Genome update: prediction of membrane proteins in prokaryotic genomes

Genomes of the month

There have been six bacterial genomes published since the last 'Genome Update' column was written. All of this month's new genomes, listed in Table 1, come from genera for which there are multiple sequenced genomes already present in the databases. *Brucella abortus* and *Chlamydophila abortus* can both cause abortion in animals, as their names imply, although via different mechanisms. Two more *Staphylococcus* genomes have been reported (*Staphylococcus aureus* and *Staphylococcus epidermidis*), as well as genomes of *Salmonella enterica* SCB67 (Chiu *et al.*, 2005) and *Wolbachia* species TRS (Foster *et al.*, 2005). From a broader perspective, it should be noted that there are many genera (indeed many bacterial phyla) which have no representative genomes sequenced and it is hoped that at least some of the future genomes being sequenced will be more reflective of the biological diversity in the environment.

Brucellosis is a zoonotic infection transmitted from animals to humans by ingestion of infected food products, direct contact with an infected animal or inhalation of aerosols. Halling *et al.* (2005) report the complete genome sequence of *Brucella abortus* (strain 9-941, 3·3 Mb, two circular chromosomes and 3296 predicted genes) and compare it with *Brucella suis* strain 1330 and *Brucella melitensis* strain 16 M. The genomes are very similar with nearly identical gene content and organization. Further analysis identified a number of insertion–deletion events and several polymorphic regions. Several genes, previously described as unique to *B. suis* or *B. melitensis*, were also observed in the *B. abortus* genome, and overall the *B. abortus* genome has more sequences in common with *B. melitensis* than with *B. suis*.

Chlamydophila abortus (formerly within the *Chlamydia psittaci* taxon) is a cause of abortion and fetal loss in sheep, cattle and goats in many countries around the world. Infection with *C. abortus* has also been associated with abortion and other clinical symptoms in humans. The genome of *C. abortus* (strain S26/3, 1·14 Mb) was sequenced by Thomson *et al.* (2005). Compared to other *Chlamydiaceae* the genome shows a high level of conserved sequences and gene content. Out of 961 predicted coding sequences, 842 are conserved with those of *Chlamydophila caviae* and *Chlamydophila pneumoniae*. These different conserved parts of the *C. abortus* genome were identified as major regions of variation and all further analyses were based on these specific loci. Genes encoding highly variable protein families, such as TMH/Inc and polymorphic membrane protein (pmp) families were located. Antibodies raised against pmps significantly reduced the activity of elementary bodies, the infectious form of *Chlamydiaceae*. Although pmps constitute only a minority of the outer-membrane proteins, the identification of these proteins could be valuable in terms of understanding mechanisms of infection. Interestingly, *C. abortus* lacks any identified toxin genes, as well as genes involved in tryptophan metabolism and nucleotide salvaging.

Most infections caused by staphylococci are due to *Staphylococcus aureus*. Nevertheless, the incidence of infections due to *Staphylococcus epidermidis* and other coagulase-negative staphylococci has been steadily increasing in recent years. *S. aureus* is responsible for numerous hospital- and community-acquired infections, whereas infections with *S. epidermidis* are often associated with implanted medical devices. Gill *et al.* (2005) have sequenced the ~2·8 Mb genome of *S. aureus* (strain COL), an early methicillin-resistant (MRSA) isolate, and the genome of *S. epidermidis* (strain RP62a, ~2·6 Mb) and have conducted comparative analysis of these two species and other staphylococcal genomes to investigate their evolution and their resistance. *S. aureus* and *S. epidermidis* share a core set of 1681 ORFs. Their virulence and resistance attributes might be due to gene transfer between staphylococci and low-GC-content Gram-positive bacteria. Integrated plasmids in *S. epidermidis* containing genes encoding resistance to cadmium and species-specific LPXTG surface proteins, and a novel genome island, which can be a potential *S. epidermidis* virulence factor, were also identified, but a significant observation was the evidence for gene transfer between staphylococci and bacilli. The *cap* operon, a major virulence factor in *Bacillus anthracis*, has integrated the genomes of *S. epidermidis* strain RP62a and ATCC 12228, possibly via plasmid-mediated gene transfer.

Microbiology Comment provides a platform for readers of *Microbiology* to communicate their personal observations and opinions in a more informal way than through the submission of papers.

Most of us feel, from time to time, that other authors have not acknowledged the work of our own or other groups or have omitted to interpret important aspects of their own data. Perhaps we have observations that, although not sufficient to merit a full paper, add a further dimension to one published by others, or we may have a useful piece of methodology that we would like to share.

Guidelines on how to submit a *Microbiology* Comment article can be found in the Instructions for Authors at http://mic.sgmjournals.org

It should be noted that the Editors of *Microbiology* do not necessarily agree with the views expressed in *Microbiology* Comment.

Charles Dorman, Editor-in-Chief

Table 1. Summary of the published genomes discussed in this update

Note that the accession number for each chromosome is the same for GenBank, EMBL and DDBJ.

Name	Length	A + T (mol%)	No. of genes	tRNAs	rRNAs	Accession no.
Brucella abortus 9-941 chromosome 1	2 124 241	42·8	2030	41	2	AE017223
Brucella abortus 9-941 chromosome 2	1 162 204	42·7	1055	14	3	AE017224
Chlamydophila abortus S26-3 main	1 144 377	60·1	961	38	1	CR848038
Staphylococcus aureus COL main	2 809 422	67·2	2615	53	6	CP000046
Staphylococcus aureus COL pT181	4 440	70·0	3	0	0	CP000045
Staphylococcus epidermidis RP62A main	2 616 530	67·9	2494	61	6	CP000029
Staphylococcus epidermidis RP62A pSERP	27 310	68·1	32	0	0	CP000028
Salmonella enterica SCB67 main	4 755 700	47·8	4445	85	7	AE017320
Wolbachia sp. TRS main	1 080 084	65·8	805	34	1	AE017321

Method of the month – prediction of membrane proteins

All living cells are encapsulated by at least one membrane. In previous Genome Updates we have described prediction of the translocation machinery and various methods for the prediction of the proteins that are translocated over the cellular membrane. This month we will discuss methods for prediction of proteins that are embedded in the membrane. Membrane proteins have a significant number of roles in cellular metabolism and cell stability.

The majority of membrane proteins have a trans-membrane α-helix domain, but a minority have a β-barrel domain. Outer-membrane proteins of Gram-negative bacteria are often β-barrel proteins.

Most prediction methods have dealt with prediction of α-helical domains and their topology. We highly recommend TMHMM which is one of the most used prediction tools for identifying transmembrane proteins and their topology (Krogh *et al.*, 2001). Many other methods are available, such as HMMTOP (Tusnady & Simon, 2001) and MEMSAT (Jones *et al.*, 1994). The performance of transmembrane helix predictors has been reviewed (Möller *et al.*, 2001). For β-barrel membrane proteins, only two methods are available: BOMP (Berven *et al.*, 2004) and PRED-TMBB (Bagos *et al.*, 2004).

One would not expect the fraction of membrane proteins to differ significantly among different phyla. All prokaryotes need membrane transporters in order to take up metabolites and essential ions from the surroundings. As can be seen in Fig. 1, the fraction of membrane proteins in the proteome of different bacteria ranges from 15 to 20 % in the majority of phyla. It is worth mentioning that in some phyla the distribution range is quite large. For example, in *Actinobacteria*, the fraction of membrane proteins predicted using
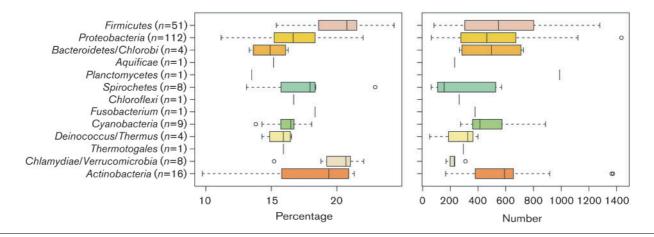

Fig. 1. Box and whisker plot of the number of predicted (TMHMM) membrane proteins in 13 different bacterial phyla. The colour scheme for the phyla is the same as found in the GenomeAtlas database. The box represents the middle 50 % of the data. The median of the data is shown by a vertical line. The 25th and 75th quartile is shown on the left and right side of the median, respectively. The whiskers cannot extend any further than 1·5 times the length of the quartiles. Outlier datapoints outside the whiskers are shown as open circles. One single vertical line is shown where only one proteome is present. For the plot on the left, the amount of predicted membrane proteins are normalized with the amount of annotated proteins of individual proteomes.

TMHMM ranges from less than 10 % in *Mycobacterium leprae* to 21 % in *Cornyebacterium glutamicum*. However, one should remember that the genome of *Mycobacterium leprae* contains many pseudogenes and hence the 'fraction of the total' in the case of this genome might have a different biological meaning than for other bacterial genomes containing few pseudogenes as part of the total gene count. The panel on the left shows the data plotted as a fraction of the predicted proteins containing transmembrane helices, normalized to the total number of proteins encoded in the genome. However, since some genomes are smaller than others, and we have shown previously that the mean genome length of some phyla can be different, we have also plotted the data in terms of the total number of transmembrane helices, without dividing by the total number of proteins in a given proteome. Again, in the example of the *Actinobacteria*, the range can be seen to differ in the plot on the right (unnormalized) compared to the fraction of the total, plotted on the left. In this case, the distribution is more evenly distributed, with a mean of around 600 proteins, although there are two outliers, with around 1370 predicted transmembrane proteins – these are the two *Streptomyces* genomes, which are quite large, containing about 7500 encoded proteins. In addition, the *Firmicutes*, with only one membrane, have a larger fraction of transmembrane proteins than the *Proteobacteria*, which have two membranes. One might expect the opposite trend here, although when the total number of proteins is examined in the right panel, the distributions seem to overlap more closely. It appears from this figure that most free-living bacteria contain roughly 400–500 transmembrane proteins.

Finally, the newly sequenced *Wolbachia* species (strain TRS) is unable to synthesize lipid A, the major component of bacterial membranes. Nevertheless, the predicted fraction of membrane proteins is 16 %, which implies that this endosymbiont still needs a large number of integral membrane proteins in order to live, even in this stable intracellular environment. As we saw in last month's Genome Update, the fraction of secreted proteins in endosymbionts is usually low compared to free-living prokaryotes.

Supplemental web pages

Additional web pages containing supplemental material related to this article can be accessed from www.cbs.dtu.dk/services/GenomeAtlas/suppl/GenUp017/

Jannick D. Bendtsen, Tim T. Binnewies, Peter F. Hallin and David W. Ussery

Center for Biological Sequence Analysis, BioCentrum-DTU, Building 208, The Technical University of Denmark, DK-2800 Kgs, Lyngby, Denmark

Correspondence: David W. Ussery (dave@cbs.dtu.dk)

Bagos, P. G., Liakopoulos, T. D., Spyropoulos, I. C. & Hamodrakas, S. J. (2004). PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins. *Nucleic Acids Res* **32**, W400–W404.

Berven, F. S., Flikka, K., Jensen, H. B. & Eidhammer, I. (2004). BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res* **32**, W394–W399.

Chiu, C. H., Tang, P., Chu, C., Hu, S., Bao, Q., Yu, J., Chou, Y. Y., Wang, H. S. & Lee, Y. S. (2005). The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. *Nucleic Acids Res* **33**, 1690–1698.

Foster, J., Ganatra, M., Kamal, I. & 23 other authors (2005). The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS Biol* **3**, e121.

Gill, S. R., Fouts, D. E., Archer, G. L. & 26 other authors (2005). Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. *J Bacteriol* **187**, 2426–2438.

Halling, S. M., Peterson-Burch, B. D., Bricker, B. J., Zuerner, R. L., Qing, Z., Li, L. L., Kapur, V., Alt, D. P. & Olsen, S. C. (2005). Completion of the genome sequence of *Brucella abortus* and comparison to the highly similar genomes of *Brucella melitensis* and *Brucella suis*. *J Bacteriol* **187**, 2715–2726.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33**, 3038–3049.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567–580.

Möller, S., Croning, M. D. & Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**, 646–653.

Thomson, N. R., Yeats, C., Bell, K. & 17 other authors (2005). The *Chlamydophila abortus* genome sequence reveals an array of variable proteins that contribute to interspecies variation. *Genome Res* **15**, 629–640.

Tusnady, G. E. & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17**, 849–850.

DOI 10.1099/mic.0.28181-0