Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv

Jean-Christophe Camus,^{1,2} Melinda J. Pryor,^{1,2} Claudine Médigue³ and Stewart T. Cole¹

Author for correspondence: Stewart T. Cole. Tel: +33 1 45688446. Fax: +33 1 40613583.
e-mail: stcole@pasteur.fr

¹ Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex, France

² Annotation-Bases de Données (PT4), Génopole, Institut Pasteur, Paris, France

³ Géoscope/UMR 8030, Atelier de Génomique Comparative, 2 rue Gaston Crémieux, 91006 Evry Cedex, France

Original genome annotations need to be regularly updated if the information they contain is to remain accurate and relevant. Here the complete re-annotation of the genome sequence of *Mycobacterium tuberculosis* strain H37Rv is presented almost 4 years after the first submission. Eighty-two new protein-coding sequences (CDS) have been included and 22 of these have a predicted function. The majority were identified by manual or automated re-analysis of the genome and most of them were shorter than the 100 codon cut-off used in the initial genome analysis. The functional classification of 643 CDS has been changed based principally on recent sequence comparisons and new experimental data from the literature. More than 300 gene names and over 1000 targeted citations have been added and the lengths of 60 genes have been modified. Presently, it is possible to assign a function to 2058 proteins (52% of the 3995 proteins predicted) and only 376 putative proteins share no homology with known proteins and thus could be unique to *M. tuberculosis*.

Keywords: mycobacteria, tuberculosis, genomics

INTRODUCTION

Since the completion of the first prokaryotic genome sequence (Fleischmann *et al.*, 1995) the number of sequencing projects and related biological databases have been increasing exponentially. This has in turn led to the development of downstream sciences that take advantage of this sequence information such as comparative genomics, transcriptomics, proteomics and metabolomics. Significant advances have been made in these fields, thereby increasing our knowledge of the functions of many gene products. It is critical, therefore, that genome annotations are frequently updated if the information they contain is to remain accurate, relevant and useful. Several other genomes have thus been re-annotated recently (Dandekar *et al.*, 2000; Serres *et al.*, 2001; Gaasterland & Oprea, 2001; Bocs *et al.*, 2002).

Mycobacterium tuberculosis H37Rv was first isolated in 1905, has remained pathogenic and is the most widely used strain in tuberculosis research. The complete genome sequence and annotation of this strain was published in 1998 (Cole *et al.*, 1998). The information from this project was incorporated into the public database TubercuList (<http://genolist.pasteur.fr/>

TubercuList/) which was created using the GenoList model (Moszer *et al.*, 2002).

In this paper we describe the re-annotation of the *M. tuberculosis* H37Rv genome. We have manually re-evaluated each of the coding sequences (CDS) previously annotated and present the combined results of recent database searches and literature surveys. This annotation also contains new comparisons with the recently completed genome sequence of *Mycobacterium leprae* (Cole *et al.*, 2001).

METHODS

Sequence analysis and annotation. Each of the *M. tuberculosis* H37Rv CDS previously predicted and annotated (Cole *et al.*, 1998) has been manually re-analysed based on the results of BLASTP (Altschul *et al.*, 1990) and FASTA (Pearson & Lipman, 1988) sequence comparisons using non-redundant data from the EMBL, TrEMBL and SWISS-PROT databases. Additional functional insight was obtained by using the PROSITE database (Falquet *et al.*, 2002) and the programs TMHMM 1.0 (Sonnhammer *et al.*, 1998) and SIGNALP (Nielsen *et al.*, 1997) to predict subcellular localization. Information about transport proteins has been incorporated from recent expert reviews (Braibant *et al.*, 2000; <http://www.biology.ucsd.edu/~ipaulsen/transport/>; Paulsen *et al.*, 2000). The coding sequence length has been re-evaluated in particular by examining *M. tuberculosis* protein families

.....
Abbreviation: CDS, protein-coding sequences.

generated with the MEME/MAST program as described previously (Tekaiia *et al.*, 1999). Re-annotation was supported by Artemis software release 4 (<http://www.sanger.ac.uk/>; Rutherford *et al.*, 2000). This update has recently been deposited in EMBL (accession no. AL123456) and in our TubercuList website (<http://genolist.pasteur.fr/TubercuList>). In this release (R4), a number of new features has been included where possible: SWISS-PROT accession number, synonyms, EC number and catalytic reaction for enzymes, protein family name, most important references from the literature with a qualifier for proteins studied experimentally, information from numerous published proteomic or transcriptomic studies, *M. leprae* orthologues, name of the putative product and a short description of the predicted function. A table with all of the predicted CDS and the corresponding functional classification, adapted from the work of Riley (1993), is available from TubercuList.

Identification of new CDS. Three independent approaches were used for detecting new potential coding sequences. In the first, some new CDS with appropriate GC content, correlation scores and codon usage were found manually during the re-annotation of the genome. In the second, a new program, AMIGA (Automatic Microbial Genome Annotation), was used to identify possible frameshifts and potential coding sequences that had been overlooked (for details see Bocs *et al.*, 2002). Briefly, AMIGA found the most likely CDS longer than 60 bp and merged the results with those generated by a modified GeneMark analysis. The combined results were then compared with the original annotation and the additional CDS detected by AMIGA investigated further using the criteria in the first approach and database searches. In the third approach, other CDS were found following TBLASTN searches of TubercuList using protein sequence data from the literature (Jungblut *et al.*, 2001; Rosenkrands *et al.*, 2000a; Corixa Corporation patent no. WO 97/09428, 1997) or personal communications (N. Stoker, P. Jungblut).

RESULTS AND DISCUSSION

Revising the number of genes in the genome

The original sequence and annotation of *Mycobacterium tuberculosis* strain H37Rv identified 3974 genes (Cole *et al.*, 1998). This included 3924 genes thought to encode proteins and 50 encoding stable RNA. Following the re-annotation, we have included 82 additional genes. All of the new genes are believed to encode polypeptides and no change has been detected in the number of RNA molecules. The numbering of the new CDS has not interfered with the labelling of the existing genes (Table 1). The new CDS use the same Rv number as the preceding CDS followed by a letter (e.g. Rv2307A, Rv2307B or Rv2307D have been introduced between CDS Rv2307c and Rv2308). The 82 new CDS are inventoried in Table 1 and comprise 75 found by re-examining the genome using manual or automatic analysis methods (see Methods) and seven that were included based on experimental results from other laboratories.

The cut-off for gene length used during the initial analysis of the genome was 100 codons (Cole *et al.*, 1998) and unknown genes smaller than this are often difficult to identify reliably by predictive methods using

hidden Markov models or by other approaches that do not use sequence similarity. Lowering the gene length threshold, coupled with checking for appropriate codon usage, resulted in the identification of 27 putative new CDS with a median length of 207 bases, e.g. Rv0979A (product: 57 aa) and Rv0634B (product: 55 aa). These two protein-coding genes share homology with the products of *rpmF* and *rpmG*, respectively, and in several other micro-organisms they encode part of the ribosomal machinery. Another gene, Rv1028A, was identified because we have changed the parameters of the BLAST program and thus detected similarities with other orthologues. This short CDS, termed *kdpF*, is localized in the *kdp* cluster involved in phosphate transport and is thought to play a role in the stabilization of the *kdp* system in *Escherichia coli* (Gassel *et al.*, 1999).

Using AMIGA, an in-house program employing *M. tuberculosis* codon usage (see Methods), we have found 30 other CDS whose putative products comprise between 45 and 211 aa (Table 1; see AMI). With these different methods, nine new CDS replaced genes identified in the original annotation, but localized on the opposite strand. In these cases, the same Rv number has been kept but the letter extension c, denoting localization on the complementary strand, has been added or removed as appropriate. AMIGA also predicts real or potential frameshifts, and these results led to the correction of four sequencing errors. The current nucleotide sequence now contains 4411532 nt.

Seven other CDS were identified using data from the literature. Thus, the proteomic study of Jungblut *et al.* (2001) identified five new proteins by two-dimensional electrophoresis and mass spectrometry. Two other CDS were uncovered using the results of antigen discovery programmes of the Corixa Corporation (Patent no. WO 97/09428, 1997) and the Statens Serum Institute (Rosenkrands *et al.*, 2000a), respectively.

All new CDS have been classified into one of the 11 functional classes used by Cole *et al.* (1998) and described in Table 2. For 60 of the putative proteins we were unable to predict a precise function and this group includes proteins classified as conserved hypothetical proteins (class 10), unknown proteins without homology (class 8), proteins putatively involved in cell-wall processes or localized in the membrane (class 3) and two PE family proteins (class 6). Functions were predicted for 22 of the new proteins, e.g. RpmF or KdpF as discussed previously and classified in classes 2 (information pathways) or 3 (cell-wall and cell processes), respectively.

Updating the annotation of CDS

To enhance the value of the annotation, systematic re-analysis has been undertaken to include principally the findings from reiterative BLASTP/FASTA searches and new scientific data from the literature. We have updated all the protein-coding genes previously identified in 1998 and tried to assign new or more precise functions when possible. This re-annotation led to changes in the

Table 1. Rv numbers of new and removed CDS, and CDS with lengths changed

New CDS identified*
<p>MAN: Rv0078A, Rv0157A, Rv0192A, Rv0236A, Rv0257, Rv0492A, Rv0500A, Rv0500B, Rv0521, Rv0609A, Rv0634B/<i>rpmG</i>, Rv0841, Rv0979A/<i>rpmF</i>, Rv1000c, Rv1028A/<i>kdpF</i>, Rv1087A, Rv1089A/<i>celA2a</i>, Rv1135A, Rv1290A, Rv1507A, Rv1508A, Rv1990A, Rv2077A, Rv2307A, Rv2307B, Rv2307D, Rv2309A, Rv2331A, Rv2438A, Rv2530A, Rv2601A, Rv2614A, Rv2922A/<i>acyP</i>, Rv2943A, Rv2970A, Rv2998A, Rv3018A/PE27A, Rv3197A/<i>whiB7</i>, Rv3198A, Rv3221A, Rv3294c, Rv3566A, Rv3705A, Rv3724A, Rv3724B</p> <p>AMI: Rv0470A, Rv0590A, Rv0724A, Rv0749A, Rv0755A, Rv1116A, Rv1322A, Rv1473A, Rv1489, Rv1489A, Rv1575A, Rv1638A, Rv1706A, Rv1765A, Rv2063, Rv2160A, Rv2219A, Rv2250A, Rv2306A, Rv2306B, Rv2401A, Rv2737A, Rv2803, Rv3022A/PE29, Rv3224A, Rv3224B, Rv3395A, Rv3678A, Rv3770A, Rv3770B</p> <p>MPI: Rv0634A, Rv0787A, Rv1159A, Rv1498A, Rv3196A</p> <p>SSI: Rv3208A</p> <p>COR: Rv3312A</p>
CDS length changes†
<p>N-terminus changed: Rv0183 (−44), Rv0492c (−55), Rv0774c (−9), Rv0966c (−30), Rv0979c (−44), Rv1046 (−65), Rv1055 (−34), Rv1105 (−9), Rv1111c (+23), Rv1171 (+26), Rv1223/<i>htrA</i> (−21), Rv1327c/<i>glgE</i> (−36), Rv1330c (−61), Rv1440/<i>secG</i> (−40), Rv1453 (−11), Rv1460 (−10), Rv1482c (−59), Rv1486c (−14), Rv1499 (−24), Rv1516c (−37), Rv1572 (+7), Rv1625 (+25), Rv1721c (+3), Rv1766 (+27), Rv1807 (+4), Rv1917c/PPE34 (+23), Rv2013 (−74), Rv2014 (−31), Rv2086 (−16), Rv2188c (−14), Rv2205c (−122), Rv2206 (+112), Rv2250c (−30), Rv2301/<i>cut2</i> (+11), Rv2332/<i>mez</i> (−104), Rv2401 (+37), Rv2438c/<i>nadE</i> (−59), Rv2770c/PPE44 (−20), Rv2870c/<i>dxr</i> (−23), Rv2922c/<i>smc</i> (−84), Rv2930/<i>fadD26</i> (−43), Rv2960c (−34), Rv3287c/<i>rsbW</i>, (−97), Rv3353c (−28), Rv3395c (−90), Rv3451/<i>cut3</i> (+15), Rv3604c (−65), Rv3674c/<i>nth</i> (+19), Rv3681/<i>whiB4</i> (+18), Rv3722c (+27)</p> <p>C-terminus changed: Rv0164 (−24), Rv0857 (+33), Rv1054 (+36), Rv1395 (−14), Rv2232 (+135), Rv2516c (+18), Rv2631 (+175), Rv3337 (+47), Rv3566c/<i>nat</i> (−49), Rv3799c/<i>accD4</i> (+5)</p> <p>Previous CDSs replaced (Rv number from original annotation): Rv0257c, Rv0521c, Rv0841c, Rv1000, Rv1489c, Rv2063c, Rv2306c, Rv2803c, Rv3294</p> <p>CDS removed (and not replaced): Rv2233</p>

* MAN, determined manually; AMI, identified by AMIGA; MPI, determined by studies of the Max-Planck-Institute, Berlin, Germany; COR, found by Corixa Corporation patent analysis; SSI, characterized by the Statens Serum Institute, Copenhagen, Denmark.

† The number between parentheses corresponds to the number of amino acids added (+) or removed (−) for the corresponding product.

functional classification of more than 600 CDS, the annotation of ~350 other CDS has been changed without affecting their classification and almost 300 gene names have been added. In addition, we have altered the related product description of 33 CDS, in many cases making this more specific, and changed the gene name of >50 existing CDS. Where gene names have been changed, the old name is retained within the note. Over 1000 targeted literature citations have been added to support the functional information.

In addition, during the new analysis, the lengths of 60 coding sequences have been altered based on the results of BLASTP/FASTA and MEME/MAST or due to correction of the nucleotide sequence (Table 1). Thus 35 CDS have been shortened and 15 extended at their 5' ends, whereas

10 have been altered at their 3' ends due to sequence corrections or errors in the original annotation. One gene, Rv2233, has been removed, but not replaced as this merged with the preceding CDS, Rv2232. When the coding sequence of a previously determined gene has been shortened, we have checked for the possible presence of a new CDS in another reading frame. So, for example, after shortening the *smc* gene, encoding a putative chromosome segregation protein, the *acyP* gene (Rv2922A) was identified.

Changes to functional classes

The functional classification of 643 of the predicted proteins of *M. tuberculosis* H37Rv identified in 1998 has been changed during the update (Fig. 1a). Not unexpect-

Table 2. Functional classification of *Mycobacterium tuberculosis* genes

Class	Function	Colour	No. of new genes	Gene no. (1998)	Gene no. (2002)
0	Virulence, detoxification, adaptation		1	91	99
1	Lipid metabolism		1	225	233
2	Information pathways		3	207	229
3	Cell-wall and cell processes		11	516	708
4	Stable RNAs		0	50	50
5	Insertion sequences and phages		6	137	149
6	PE and PPE proteins		2	167	170
7	Intermediary metabolism and respiration		6	877	894
8	Proteins of unknown function		14	606	272
9	Regulatory proteins		3	188	189
10	Conserved hypothetical proteins		35	910	1051

tedly, the two functional classes exhibiting the greatest number of transfers to other classes were the unknown category with 354 changes (class 8) and the conserved hypotheticals category with 183 changes (class 10). Figs 1(b, c) show the functional categories to which the protein-coding genes annotated as unknown and conserved hypothetical proteins in 1998 have moved.

During the re-annotation, functions have been postulated for 94 unknown proteins based on new sequence similarities with other proteins or experimental data from the literature (studies on *M. tuberculosis* or other organisms). For example, the putative protein encoded by Rv2476c (class 8) now shows consistent homology with several *gdh* products, NAD⁺-dependent glutamate dehydrogenases (EC 1.4.1.2), which have been well characterized recently in other bacteria (Kersten *et al.*, 1999; Minambres *et al.*, 2000; Lu & Abdelal, 2001). Consequently, Rv2476c has been transferred to class 7. However, the majority of the class 8 proteins (245) have been reclassified now as conserved hypothetical proteins without any indication of a function. These changes are principally due to an increase in genomic data generated from sequencing projects in the last 4 years. Notably, many of the *M. tuberculosis* CDS have orthologues in *M. leprae* (Cole *et al.*, 2001; <http://genolist.pasteur.fr/Leproma/>) and *Streptomyces coelicolor* (Bentley *et al.*, 2002; http://www.sanger.ac.uk/Projects/S_coelicolor/).

Many of the class 10 and class 8 proteins have been reclassified in the cell wall and cell process category (class 3) (Fig. 1b, c). There are two reasons for these changes. First, the criteria for classification in this group (class 3) have been amended since 1998. Class 3 now comprises all predicted membrane proteins or proteins

believed to be involved in a cell process (including secreted and transmembrane proteins), regardless of whether they have similarities with other proteins or a predicted function. For example in 1998, Rv0970, encoding an integral membrane protein with no similarity to other proteins, was classified as unknown (class 8). Second, some of the predicted proteins have moved to the cell-wall and cell processes group because of new information from the databases or new data from research. For example, Rv2450c is believed to encode a bacterial growth factor or cytokine involved in promoting the resuscitation and growth of dormant cells (Mukamalova *et al.*, 1998) and has been renamed *rpfE* (M. Young, personal communication).

There have also been transfers from the other functional groups as follows (Fig. 1a): 63 transfers from class 7 (intermediary metabolism and respiration), 24 transfers from class 3, 11 transfers from class 9 (regulatory proteins), 6 transfers from class 1 (lipid metabolism), 1 transfer from class 0 (virulence, detoxification or adaptation) and 1 transfer from class 2 (information pathways). These transfers often involve a change from a predicted function in one class to a more precise function in another (e.g. 7 to 1), but can also involve a regression. For example, Rv3522 was predicted to be a transcriptional regulatory protein in the first annotation (class 9), but re-analysis of its sequence leads us to consider it to be involved in lipid metabolism (class 1). The total number of regressions was 82, the largest group involving transfer from class 7 to 10. The original annotation of these putative proteins involved either incorrect analysis of the results or the existence of errors in the database used. For example, the CDS Rv0382c was originally annotated as a probable uridine 5'-monophosphate synthase based on the best similarities

Fig. 1. Changes to the functional classification of protein-coding genes following the re-annotation of the *M. tuberculosis* H37Rv genome. (a) The percentage of protein-coding genes changed in each of the functional classes. (b) The functional classes where the protein-coding genes originally annotated as unknown function (class 8) have moved. (c) The functional classes where the protein-coding genes originally annotated as conserved hypothetical proteins (class 10) have moved.

after BLASTP or FASTA analysis. However, this assignment is misleading as uridine 5'-monophosphate synthase is a bifunctional enzyme containing orotate phosphoribosyltransferase activity (EC 2.4.2.10) at its N terminus and orotidine 5'-phosphate decarboxylase activity (EC 4.1.1.23) at its C terminus. Several proteins

annotated as uridine 5'-monophosphate synthases in the databases share similarity only at their N terminus and these are therefore incorrectly annotated because they lack the second domain found in authentic uridine 5'-monophosphate synthases. We thus consider the Rv0382c product to be an orotate phosphoribosyltransferase and not a uridine 5'-monophosphate synthase as predicted in the first annotation. Note that in certain regressions, gene names attributed in 1998 have been removed.

Changes within functional classes

Updating the genome annotation of *M. tuberculosis* H37Rv has also resulted in many changes within the functional classes usually due to new information from the literature. These have included updating the product names, changing 58 specific gene names and introducing appropriate new citations. Gene names were included when there was significant similarity to a pertinent publication. For example, the previously annotated *umaA2* (Rv0470c), unknown mycolic acid synthase, has recently been changed to *pcaA*, encoding an S-adenosyl methionine (SAM)-dependent methyl transferase, required for α -mycolic acid cyclopropanation and lethal chronic persistence in *M. tuberculosis* infection (Glickman *et al.*, 2000). On occasion, a gene name has been given to a CDS which previously had just an Rv number, e.g. Rv0981 is identified now as *mpr* (Zahrt & Deretic, 2001).

Changes have also been generated from new sequences in the databases and from detailed studies of synteny, particularly with the related pathogen *M. leprae*. For example, Rv1860 was previously known as *apa*, encoding a 45–47 kDa secreted antigen (Laqueyrie *et al.*, 1995), located at the end of the *modABC* operon (molybdate transport system). Its name was recently changed to *modD*, on the basis of its proximity to *modABC*, and it is now part of the ModD family described by SWISS-PROT (e.g. P46842, Q50906). However, the protein which has been shown to be glycosylated and to have fibronectin-binding activity (Schorey *et al.*, 1995) shares no significant sequence similarity with other proteins involved in molybdate uptake. Furthermore, all of the functions involved in molybdate transport, and the enzymes that synthesize or require molybdopterin for activity, have been inactivated or lost from *M. leprae* (Eiglmeier *et al.*, 2001) with the exception of *modD*. This strongly suggests that ModD does not participate in molybdate uptake and we propose, therefore, that the name *apa* should be maintained.

Functional distribution of predicted CDS in 1998 and 2002

The new genomic annotation of *M. tuberculosis* H37Rv has incorporated many changes to the functional classifications of the predicted proteins. A comparison of the number of predicted proteins in each of the functional

categories between 1998 and 2002 is shown in Table 2. An important change has been the decrease in the number of unknown proteins from 606 to 272. Presently we are able to predict a function for 2058 proteins (52 % of the proteome) and more than 150 of these have been experimentally proven in mycobacterial research. The number of conserved hypothetical proteins has changed from 910 in 1998 to 1051 today. 376 putative proteins show no similarity to known proteins from other organisms and some of them may be specific to *M. tuberculosis*. To date, more than 400 *M. tuberculosis* proteins have been detected experimentally, most of them by proteomic studies (Weldingh *et al.*, 1998; Jungblut *et al.*, 1999; Mollenkopf *et al.*, 1999; Rosenkrands *et al.*, 2000b; Betts *et al.*, 2000). In the coming years, the number of unknown proteins should continue to decrease as more similarities are found by database searches or as functions are identified for some of these potentially *M. tuberculosis*-specific proteins. The structural genomics programmes currently under way on mycobacteria should have a significant impact in this respect (<http://www.doe-mbi.ucla.edu/TB/>; <http://www.pasteur.fr/recherche/X-TB/>).

ACKNOWLEDGEMENTS

We thank Q. T. Huynh, B. Caudron, L. Jones and N. Joly for their generous assistance with informatics, and J. Parkhill and K. D. James for help with sequence analysis. Special thanks to N. Stoker, P. Jungblut, I. Rosenkrands, A. Wietzorrek, A. Marcel, L. Frangeul, T. Garnier and T. Stinear, and the members of the mycobacterial research community who provided helpful comments. Financial support for this work was from the Wellcome Trust and the Génopole programme.

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403–410.
- Bentley, S. D., Chater, K. F., Cerdano-Tarraga, A. M. & 40 other authors (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147.
- Betts, J. C., Dodson, P., Quan, S., Lewis, A. P., Thomas, P. J., Duncan, K. & McAdam, R. A. (2000). Comparison of the proteome of the *Mycobacterium tuberculosis* strain H37Rv with clinical isolate CDC1551. *Microbiology* **146**, 3205–3216.
- Bocs, S., Danchin, A. & Medigue, C. (2002). Re-annotation of genome microbial CoDing-Sequences: finding new genes and inaccurately annotated genes. *BMC Bioinformatics* **3**, 1–5.
- Braibant, M., Gilot, P. & Content, J. (2000). The ATP binding cassette (ABC) transport systems of *Mycobacterium tuberculosis*. *FEMS Microbiol Rev* **24**, 449–467.
- Cole, S. T., Brosch, R., Parkhill, J. & 39 other authors (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544.
- Cole, S. T., Eiglmeier, K., Parkhill, J. & 40 other authors (2001). Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011.
- Dandekar, T., Huynen, M., Regula, J. T. & 10 other authors (2000). Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. *Nucleic Acids Res* **28**, 3278–3288.
- Eiglmeier, K., Parkhill, J., Honore, N., & 12 other authors (2001). The decaying genome of *Mycobacterium leprae*. *Lepr Rev* **72**, 387–398.
- Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K. & Bairoch, A. (2002). The PROSITE database, its status in 2002. *Nucleic Acids Res* **30**, 235–238.
- Fleischmann, R. D., Adams, M. D., White, O. & 37 other authors (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.
- Gaasterland, T. & Oprea, M. (2001). Whole-genome analysis: annotations and updates. *Curr Opin Struct Biol* **11**, 377–381.
- Gassel, M., Mollenkamp, T., Puppe, W. & Altendorf, K. (1999). The KdpF subunit is part of the K(+) -translocating Kdp complex of *Escherichia coli* and is responsible for stabilization of the complex *in vitro*. *J Biol Chem* **274**, 37901–37907.
- Glickman, M. S., Cox, J. S. & Jacobs, W. R., Jr (2000). A novel mycolic acid cyclopropane synthetase is required for cording, persistence, and virulence of *Mycobacterium tuberculosis*. *Mol Cell* **5**, 717–727.
- Jungblut, P. R., Schaible, U. E., Mollenkopf, H. J. & 7 other authors (1999). Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: towards functional genomics of microbial pathogens. *Mol Microbiol* **33**, 1103–1117.
- Jungblut, P. R., Muller, E. C., Mattow, J. & Kaufmann, S. H. (2001). Proteomics reveals open reading frames in *Mycobacterium tuberculosis* H37Rv not predicted by genomics. *Infect Immun* **69**, 5905–5907.
- Kersten, M. A., Muller, Y., Baars, J. J., Op den Camp, H. J., van der Drift, C., Van Griensven, L. J., Visser, J. & Schaap, P. J. (1999). NAD⁺-dependent glutamate dehydrogenase of the edible mushroom *Agaricus bisporus*: biochemical and molecular characterization. *Mol Gen Genet* **261**, 452–462.
- Laqueyrie, A., Militzer, P., Romain, F., Eiglmeier, K., Cole, S. T. & Marchal, G. (1995). Cloning, sequencing, and expression of the *apa* gene coding for the *Mycobacterium tuberculosis* 45/47-kilodalton secreted antigen complex. *Infect Immun* **63**, 4003–4010.
- Lu, C. D. & Abdelal, A. T. (2001). The *gdhB* gene of *Pseudomonas aeruginosa* encodes an arginine-inducible NAD(+) -dependent glutamate dehydrogenase which is subject to allosteric regulation. *J Bacteriol* **183**, 490–499.
- Minambres, B., Olivera, E. R., Jensen, R. A. & Luengo, J. M. (2000). A new class of glutamate dehydrogenases (GDH). Biochemical and genetic characterization of the first member, the AMP-requiring NAD-specific GDH of *Streptomyces clavuligerus*. *J Biol Chem* **275**, 39529–39542.
- Mollenkopf, H. J., Jungblut, P. R., Raupach, B., Mattow, J., Lamer, S., Zimny-Arndt, U., Schaible, U. E. & Kaufmann, S. H. (1999). A dynamic two-dimensional polyacrylamide gel electrophoresis database: the mycobacterial proteome via Internet. *Electrophoresis* **20**, 2172–2180.
- Moszer, I., Jones, L. M., Moreira, S., Fabry, C. & Danchin, A. (2002). SubtiList: the reference database for the *Bacillus subtilis* genome. *Nucleic Acids Res* **30**, 62–65.
- Mukamolova, G. V., Kaprelyants, A. S., Young, D. I., Young, M. & Kell, D. B. (1998). A bacterial cytokine. *Proc Natl Acad Sci USA* **95**, 8916–8921.
- Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* **10**, 1–6.
- Paulsen, I. T., Nguyen, L., Sliwinski, M. K., Rabus, R. & Saier,

- M. H., Jr (2000).** Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J Mol Biol* **301**, 75–100.
- Pearson, W. R. & Lipman, D. J. (1988).** Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85**, 2444–2448.
- Riley, M. (1993).** Functions of the gene products of *Escherichia coli*. *Microbiol Rev* **57**, 862–952.
- Rosenkrands, I., Weldingh, K., Jacobsen, S., Hansen, C. V., Florio, W., Gianetri, I. & Andersen, P. (2000a).** Mapping and identification of *Mycobacterium tuberculosis* proteins by two-dimensional gel electrophoresis, microsequencing and immunodetection. *Electrophoresis* **21**, 935–948.
- Rosenkrands, I., King, A., Weldingh, K., Moniatte, M., Moertz, E. & Andersen, P. (2000b).** Towards the proteome of *Mycobacterium tuberculosis*. *Electrophoresis* **21**, 3740–3756.
- Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. (2000).** Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944–945.
- Schorey, J. S., Li, Q., McCourt, D. W., Bong-Mastek, M., Clark-Curtiss, J. E., Ratliff, T. L. & Brown, E. J. (1995).** A *Mycobacterium leprae* gene encoding a fibronectin binding protein is used for efficient invasion of epithelial cells and Schwann cells. *Infect Immun* **63**, 2652–2657.
- Serres, M. H., Gopal, S., Nahum, L. A., Liang, P., Gaasterland, T. & Riley, M. (2001).** A functional update of the *Escherichia coli* K-12 genome. *Genome Biol* **2**, 0035.1–0035.7.
- Sonnhammer, E. L., von Heijne, G. & Krogh, A. (1998).** A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* **6**, 175–182.
- Tekaia, F., Gordon, S. V., Garnier, T., Brosch, R., Barrell, B. G. & Cole, S. T. (1999).** Analysis of the proteome of *Mycobacterium tuberculosis* in silico. *Tuber Lung Dis* **79**, 329–342.
- Weldingh, K., Rosenkrands, I., Jacobsen, S., Rasmussen, P. B., Elhay, M. J. & Andersen, P. (1998).** Two-dimensional electrophoresis for analysis of *Mycobacterium tuberculosis* culture filtrate and purification and characterization of six novel proteins. *Infect Immun* **66**, 3492–3500.
- Zahrt, T. C. & Deretic, V. (2001).** *Mycobacterium tuberculosis* signal transduction system required for persistent infections. *Proc Natl Acad Sci USA* **98**, 12706–12711.
-
- Received 2 April 2002; revised 27 June 2002; accepted 18 July 2002.