

Genome update: prediction of secreted proteins in 225 bacterial proteomes

Genomes of the month

There have been three genomes published since the last 'Genome Update' column was written: *Bacteroides fragilis*, an opportunistic bacterial pathogen; *Cryptococcus neoformans*, an encapsulated yeast; and *Entamoeba histolytica*, an intestinal parasite.

The *Bacteroides fragilis* NCTC 9343 genome is the fifth member of the *Bacteroidetes/Chlorobi* group to be sequenced and the second *Bacteroides fragilis* strain. The *Bacteroides fragilis* NCTC 9343 genome shows considerable variation even within the same strain and 'invertible promoters' appear to regulate much of this variance in many surface-exposed and secreted proteins (Cerdeño-Tárraga *et al.*, 2005). Characteristics of the genome are given in Table 1. As discussed below, members of the *Bacteroidetes/Chlorobi* group appear to have a large fraction of proteins in their genome which are predicted to be secreted compared to other bacterial phyla.

Cryptococcus is an encapsulated yeast. The genus includes around 37 species, but *Cryptococcus neoformans* is the only species that is pathogenic in humans. This has made this organism not only a biological curiosity, but also a medical curiosity. The disease takes advantage of the increasing numbers of patients with immunosuppressive diseases, like AIDS, and the use of drugs that suppress the immune system supports its dissemination. Loftus *et al.* (2005b) sequenced and published the ~19 Mbp genome of two strains (JEC21 and B-3501A). Around 5% of this genome consists of transposons, which can cause karyotype instability and variation in the phenotype. Strain B-3501A is more thermotolerant in animal models than JEC21. Other remarkable features were identified, such as a special link between virulence and mating type, controlled by the MAT locus, identification of more than 30

new genes likely to be involved in capsule biosynthesis, and differences between *Saccharomyces cerevisiae* and *Cryptococcus neoformans* in their mechanisms of cell-wall protein association. A comparison between *Cryptococcus neoformans* and *Candida albicans* revealed that the organisms use different mechanisms to bind host cells.

Entamoeba histolytica, an intestinal parasite which causes amoebiasis, is the first fully sequenced amoeba genome (Loftus *et al.*, 2005a). This disease is most common in developing countries with poor sanitary conditions, but it has also been observed in industrial countries. The most dramatic incident in the USA was the Chicago World's Fair outbreak in 1933 caused by contaminated drinking water. Defective plumbing permitted sewage to contaminate the drinking water which ultimately led to 1000 cases with 58 deaths. The most striking results from the sequencing project is the amount of lateral transferred genes. *Entamoeba histolytica* has apparently found a way to increase its range of substrates for energy generation, simple by 'borrowing' genes from prokaryotes (mostly from the *Cytophaga/Flavobacterium/Bacteroides* group). It also shows an interesting arsenal of tyrosine and serine/threonine kinases, normally not present in protists. Having kinases from all eukaryotic protein kinase superfamilies results in a complex mix of signal transduction systems for interaction with different environments.

Method of the month – prediction of secreted proteins

As mentioned above, members of the *Bacteroidetes/Chlorobi* group appear to have a large fraction of proteins in their genome which are predicted to be secreted, as can be seen in Fig. 1. How do we know this, i.e. how are secreted proteins predicted? This month we look into prediction of secretory proteins as a natural follow up of the previous Genome Update where we looked at bacterial secretion systems. Interestingly, there is a correlation between the fraction of secreted proteins and the environment in which the bacteria live.

Here we have focused on the prediction of N-terminal Sec signal peptides, Tat signal peptides and lipoprotein signal peptides. For this purpose we have used SignalP (www.cbs.dtu.dk/services/SignalP) for prediction of Sec signal peptides (Bendtsen *et al.*, 2004). LipoP (www.cbs.dtu.dk/services/LipoP) was used for prediction of lipoprotein signal peptides (Juncker *et al.*, 2003). For proteins carrying a twin-arginine (Tat) signal peptide we have used an unpublished in-house tool. These methods for signal peptide prediction are partly overlapping and therefore we have developed a rule-based classification scheme which does not allow overlapping predictions from the different methods.

We have explicitly excluded analysis of

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*.

Chris Thomas, 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.
<i>Bacteroides fragilis</i> NCTC 9343 (main)	5 205 140	56.8	4260	73	6	CR626927
<i>Bacteroides fragilis</i> NCTC 9343 (plasmid)	36 560	67.8	48	0	0	CR626928
<i>Cryptococcus neoformans</i> JEC21	19 051 841	51.4	~ 6500	141	~ 5 %*	AE017341–AE017356
<i>Cryptococcus neoformans</i> B-3501A	~ 18 500 000	~ 51	~ 6500	?	~ 5 %*	AAEY00000000
<i>Entamoeba histolytica</i>	23 751 783	77.6	9938	?	~ 10 %*	AAFB01000001-1819

*Fraction of the genome.

proteins which are secreted via a non-classical secretory pathway, such as the ESAT-6 proteins of *Mycobacterium tuberculosis*. These proteins do not have any known sequence motif which would enable targeting to the extracellular environment, thus they are not easily predictable.

For more than 200 bacterial genomes, in 13 phyla, we have estimated the fraction of proteins carrying Sec signal peptides, lipoprotein signal peptides and Tat signal peptides (see Fig. 1). For example, proteins carrying classical Sec signal peptides are found in a range of 4–17 % for *Proteobacteria* and slightly less for *Firmicutes*; these two phyla account for 75 % of the genomes in our database. The fraction of lipoproteins found in *Proteobacteria* approximately ranges from 0 to 3 % whereas *Firmicutes* are predicted to secrete slightly more lipoproteins, ranging from 1 to 4 %, with some outliers (Fig. 1). Members of the phylum *Bacteroidetes* seem to secrete more lipoproteins than the other phyla, which could be correlated to the

living environment of these organisms. Attachment of *Bacteroides fragilis* to the host tissue may be the result of secretion of lipoproteins which are often adhesins. Degradation of the host extracellular matrix also involves a large fraction of secreted proteins.

Most phyla have a narrow range of proteins carrying classical Sec signal peptides, lipoprotein signal peptides and Tat signal peptides, respectively. This indicates a bias of protein secretion within phyla, which may be a result of the environmental conditions for the bacteria. Nevertheless, some outliers in the low range are seen for the *Proteobacteria* (Fig. 1). This shows that some endosymbiotic bacteria (intracellular in eukaryotic cells) secrete far less proteins to the extracellular environment than other bacteria. ‘*Candidatus Blochmannia floridanus*’, *Buchnera aphidicola* and ‘*Wigglesworthia brevipalpis*’ are examples of intracellular bacteria which only secrete approximately 2 % of their entire proteome. Moreover, ‘*Candidatus Blochmannia*

floridanus’ has only part of the Type I and II secretion systems.

For another proteobacterium, *Bdellovibrio bacteriovorus*, we predict that 23 % of the proteins carry Sec signal peptides and 5 % to be lipoproteins. *Bdellovibrio bacteriovorus* is a predatory bacterium that preys on other Gram-negative bacteria, so it makes sense that it should encode many secretory proteins for degradation and uptake of foreign proteins (Rendulic *et al.*, 2004).

In general, all investigated phyla have far fewer proteins carrying a Tat signal peptide (Berks, 1996) than Sec signal peptides or lipoprotein signal peptides, although a few outliers are found. *Bradyrhizobium japonicum* has 116 predicted Tat signal peptides. *Bradyrhizobium japonicum* is a nitrogen-fixing bacterium which has a large genome carrying 8317 genes. Nitrogen fixation involves co-factor-binding proteins which are known to be secreted via the Tat secretory pathway. Two outliers are found

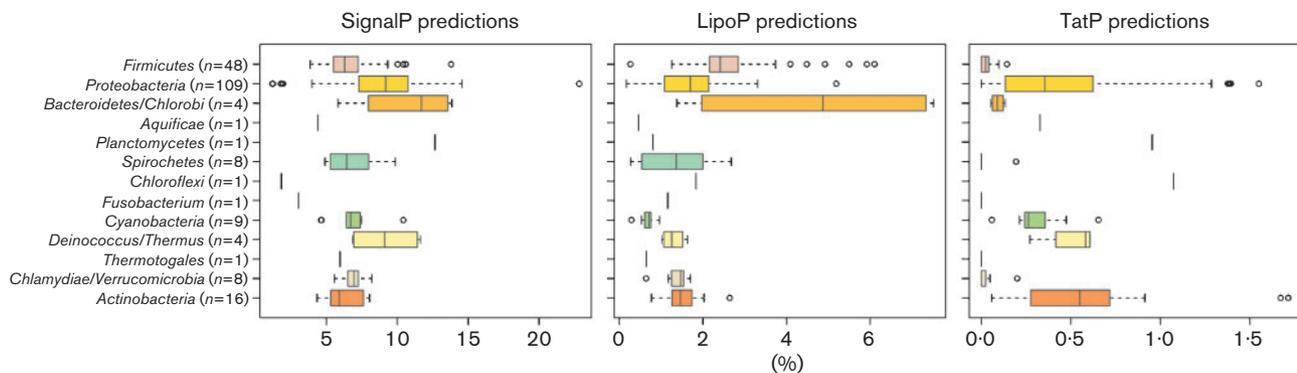


Fig. 1. Proteome fractions of proteins carrying Sec signal peptides, Tat signal peptides or lipoproteins in 13 different phyla. One single vertical line is shown where only one proteome is present. Outliers are shown by open circles.

for the phylum *Actinobacteria*. *Streptomyces avermitilis* and *Streptomyces coelicolor* have around 130 predicted Tat substrates. As for *Bradyrhizobium japonicum*, the two *Streptomyces* genomes are far larger than ordinary actinobacterial genomes, having more than 7500 genes. In summary, we find that the fraction of secretory proteins is to some extent correlated to the environment where the bacteria live.

Supplemental web pages

Additional web pages containing supplemental material related to this article can be accessed from www.cbs.dtu.dk/services/GenomeAtlas/suppl/GenUp016/

Acknowledgements

This work was supported by a grant from the Danish Center for Scientific Computing.

Jannick D. Bendtsen, Tim T. Binnewies, Peter F. Hallin, Thomas Sicheritz-Pontén 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)

Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783–795.

Berks, B. C. (1996). A common export pathway for proteins binding complex redox cofactors? *Mol Microbiol* **22**, 393–404.

Cerdeño-Tárraga, A. M., Patrick, P., Crossman, L. C. & 22 other authors (2005). Extensive

DNA inversions in the *B. fragilis* genome control variable gene expression. *Science* **307**, 1463–1465.

Juncker, A. S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H. & Krogh, A. (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci* **12**, 1652–1662.

Loftus, B., Anderson, I., Davies, R. & 51 other authors (2005a). The genome of the protist parasite *Entamoeba histolytica*. *Nature* **433**, 865–868.

Loftus, B. J., Fung, E., Roncaglia, P. & 51 other authors (2005b). The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. *Science* **307**, 1321–1324.

Rendulic, S., Jagtap, P., Rosinus, A. & 10 other authors (2004). A predator unmasked: life cycle of *Bdellovibrio bacteriovorus* from a genomic perspective. *Science* **303**, 689–692.

DOI 10.1099/mic.0.28029-0