Fig. 1. Multiple alignment showing representative members of the SLH-domain family. Residues not belonging to an SLH domain are shown as lower-case letters and residues conserved in more than half of all known SLH domains are shown in bold. The proteins are: *Thermotoga maritima* Ompα, Tm Omp; *Thermus thermophilus* S-layer protein, Tt SLP; *Acetogenium kivui* S-layer protein, Ak SLP; *Bacillus sphaericus* 2362 S-layer protein, Bs SLP; *C. thermocellum* exo-β-1,4-cellobiohydrolase, Ct XynX; *Bacillus* sp. XAL-601 endo-α-1,4- and exo-α-1,6-glucosidase (α-amylase/pullulanase), B AapT; *C. thermocellum* outer layer protein B, Ct OlpB; *C. thermocellum* ORF2 3′ of *cipA*, Ct Orf2; and *C. thermocellum* ORF3 3′ of *cipA* (putative cellulosome anchoring protein), Ct Orf3. The numbering (in which signal sequences are included) refers to the first residue in each sequence. Amino-termini are marked by (^) and carboxy-termini by (*). The predicted secondary structure (H = α-helix, S = β-strand) is the average of the consensus predictions for the entire SLH-domain family by the methods of Chou and Fasman (1978, *Annu Rev Biochem* **47:** 251–276), Garnier *et al.* (1978, *J Mol Biol* **120:** 97–120), Rost and Salander (1993, *J Mol Biol* **232:** 584–599) and Solovyev and Salamov (1994, *CABIOS* **10:** 661–669).

may have occurred in the evolution of these proteins. This would imply that the three SLH domains form a compact, circular structure in which the amino-terminus of the first domain is close to the carboxy-terminus of the third domain, rather than a loose, linear arrangement. Such a compact structure is in agreement with the globular appearance of SLH domains in electron micrographs (Engel *et al.*, 1992, *EMBO J* **11:** 4369–4378; Lupas *et al.*, 1994, ibid.).

Andrei Lupas
Max-Planck-Institut für Biochemie, D-82152 Martinsried, Germany.
E-mail: lupas@vms.biochem.mpg.de; Tel (89) 8578 2646; Fax (89) 8578 2641.

Received 20 February, 1996; accepted 29 February, 1996.

Novelties from the complete genome of *Mycoplasma genitalium*

Sir,

With the recent publication of the complete genome of *Mycoplasma genitalium* (Fraser *et al.*, 1995, *Science* **270:** 397–403), we are able to explore the functional potential of a minimal cell, indeed much smaller than a related species, *Mycoplasma capricolum,* that we have recently analysed (Bork *et al.*, 1995, *Mol Microbiol* **16:** 955–967). To supplement or correct the original functional annotations of the genes of *M. genitalium* (Fraser *et al.*,

1995, *Science* **270:** 397–403), we report here additional function predictions for a number of gene products, the result of a thorough analysis of the complete genome immediately after publication. Apart from listing new functions, we also discuss the accuracy of different approaches to sequence analysis, as well as genome composition. We assume that our careful manual analysis, using the automatically generated data, yields the best results; taking these as the 'benchmark', we can obtain estimates for the accuracy of function predictions, using various approaches. As we have analysed this genome automatically using the GENEQUIZ software system (Casari *et al.*, 1995, *Nature* **376:** 647–648; Casari *et al.*, 1996, *First Annual Pacific Symposium on Biocomputing*, pp. 707–709; Scharf *et al.*, 1994, *Intelligent Systems for Molecular Biology 1994*, pp. 348–353) as well as manually with great care, we are in a position to estimate the overall accuracy of our analyses. Benchmarking results show that the accuracy of the system has reached a level of about 96% (1 false negative and 11 false positives out of 285 assignments), in contrast to the TIGR annotations, which reach an accuracy of 86% (40 false assignments out of 285). This margin of 10 percentage points can represent missing a couple of hundred new functional annotations in a genome such as that of *Haemophilus influenzae* (Casari *et al.*, 1995, *Nature* **376:** 647–648) or even thousands of proteins for eukaryotic genomes.

The technical advantages of the GENEQUIZ analysis come mainly from two sources: (i) full automation and therefore the availability of the latest and most updated database releases (Scharf *et al.*, 1994, *Intelligent Systems for Molecular Biology 1994*, pp. 348–353), and (ii) a combination of sophisticated algorithms that screen out false positives in combination with a number of well-tested empirical rules (Casari *et al.*, 1996, *First Annual Pacific Symposium on Biocomputing*, pp. 707–709). Below, we show 21 new functions identified by GENEQUIZ (Table 1) and 29 more cases for which no function should be associated with the corresponding open reading frame (ORF) (Table 2). We list the sources of error that have led to these incorrect annotations.

First, the 21 new functions represent various activities that are found in other bacterial species. The functional assignments vary between precise (such as enzyme names) and imprecise (such as gene names) predictions, where the function may be implied by sequence similarity to other proteins. The difficulties of function assignment by sequence similarity have been discussed elsewhere (Ouzounis *et al.*, 1995, *Prot Sci* **4:** 2424–2428). Some of the most interesting findings are as follows: arginine deiminase (mg123), the second amino-acid-metabolising enzyme that has been identified in this species and known to exist in *Mycoplasma arginini* (Kondo *et al.*, 1990, *Mol Gen Genet* **221:** 81–86; Ohno *et al.*, 1990, *Infect Immun* **58:** 3788–3795) and *Mycoplasma hominis* (Harasawa *et al.*, 1992, *Microbiol Immunol* **36:** 661–665), two phoshodiesterases (mg333, mg385), two aminoacyl-tRNA synthetase homologues (mg237, mg449), and DNA polymerase I (mg468) (Table 1).

Table 1. New functions found by GENEQUIZ, which were previously characterized either as having 'no database match' or matching a 'hypothetical protein' (Fraser *et al.*, 1995).

mg#	TIGR annotation	GQ annotation
mg123	Hypothetical	Arginine deiminase
mg125	Hypothetical	Hydrolase
mg132	Hypothetical	Hit1 protein
mg139	Hypothetical	Amps (fragment)
mg225	Hypothetical	Histidine permease?
mg245	Hypothetical	5,10-metTHF synthetase
mg263	Hypothetical	Hydrolase
mg265	Hypothetical	Hydrolase
mg270	Hypothetical	Lipoate-protein ligase A
mg294	Hypothetical	NarK, nitrite extrusion protein
mg326	Hypothetical	DegV protein
mg442	Hypothetical	GTP-binding protein
mg464	Hypothetical	Stage III sporulation protein J
mg140	Unknown	DNA-binding protein S mu bp-2
mg237	Unknown	Isoleucyl-tRNA synthetase domain
mg329	Unknown	GTP-binding protein
mg333	Unknown	ACP phosphodiesterase
mg377	Unknown	Zinc protease
mg385	Unknown	Glycerol-P diester phosphodiesterase
mg449	Unknown	Phe-tRNA synthetase N-terminal
mg468	Unknown	DNA polymerase I[a]

These represent false negative assignments by the original authors, since various functions (at different levels of precision) can be identified.
a. Identical to mg262 (unclear case — may be due to contig assembly).

Second, we have identified a fair number of false positives, i.e. 29 functional assignments for which insufficient information exists in support. These assignments come from the wrong interpretation of similarity searches usually with the introduction of composition bias effects and long gaps. These are incorrect predictions, with or without homology, and also over- or under-predictions, errors due to wrong database annotations, e.g. from neighbouring ORFs, or simply erroneous manual annotations (Table 2).

There is only one case, oligoendopeptidase F (mg183), that our automatic analysis has missed. Therefore, with 11 false positives (which we have now corrected) and one false negative out of 285 clear assignments, we have an error rate of only 4.2%, or an accuracy of 95.8%, 10 percentage points higher than the original annotation.

The most extraordinary elements in the genome of *M. genitalium* seem to be four genes with eukaryotic homologues with no known bacterial counterparts either in well-studied species such as *Escherichia coli*, or in the complete genome of *H. influenzae* (data not shown). In the original publication, no comment was provided for these cases, although they represent some intriguing homologies. These are three previous findings: a pre-B-cell colony-enhancing factor

Table 2. Corrections provided to the original TIGR annotations (Fraser *et al.*, 1995), based on evidence from our analysis (type of error also listed).

mg#	TIGR annotation	GO annotation: false positives	GO annotation: type[a]
mg032	AddA protein		E
mg061	UhpT protein		E
mg067	SPase		E
mg085	Reductase		E
mg090	Ribosomal protein S6		G
mg098	p48 eggshell protein		D, E
mg120	rbsC protein		E
mg219	IgA1 protease		E
mg220	Pre-procytoxin vacA		E
mg269	Surface antigen pag		E
mg288	Protein L		E
mg328	Protein V (fcrV)		E
mg364	Mobilization protein mob13		D
mg406	Transport permease P69		E
mg459	Surface exclusion protein prgA		E
		Other errors	
mg137	RfbD, reductase	Amine oxidase	A
mg217	XynA, xylanase	P65 protein	A, D
mg278	Rel protein	Pyrophosphohydrolase	A
mg310	Proline iminopeptidase	Triacylglycerol lipase	A
mg318	Fibronectin-binding protein	Adhesin-related protein	E
mg396	Galactosidase acetyl-transferase	g6p isomerase	A
mg409	PhoU membrane protein	Pho negative regulator	A, F
mg006	Thymidylate kinase	Putative kinase	B
mg041	Phosphotransferase	*ptsH* gene, HPR	B
mg248	Sigma factor	OrfA adjacent to sigma	B, F
mg356	Lic-1	Unclear	B, G
mg099	Aux2 hydrolase	Indoleacetamide hydrol	C
mg145	protein X	FAD synthetase	C
mg194	PheS beta chain	PheS alpha chain	F

a. Types of error: A, inaccurate prediction; B, overprediction; C, underprediction; D, composition bias; E, low gap penalties/too many gaps; F, manual annotation; G, unclear.

(mg037), a PET112 homologue (mg100), and a serine/threonine protein kinase (mg109), complemented by a new one, i.e. the identified homologue of human S mu bp-2 protein (mg140).

The pre-B-cell colony-enhancing factor is a putative cytokine for early B cells (Samal *et al.*, 1994, *Mol Cell Biol* **14:** 1431–1437). PET112 is involved in regulating mitochondrial transcripts, and, in particular, cytochrome *c* oxidase subunit II (Mulero *et al.*, 1994, *Curr Genet* **25:** 299–304). Protein kinases are very unusual in bacteria (Munoz-Dorado *et al.*, 1991, *Cell* **67:** 995–1006). Finally, the homology to the human protein involved in DNA binding of immunoglobulin mu (Fukita *et al.*, 1993, *J Biol Chem* **268:** 17463–17470), is peculiar to this bacterium. The homology to two proteins involved in immune-system regulation may be linked to the pathogenicity of this prokaryote.

Another interesting fact is the absence of the transcription factors from *M. genitalium*. For a comparison, *E. coli* has 55 known transcriptional activators and 58 repressors (C. Ouzounis, unpublished observations). Apart from *nusA* (mg141) and *nusG* (mg054), *M. genitalium* contains a single transcription elongation factor, GreA (mg282), incorrectly classified by TIGR as a translation factor (Fraser *et al.*, 1995, *Science* **270:** 397–403). It is not clear how transcription is regulated in this organism, with such a small number of factors.

With these new findings, the functional composition of the genome can be slightly modified, but the major classes (Ouzounis *et al.*, 1995, *European Conference on Artificial Life 1995 (ECAL95)*, pp. 843–851; Tamames *et al.*, 1996, submitted) remain the same: 35% of all products are involved in translation, 29% in metabolism, 17% in DNA and RNA-related processes (replication, repair, transcription and regulation), and another 10% in transport. The remaining 9% of the genome codes for proteins involved in protein processing, signalling, structural roles and communication with the environment. It is interesting that such a reduced genome has lost components from almost every cellular process, especially transcription, while keeping translation intact, dedicating most of its genes to this indispensable process.

Acknowledgements

C.O. is a recipient of a long-term Human Frontiers Science Programme Fellowship.

Note

The corrected TIGR table can be accessed on the World Wide Web at the URL: http://www.ai.sri.com/~ouzounis/mg_embl.html and the GENEQUIZ results at the URL: http://saturn.ebi.ac.uk:8421/mycogen.html. The effective date of our searches was October 24th, 1995 (four days after publication date) and updates will be regularly provided at the above site.

Christos Ouzounis,[1]* Georg Casari,[2] Alfonso Valencia[3] and Chris Sander[4]

[1]*AIC–SRI International, 333 Ravenswood Avenue, Menlo Park, California 94025-3493.*

[2]*EMBL, Heidelberg, Germany.*

[3]*CNB–CSIC, Madrid, Spain.*

[4]*EMBL–EBI, Cambridge, UK.*

**For correspondence. E-mail ouzounis@ai.sri.com; Tel. (415) 859 2159; Fax (415) 859 3735.*

Received 22 January, 1996; accepted 15 February, 1996.