the complete sequence (751,719bp) of *Ureaplasma urealyticum* (parvum biovar), another mucosal human pathogen. It is a common commensal of the urogenital tract but it can cause opportunistic infections, e.g. during pregnancy. Differential genome analysis becomes more powerful considering these three well-annotated public genomes together. Several features make *U. urealyticum* unique among Mollicutes and all bacteria. Almost all ATP synthesis in this organism is the result of urea hydrolysis, which generates an energy-producing electrochemical gradient. Some highly conserved eubacterial enzymes appear not to be encoded by *U. urealyticum,* including the cell-division protein FtsZ, chaperonins GroES and GroEL, and ribonucleoside-diphosphate reductase. *U, urealyticum* has six closely related iron transporters, which apparently arose through gene duplication, suggesting that it has a kind of respiratory system not present in other small bacteria. The genome is only 25.5% G+C in nucleotide content, and the G+C content of individual genes may be used to predict how essential those genes are to Ureaplasma survival. There are 613 *U. urealyticum* protein-coding genes15, only 324 are homologous to *M. genitalium* genes or *M. pneumoniae* genes. No function could be predicted for 77 of the genes shared by all three Mollicutes; for a list see genome.microbio.uab.edu/uu.

Using rapid automated procedures for genome comparisons unique and specific genes for this Mollicute genome triplet or quartet (considering then in addition *M .pulmonis,* see below) but not present in 49 other genomes (Eucaryota, Archaea, Eubacteria with known genome sequence) are calculated and shown in Table 3. Furthermore, some Mollicute specific conserved small gene clusters become apparent (Table 4). This interesting set of genes should be further characterized experimentally, their unique nature makes it difficult (legend to Figure 3) to deduce their function by sequence comparisons alone.

Figure 2 shows dotplot comparisons of the orthologous genes in all four Mollicute genomes using always *M. pneumoniae* as the horizontal axis. The tree shown, based on 263 orthologues shared between all four genomes confirms the Mollicute phylogeny found in Figure 1a-c by different methods.

Table 3. Synteny regions in four compared Mollicute genomes[1]

Group I. Ribosomal proteins				
MPN164	MG150	UU232	MYPU_5900	S10
MPN165	MG151	UU233	MYPU_5890	L3
			MYPU_5880	(unknown OW)
MPN166	MG152	UU234	MYPU_5870	L4
MPN167	MG153	UU235	MYPU_5860	L23
MPN168	MG154	UU236	MYPU_5850	L2
MPN169	MG155	UU237	MYPU_5840	S19
MPN170	MG156	UU238	MYPU_5830	L22
			MYPU_5820	(unknown ORF)
MPN171	MG157	UU239	MYPU_5810	S3
MPN172	MG158	UU240	MYPU_5800	L16
MPN173	MG159	UU241	MYPU_5790	L29
MPN174	MG160	UU242	MYPU_5780	S17
MPN175	MG161	UU243	MYPU_5770	L14
MPN176	MG162	UU244	MYPU_5760	L24
MPN177	MG163	UU245	MYPU_5750	L5
MPN178	MG164	UU246	MYPU_5740	S14
MPN179	MG165	UU247	MYPU_5730	S8
MPN180	MG166	UU248	MYPU_5720	L6
MPN181	MG167	UU249	MYPU_5710	L18
MPN182	MG168	UU250	MYPU_5700	S5
MPN183	MG169	UU251	MYPU_5690	L15
MPN184	MG170	UU252	MYPU_5680	SecY
MPN185	MG171	UU253	MYPU_5670	adenylate kinase
MPN186	MG172	UU254	MYPU_5660	methionine amino peptidase
MPN187	MG173	UU255	MYPU_5650	initiation factor 1
MPN188	MG174	UU256	MYPU_5640	L36
MPN189	MG175	UU257	MYPU_5630	S13
MPN190	MG176	UU258	MYPU_5620	S11
MPN191	MG177	UU259	MYPU_5610	RNA polymerase alpha core subunit
MPN192	MG178	UU260	MYPU_5600	L17
Group II (oligopeptide transporter)				
MPN215	MG077	UU568	MYPU_2830	oligopeptide transport system permease
MPN216	MG078	UU567	MYPU_2840	oligopeptide transport system permease
MPN217	MG079	UU566	MYPU_2850	Oligopeptide transport ATP-binding protein
MPN218	MG080	UU565	MYPU_2860	oligopeptide transport ATP-binding
Group III (kinase region)				
MPN246	MG107	UU214	MYPU_6870	guanylate kinase
		UU215		(conserved hypothetical)
MPN247	MG108	UU216	MYPU_6860	protein phoshatase
MPN248	MG109	UU217	MYPU_6850	Ser/Thr protein kinase
MPN249	MG110	UU218	MYPU_6840	conserved hypothetical

Group III (kinase region)				
Group IV (ATP synthase)				
MPN597	MG398	UU127	MYPU 2650	ATP synthase ε-chain
MPN598	MG399	UU128	MYPU 2660	ATP synthase β-chain
MPN599	MG400	UU129	MYPU 2670	ATP synthase y-chain
		UU130		(unique hypothetical)
MPN600	MG401	UU131	MYPU 2680	ATP synthase a-chain
		UU132		ATP synthase δ-chain (C-term)
			MYPU 2690	ATP synthase δ-chain
MPN601	MG402	UU133		ATP synthase δ-chain
MPN602	MG403	UU134	MYPU 2700	ATP synthase B chain
MPN603	MG404	UU135	MYPU 2710	ATP synthase C chain
MPN604	MG405	UU136	MYPU 2720	ATP synthase A chain

MPN – M. *pneumoniae; MG – M. genitalium;* UU – *U. urealyticum; MYPU – M. pulmonis*

Furthermore, readily apparent from this plot for all four genomes are the conserved regions for ribosomal proteins and three other synteny regions *(Table 3):* Oligopeptide transporter synteny region, ATP synthase coding region and an interesting regulatory locus with kinases, a phosphatase and a conserved hypothetical protein. *U. urealyticum* is less conserved in genome order than are *M. pneumoniae* and *M. genitalium* amongst each other, M. *pulmonis* is even more distant (see Figure *2).* Thus no other big synteny regions stand out. Genome specific genes are shown as blue circles on the axis, but are not shown in the *M. pneumoniae* self comparison. These are also not shown for M. *genitalium* as it contains no open reading frames without similarity to M. *pneumoniae*[20]. M. *genitalium* has been described as the smallest living organism. However, there are 74 genes in M. *genitalium* that do not have orthologues in *U. urealyticum.* Ten of these genes concern energy metabolism[15], which is probably a result of the adaptation of *U. urealyticum* to produce ATP by urea hydrolysis. This is simpler than glycolysis in M. *genitalium.* Global transposon mutagenesis in *M. genitalium*[21] showed that 129 of the 480 protein encoding genes from M. *genitalium* were not essential: Any of them could be deleted without causing death of the cell. However, one has to stress that gene deletion combinations were not tested and may be lethal in many cases. From the *351* genes remaining, *265* were found to be with high probability essential. In *U. urealyticum* there are 69 proteins homologous to the unessential genes from *M. genitalium* and there are *255* of the possibly essential genes also contained in *U. tirealyticum.* Moreover, *289* genes from *U. urealyticum* have no homologues in *M. genitalium. Mycoplasma pulmonis* causes respiratory disease. It is now[7] the fourth pathogenic Mollicute whose sequence is available in a well annotated form. The strain sequenced, UAB CTIP is

composed of a single circular chromosome (length: 963 879bp) with a very low GC content (26.6 mol%), in the same range as *Ureaplasma urealyticum*.

Table 4. Unique Mollicute genes

(a)Unique genes in the Mollicute Triplet *U. urealyticum*, *M. genitalium* and *M. pneumoniae*.

RMG055	MPN068	UU580
MG068	MPN084	UU045
MG144	MPN157	UU322
MG149_1	MPN163	UU092
MG255	MPN358	UU572
MG284	MPN403	UU352
MG286	MPN405	UU505
MG306	MPN435	UU450
MG350_1	MPN527	UU122
MG377	MPN555	UU068
MG384_1	MPN565	UU462
MG397	MPN596	UU292
MG423	MPN621	UU509

(b) mollicutes specific clusters[1]

Extra results from 2 iteration(s); converged after 1 iterations			
1	MG046	MPN059	UU411
2	MG047	MPN060	UU412
Extra results from 2 iteration(s); converged after 1 iterations			
1	MG127	MPN266	UU176
2	MG128	MPN267	UU177
3	MG129	MPN268	UU178
Extra results from 2 iteration(s); converged after 1 iterations			
1	MG313	MPN448	UU330
2	MG315	MPN450	UU328
Extra results from 2 iteration(s); converged after 1 iterations			
1	MG324	MPN470	UU532
2	MG325	MPN471	UU533
Extra results from 2 iteration(s); converged after 1 iterations			
1	MG371	MPN549	UU417
2	MG372	MPN550	UU418
3	MG373	MPN551	UU419

[1] These conserved clusters were found by carrying out a large scale STRING analysis[39] on the whole *M. genitalium* genome where we required that the whole detected cluster only occurred in the Mollicutes and in none of the other 40 complete genomes (in contrast to the larger synteny regions shown in Figure 2 which are found also in other genomes).

(c) Unique genes in the Mollicute quartet: three genomes above and in addition that of *M. pulmonis*)

MG068	MPN581	MYPU_5040	UU045
MG350_1	MPN527	MYPU_5910	UU122
MG423	MPN621	MYPU_7050	UU509
MG045	MPN058	MYPU_4220	UU110
MG373	MPN551	MYPU_4770	UU419

The above mentioned genes (gene identifier given) are all commonly present in the compared mollicute genomes but **not** in the other 49 known genomes compared to (including eucaryotes). This type of specific unique conservation makes it difficult to identify the function by genome analysis and only suggestions for function and reason for conservation can be given for the specific clusters shown in b: MG234 and MG235 are an arninopeptidase and rpL33, respectively; both are involved in translation. In MG046 and MG047, one is a sialoglycoprotease and the other is the S-adenosylmethione synthesis protein metK. MG371 is homologous to phosphoesterases. MG372 is homologous to thiI.

There are 782 protein reading frames covering 91.4% of the genome. A function could be assigned to 486 ORFs (62.2% of all). Of the remaining 37.8%, 92 ORFs (11.8%) are conserved hypothetical proteins. 204 open reading frames remained without any significant database match (26%). Only 29 tRNAs genes and one ribosomal operon are present. Chambaud *et al*. (2001) identified several pathogenicity factors by their genome analysis: sequence polymorphisms within stretches of repeated nucleotides generate phase-variable protein antigens. A recombinase gene is likely to catalyse the site-specific DNA inversions in major *M. pulmonis* surface antigens (see chapter on Genetic Mechanisms of Antigenic Variation by Yogev *et al.)*. Predicted virulence factors include a hemolysin, secreted nucleases and a glycoprotease. Eight genes previously reported to be essential for a self replicating cell in *M. genitalium* are missing from the larger *M. pulmonis* genome (spoT,cpsG, gtaB, fruK, groEL, groES, udk, galE); the lack of the stringent response gene in *M. pulmonis* (spoT) perhaps being the most striking. GroEL and GroES are also missing in *Ureaplasma urealyticum.* The set of unique genes shared by all Mollicutes in the quartet is small (Table 3c).

5. COMPARATIVE ANALYSIS TO REVEAL COMMON FEATURES OF MOLLICUTES

After characterizing genome specific features apparent in different Mollicutes it is appropriate to discuss several general features present in most of the sequenced and annotated Mollicute genomes. More general statements have to be considered with care, as this presents only a minority of all mollicutes. The wall-less Mollicutes as suggested by their phylogenetic position clearly descended from low GC content gram-positive bacteria. This is further supported by the phylogenetic analysis as discussed above (Figure la-c; Figure 2). Some Mollicutes are exceedingly small (0.3 micron in diameter). Their genomes are equally small. However, their gene density is roughly similar (Table lb): The smallest, *Mycoplasma genitalium* contains 580,070 base pairs and currently 480 ORFs, corresponding to 0.83 protein

genes per kilobase (according to our updated data, www.bork.emb1-heidelberg.de/Annot/MG/). The *Escherichia coli* genome has 4,639,000bp and contains 4290 identified protein ORFs and 115 RNA genes[4]; latest genome annotation at www.genome.wisc.edu). This corresponds to 0.92 protein ORFsAb and 0.94 geneslkb including RNA genes. Mollicutes are considered models for describing the minimal metabolism necessary to sustain independent life[31]. Nevertheless, *M. pneurnoniue* contains approximately 25% of the total number of known protein folds with single domains[47]. The Mollicutes sequenced to date have no cytochromes or the TCA cycle except for malate dehydrogenase activity. Some uniquely require cholesterol for growth, some require urea and some are anaerobic. They fix CO_2 in anaplerotic or replenishing reactions. Some require pyrophosphate instead of ATP as an energy source for reactions, including the rate-limiting step of glycolysis: 6-phosphofi-uctokinase. Mollicutes scavenge for nucleic acid precursors and apparently do not synthesize pyrimidines or purines *de novo*. Some species uniquely lack dUTPase activity such as *Entornoplasma ellychniae* ELCN-1T, *Entomoplasma melaleucae* M-1T, *Mesoplusrnu seiffertii* F7T, *Mesoplusrnu entornophilurn* TACT, *Mesoplasma florum* L1T, *Mycoplasma fermentans* PGl8T, *M. pneumoniae, M.genitalium* and *Acholeplasma rnultilocale* PN525T, but still have uracil-DNA glycosylase. Some other species also lack DNA uracil-DNA glycosylase such as *Mesoplasma entornophilurn.* The absence of the latter two reactions that limit the incorporation of uracil or remove it from DNA may be related to the marked mutability of the Mollicutes and their tachytelic or rapid evolution (e.g. these enzyme activities are lacking in *M. genituliurn* and *M. pneumoniue).* Approximately 150 cytoplasmic activities have been identified in these organisms, 225 to 250 are presumed to be present (according to Pollack *et al.*, 1997, see also his chapter Central Carbohydrate Pathways: Metabolic Flexibility and the Extra Role of Some "Housekeeping" Enzymes). About 100 of the core reactions can be graphically linked to a metabolic map. This includes glycolysis, pentose phosphate pathway, arginine dihydrolase pathway, transamination, and purine, pyrimidine, and lipid metabolism. Reaction sequences or loci of particular importance are: phosphofructokinases, NADH oxidase, thioredoxin complex, deoxyribose-5-phosphate aldolase, and lactate, malate, and glutamate dehydrogenases. Enzymatic activities of the Mollicutes can be grouped according to metabolic similarities that are taxonomically discriminating. A pathway listing of all relevant enzymes encoded in the M. *pneurnoniue* genome including available *M. genitalium* orthologues is found at www.bork.emb1-heidelberg.de/Annot/MP/.

Mollicute genomes have developed specialized cell reproduction cycles adapted to the limited genome information and a parasitic life style[29]. Thus

DNA replication in *Mycoplasma capricolum* starts at a fixed site neighboring the dnaA gene and proceeds to both directions after a short arrest in one direction. The initiation frequency fits the slow speed of the replication fork, setting the DNA content constant. The replicated chromosomes migrate to one and three quarters of cell length before cell division to ensure delivery of the replicated DNA to daughter cells. The cell reproduction is based on binary fission but a branch is formed when DNA replication is inhibited. *Mycoplasma pneumoniae* has a terminal structure, designated as an attachment organelle, responsible for both host cell adhesion and gliding motility[2]. Behaviour of the organelle in a cell implies coupling of organelle formation to the cell reproduction cycle. Several proteins coded in three operons are delivered sequentially to a position neighboring the previous organelle and a nascent one is formed. One of the duplicated attachment organelles migrates to the opposite pole of the cell before cell division (see chapters Cell Division and Cytoskeleton, and Cell Division).

Karlin and Campbell (1994) collected data on oligonucleotide abundance (see also the chapters on cell division and cytosceleton). Though we do not support their hypothesis that a Mollicute- or *Sulfolobus-like* endosymbiont rather than an alpha-proteobacterium is the ancestor of animal mitochondrial genomes, there are pronounced similarities in extremes of oligonucleotide abundance common to animal mtDNA, *Sulfolobus,* and *Mycoplasma capricolum.* Furthermore, there are pronounced discrepancies of these relative abundance values with respect to alpha-proteobacteria. In addition, genomic dinucleotide relative abundance measures place *Sulfolobus* and *M. capricolum* among the closest to animal mitochondria1 genomes, whereas the classical eubacteria, especially the alpha-proteobacteria, are at excessive distances. There are also considerable molecular and cellular phenotypic analogies among mtDNA, *Sulfolobus,* and *M. capricolum.*

Another observation related to nucleotide composition is the moderate avoidance of palindromes in Mollicutes, examined in detail for *Mycoplasma genitalium*[14]. Short palindromic sequences (4, 5 and 6bp palindromes) are avoided at a statistically significant level in the genomes of several bacteria, including the completely sequenced *Haemophilus influenzae* and *Synechocystis* sp. genomes and in the complete genome of the archaeon *Methanococcus jannaschii.* In contrast, there is no detectable avoidance in the genomes of chloroplasts and mitochondria. The sites for type II restriction-modification enzymes detected in the above given species tend to be among the most avoided palindromes in a particular genome, indicating a direct connection between the avoidance of short oligonucleotide words and restriction-modification systems with the respective specificity. Palindromes corresponding to sites for restriction enzymes from other species are also

avoided, albeit less significantly, suggesting that in the course of evolution bacterial DNA has been exposed to a wide spectrum of restriction enzymes. This is probably the result of lateral transfer mediated by mobile genetic elements, such as plasmids and prophages. Palindromic words appear to accumulate in DNA once it becomes isolated from restriction-modification systems, as demonstrated by the case of organellar genomes. We note furthermore that there is good genome evidence for an elaborate type I restriction enzyme system both in M. *pneumoniae*[19, 8] and *M. pulmonis*[16].

6. COMBINING GENOMIC FEATURES WITH BIOCHEMICAL KNOWLEDGE

Several of the conclusions apparent from Mollicute genome analysis can be enhanced or critically re-examined when combined with biochemical knowledge and functional analysis. Thus M. *genitalium* is often pictured as a minimal genome[31, 21]. However, this is a relative definition as this depends on the biochemical context of the environment. A good case can only be made for genes essential for a self-replicating cell under any circumstances. For instance, primordial tRNA modifications are required to prevent frame-shifting just besides the anti-codon. A suitable G37-tRNA-methyltransferase gene is also found in Mollicute genomes including M. *genitalium*[3].

The minimalistic setup of the Mollicutes is interesting from a biotechnological point of view. Potentially they could be engineered just to produce more and more of the desired protein with only a small burden of their own standard architecture.

The small Mollicute genomes should use their genomic information efficiently[35]. Evidence for this are the broad aminotransferase acitivities in M. *pneumoniae* and the multifunctional enzyme encoded by $MPN158_{MP674}$, (a riboflavin kinase, an FMN adenylyltransferase; a predicted nicotinate-nucleotide adenylyltransferase would furthermore complete the synthesis from imported nicotinic acid to NAD, a pathway otherwise incomplete). *M. pneumoniae* pseudogenes for arginine deiminase, $MPN305_{MP531}$ and $MPN304_{MP532}$ are an example for remnant enzymes[35] from genome reduction due to the parasitic life style.

For more detailed genome studies, homologous recombination should be developed further. For this, a plasmid that replicates in *Escherichia coli* but not in M. *genitalium* was constructed and used to disrupt the cytadherence-related gene MG218 of M. *genitalium*[9].

There are discrepancies between genes annotated in the genome and biochemical activities measured. This is partly caused by strain variation (see above) and partly because it is difficult to assign and detect function just

by inference from sequence similarities for those open reading frames only occurring in Mollicutes. This includes the biochemically measured enzymes such as aspartate aminotransferase activity (E.C. 2.6.1.1.) well measured biochemically previously in *M. pneumoniae*[28] or 5 prime nucleotidase activity in M. *genitalium*. Both were not detected in the respective original genome annotation, however candidate genes are now available[8, 23]. There are also examples for the opposite case, thus there is genome evidence for ammonia production (arginine deaminase $MPN560_{MP282}$) in M. *pneumoniae* but enzyme activities of this sort have not yet been unequivocally determined yet[33]. Examination of genomic or enzymatic activity data alone neither provides a complete picture of metabolic function or potential, nor confidently reveals sites amenable to inhibition. *Ureaplasma urealyticum* (parvum) provides several examples[34]. Combining evidence from genomic sequence, transcription, translational phenomena, structure and enzymatic activity gives the best picture of the organism's metabolic capabilities.

Non-orthologous gene displacement complicates function assignment by sequence analysis. For example polymerase type I has been duplicated in *M. pneumoniae*[26] and seems to replace the missing function for DNA repair. Biochemical experiments are necessary to confirm this hypothesis. However, complete genome sequences from close by non-mollicute genomes will enhance bioinfornatical analyses such as those of *Streptococcus pyogenes*[10] and *Lactococcus lactis*[5]. The combination of analysis tools should in particular give more complete information about pathogenicity factors, the cytoskeleton related proteins and proteins comprising species-specific adaptations such as the attachment organelle in *M. pneumoniae* (besides the well known cytadherence operon a total of 27 proteins in the genome are predicted to be involved in cytadherence in *M. pneumoniae;* see chapter on Cytadherence and Cytoskeleton).

Comparative genome analysis of Mollicutes is an ongoing exercise. In the light of new genomes, more experimental data and better software, more genome information can be deciphered[15, 22, 8]. The exercise of re-annotation and re-evaluation of gene information will become more important. The genome variety of Mollicutes is, despite their small genomes rather considerable and includes many specific adaptations. Thus, comparative genome analysis will remain an important tool to broaden our knowledge on these organisms.

REFERENCES

1. Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

2. Aravind, L. and E. V. Koonin. 1998. The HD domain defines a new superfamily of metal-dependent phosphohydrolases. Trends Biochem. Sci. 23:469-472.
3. Bjork, G. R., K. Jacobsson, K. Nilsson, M. J. Johansson, A. S. Bystrom, and O. P. Persson. 2001. A primordial tRNA modification required for the evolution of life? EMBO J. 20:23 1-239.
4. Blattner, F. R., , C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453-1474.
5. Bolotin, A., S. Mauger, K. Malarme, S. D. Ehrlich, and A. Sorokin. 1999. Low-redundancy sequencing of the entire *Lactococcus lactis* IL1403 genome. Antonie Van Leeuwenhoek 76:27-76.
6. Bork, P., C. Ouzounis, G. Casari, R. Schneider, C. Sander, M. Dolan, W. Gilbert, and P. M. Gillevet. 1995. Exploring the *Mycoplasma capricolum* genome: a minimal cell reveals its physiology. Mol. Microbiol. 16:955-967.
7. Chambaud, I., R. Heilig, S. Ferris, V. Barbe, D. Samson, F. Galisson, I. Moszer, K. Dybvig, H. Wroblewski, A. Viari, E. P. Rocha, and A. Blanchard. 2001. The complete genome sequence of the murine respiratory pathogen *Mycoplasmapulmonis*. Nucleic Acids Res. 29:2145–2153.
8. Dandekar, T., M. Huynen, J. T. Regula, B. Ueberle, C. U. Zimmermann, M. A. Andrade, T. Doerks, L. Sanchez-Pulido, B. Snel, M. Suyama, Y. P. Yuan, R. Herrmann, and P. Bork. 2000. Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. Nucleic Acids Res. 28:3278-3288.
9. Dhandayuthapani, S., W. G. Rasmussen, and J. B. Baseman. 1999. Disruption of gene mg218 of *Mycoplasma genitalium* through homologous recombination leads to an adherence-deficient phenotype. Proc. Natl. Acad. Sci. USA 96:5227-5232.
10. Ferretti, J. J., W. M. McShan, D. Ajdic, D. J. Savic, G. Savic, K. Lyon, C. Primeaux, S. Sezate, A. N. Suvorov, S. Kenton, H. S. Lai, S. P. Lin, Y. Qian, H. G. Jia, F. Z. Najar, Q. Ren, H. Zhu, L. Song, J. White, X. Yuan, S. W. Clifton, B. A. Roe, and R. McLaughlin. 2001. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc. Natl. Acad. Sci. USA 98:46584663.
11. Fischer, D. and D. Eisenberg. 1997. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. Proc. Natl. Acad. Sci. USA 94:11929-11934.
12. Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, and J. M. K. and. 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270:397-403.
13. Garcia-Vallve, S., A. Romeu, and J. Palau. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. 10:1719–1725.
14. Gelfand, M. S. and E. V. Koonin. 1997. Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. Nucleic Acids Res. 25:2430-2439.
15. Glass, J. I., E. J. Lefkowitz, J. S. Glass, C. R. Heiner, E. Y. Chen, and G. H. Cassell. 2000. The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. Nature 407:757-762.
16. Gumulak-Smith, J., A. Teachman, A. H. Tu, J. W. Simecka, J. R. Lindsey, and K. Dybvig. 2001. Variations in the surface proteins and restriction enzyme systems of *Mycoplasma pulmonis* in the respiratory tract of infected rats. Mol. Microbiol. 40: 1037-1044.

17. Hamet, M., C. Bonissol, and P. Cartier. 1979. Activities of enzymes of purine and pyrimidine metabolism in nine Mycoplasma species. Adv. Exp. Med. Biol. 122B:231-235.
18. Herrmann, R. and B. Reiner. 1998. Mycoplasmapneumoniae and Mycoplasma genitalium: a comparison of two closely related bacterial species. Curr. Opin. Microbiol. 1:572-579.
19. Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B. C. Li, and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium Mycoplasmapneumoniae. Nucleic Acids Res. 24:4420-4449.
20. Himmelreich, R., H. Plagens, H. Hilbert, B. Reiner, and R. Herrmann. 1997. Comparative analysis of the genomes of the bacteria Mycoplasmapneumoniae and Mycoplasma genitalium. Nucleic Acids Res. 25:701-712.
21. Hutchison III C. A., S. N. Peterson, S. R. Gill, R. T. Cline, O. White, C. M. Fraser, H. O. Smith, and J. C. Venter. 1999. Global transposon mutagenesis and a minimal Mycoplasma genome. Science 286:2165–2169.
22. Huynen, M., B. Snel, , and P. Bork. 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res. 10:1204-12 10.
23. Hnynen, M., T. Doerks, F. Eisenhaber, C. Orengo, S. Sunyaev, Y. Yuan, and P. Bork. 1998. Homology-based fold predictions for Mycoplasma genitalium proteins. J. Mol. Biol. 280:323-326.
24. Karlin, S. and A. M. Campbell. 1994. Which bacterium is the ancestor of the animal mitochondria1 genome? Proc. Natl. Acad. Sci. USA 91:12842-12846.
25. Kokotovic, B., G. Bolske, P. Ahrens, and K. Johansson. 2000. Genomic variations of Mycoplasma capricolum subsp. capripneumoniae detected by amplified fragment length polymorphism (AFLP) analysis. FEMS Microbiol. Lett. 184:63-68.
26. Koonin, E. V. and P. Bork. 1996. Ancient duplication of DNA polymerase inferred from analysis of complete bacterial genomes. Trends Biochem. Sci. 21:128–129.
27. Krause, D. C. 1996. Mycoplasmapneumoniae cytadherence: unravelling the tie that binds. Mol. Microbiol. 20:247-253.
28. Manolukas, J. T., M. F. Barile, D. K. Chandler, and J. D. Pollack. 1988. Presence of anaplerotic reactions and transamination, and the absence of the tricarboxylic acid cycle in mollicutes. J. Gen. Microbiol. 134 (Pt 3):791-800.
29. Miyata, M. and S. Seto. 1999. Cell reproduction cycle of mycoplasma. Biochimie 81:873-878.
30. Musco, G., G. Stier, C. Joseph, M. A. Castiglione Morelli, M. Nilges, T. J. Gibson, and A. Pastore. 1996. Three-dimensional structure and stability of the KH domain: molecular insights into the fragile X syndrome. Cell 85:237-245.
31. Mushegian, A. R. and E. V. Koonin. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc. Natl. Acad. Sci. USA 93: 10268-10273.
32. Neimark, H. C. and C. S. Lange. 1990. Pulse-field electrophoresis indicates full-length Mycoplasma chromosomes range widely in size. Nucleic Acids Res. 18:5443-5448.
33. Pollack, J. D. 1997. Mycoplasma genes: a case for reflective annotation. Trends Microbiol. 5:413-419.
34. Pollack, J. D. 2001. *Ureaplasma urealyticum:* an opportunity for combinatorial genomics. Trends Microbiol. 9:169–175.
35. Pollack, J. D., M. V. Williams, and R. N. McElhaney. 1997. The comparative metabolism of the mollicutes (Mycoplasmas): the utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. Crit. Rev. Microbiol. 23:269-354.

36. Pyle, L. E., L. N. Corcoran, B. G. Cocks, A. D. Bergemann, J. C. Whitley, and L. R. Finch. 1988. Pulsed-field electrophoresis indicates larger-than-expected sizes for mycoplasma genomes. Nucleic Acids Res. 16:6015-6025.
37. Robertson, J. A., L. E. Pyle, G. W. Stemke, and L. R. Finch. 1990. Human ureaplasmas show diverse genome sizes by pulsed-field electrophoresis. Nucleic Acids Res. 18:1451-1455.
38. Rychlewski, L., B. Zhang, and A. Godzik. 1998. Fold and function predictions for *Mycoplasma genitalium* proteins. Fold Des. 3:229-238.
39. Snel, B., G. Lehmann, P. Bork, and M. A. Huynen. 2000. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res. 28:3442-3444.
40. Snel, B., P. Bork, and M. A. Huynen. 1999. Genome phylogeny based on gene content. Nat. Genet. 21:108–110.
41. Sokal, R. R. and C. D. Michener. 1958. A statistical method of evaluating systematic relationships. Univ. Kansas Sci. bull. 28: 1409-1438.
42. Suyama, M. and P. Bork. 2001. Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet. 17: 10-13
43. Teichmann, S. A. and G. Mitchison. 1999. Is there a phylogenetic signal in prokaryote proteins? J. Mol. Evol. 49:98-107.
44. Teichmann, S. A., A. G. Murzin, and C. Chothia. 2001. Determination of protein function, evolution and interactions by structural genomics. Curr. Opin. Struct. Biol. 11:354-363.
45. Teichmann, S. A., C. Chothia, and M. Gerstein. 1999. Advances in structural genomics. Curr. Opin. Struct. Biol. 9:390-399.
46. Trachtenberg, S. 1998. Mollicutes-wall-less bacteria with internal cytoskeletons. J. Struct. Biol. 124:244-256.
47. *Wolf, Y. I., N. V. Grishin, and E. V. Koonin.* 2000. Estimating the number of protein folds and families from complete genome data. J. Mol. Biol. 299:897-905.