

### Genome Update: 2D clustering of bacterial genomes

#### Genomes of the month

Four new microbial genomes have been published since last month's Genome Update was written: that of an environmental bacterium which degrades aromatic compounds (*Azoarcus* sp. strain EbN1), that of a spirochaete that causes Lyme disease (*Borrelia garinii*) and those of two more *Streptococcus* species. A brief overview of each of these genomes is given below.

*Azoarcus* sp. strain EbN1 can metabolize many aromatic compounds, including hydrocarbons (Rabus *et al.*, 2005). Although other organisms such as *Pseudomonas putida* KT2440 can also catabolize aromatics (Jimenez *et al.*, 2002), *Azoarcus* sp. strain EbN1 can catabolize aromatics under both aerobic and anaerobic conditions. The main chromosome is ~4.3 Mbp long and is GC-rich (see Table 1). There are also two large plasmids, each around 200 kbp in length. Genes encoding more than 150 proteins involved in aromatic degradation were found, in keeping with this organism's known biochemical lifestyle.

Infections with *B. garinii* are associated with neurological symptoms and, in Europe, it is one of the major contributors to reported cases of Lyme disease, also referred to as Lyme borreliosis, which is a multi-system disorder. In the United States, *Borrelia burgdorferi sensu stricto* is the only known causative agent of this disease. Both bacteria live in the gastro-intestinal tract of blood-feeding ticks that may infect their hosts by biting them. *B. garinii* strain PBI has been sequenced (Glockner *et al.*, 2004). The main chromosome is ~904 kbp long, and two plasmids (lp54, A and cp26, B) of ~56 kbp and ~27 kbp, respectively, were also sequenced, as shown in Table 1. Both plasmids and the chromosome are collinear to the chromosome, the linear

plasmid lp54 and the circular plasmid cp26 of *B. burgdorferi* strain B31.

*Streptococcus thermophilus* is a lactic acid bacterium that is widely used for the production of yoghurt and cheese. Although the genus *Streptococcus* includes several pathogenic species such as *Streptococcus pyogenes* and *Streptococcus pneumoniae*, the dairy bacterium *S. thermophilus* has apparently lost its pathogenicity somewhere in its evolutionary history. To get a better understanding of this evolutionary path and to assess the potential for virulence, two genomes of *S. thermophilus* (of strains CNRZ 1066 and LMG 13811) were isolated, sequenced (Bolotin *et al.*, 2004) and compared with previously sequenced pathogenic streptococci (*S. pneumoniae*, *S. pyogenes*, *Streptococcus agalactiae* and *Streptococcus mutans*).

*S. thermophilus* strains CNRZ 1066 and LMG 13811 both contain a circular chromosome of ~1.8 Mbp and about 1900 coding sequences (1 796 226 bp/1915 genes and 1 796 846 bp/1889 genes, respectively) with an AT content of 60.9% (Bolotin *et al.*, 2004). The two strains are involved in the same dairy process so therefore it is not a surprise that they have greater than 90% of coding sequences in common. Of these coding sequences, nearly 1500 (80%) genes are orthologous to other streptococcal genes.

This indicates that *S. thermophilus* and its pathogenic relatives still share a large part of their general physiology and metabolism. Many streptococcal virulence-related genes are absent from *S. thermophilus* or only present as pseudogenes, except when they code for proteins that are involved in basic cellular functions. Adaptation to the constant dairy environment appears to support the stabilization of the genome structure. *S. thermophilus* is the first organism for which regressive evolution has been observed in a food niche rather than in a pathogen-host situation.

#### Method of the month – 2D clustering of genomic properties

This month we present an old method as a new tool for comparison of genomes. Clustering is a method for visualizing trends in multidimensional data and it is an unsupervised learning algorithm where class labels are not assigned beforehand. By cluster analysis, meaningful patterns can be discovered and the method may thus be used for discovery of new and unforeseen classes and/or confirmation of known classification. In phylogeny, clustering is mostly used for 16S rRNA alignments to create phylogenetic trees. Here, we demonstrate how it may also be used to cluster organisms that share a similar pattern of various genomic traits. These traits may be anything from genome size to G+C content. Depending

**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 the DDBJ.

Name	Length	AT content (%)	No. of genes	tRNAs	rRNAs	Accession no.
<i>Azoarcus</i> species EbN1 main chromosome	4 296 230	34.9	4133	58	4	CR555306
<i>Azoarcus</i> species EbN1 plasmid 1	207 355	42.4	274	0	0	CR555307
<i>Azoarcus</i> species EbN1 plasmid 2	223 670	36.9	196	0	0	CR555308
<i>Borrelia garinii</i> PBi main chromosome	904 246	71.7	832	33	1	CP000013
<i>Borrelia garinii</i> PBi plasmid lp54	55 560	73.5	74	0	0	CP000015
<i>Borrelia garinii</i> PBi plasmid cp26	27 108	74.4	26	0	0	CP000014
<i>Streptococcus thermophilus</i> LMG 18311	1 796 846	60.9	1889	67	6	CP000023
<i>Streptococcus thermophilus</i> CNRZ 1066	1 796 226	60.9	1915	67	6	CP000024

on which properties one tries to cluster by, clusters may be similar to the phylogenetic classes or they may be more related to, for example, the environment in which the organisms live.

Here we use a hierarchical clustering scheme, which is a deterministic method that represents all pairwise distances between organisms or feature traits in a dendrogram. This way, the distances between organisms plotted in an  $n$ -dimensional space ( $n$  being the number of features) is determined and organisms that are closest to each other in this  $n$ -dimensional space will subsequently cluster most closely together in the derived 2D dendrogram representation of the organisms. The distances were determined using Pearson correlation distances to capture specific trends or patterns in the data. Also, we use average-linkage, where the distance between any two clusters is considered equal to the average distance from any member of one cluster to any member of the other cluster.

Perhaps the best way to illustrate this is to have a look at some examples. Fig. 1 displays two cluster plots of various genomic properties for the two newly

sequenced *S. thermophilus* strains (as discussed above, used in the yoghurt and cheese industries) as well as the other sequenced *Streptococcus* species and sequenced members of the 'dairy bacteria'. Values for each feature trait were centred and scaled so that they could be represented by the same colour scale. Organisms are clustered vertically and the feature traits are clustered horizontally, hence the term 2D clustering. In the middle coloured part of the plots, each coloured square represents the scaled value of the feature trait written directly below and the organism written directly to the right. In this way, more-intense green colours correspond to extremely low values for the given feature trait as compared to the values for the other included organisms and *vice versa* for the more-intense blue colours.

By clustering according to Pearson correlation distances, organisms with the same trend are clustered together instead of clustering organisms with similar absolute values of these traits (Euclidian distances). In both cluster plots, it is seen that organisms of the same genus and species tend to cluster together and the difference between the two newly

sequenced *S. thermophilus* strains is negligible and they cluster very tightly, as expected.

In Fig. 1(a), we illustrate the clustering of some traits that are very often reported for newly sequenced genomes. These are AT content, number of genes, coding fraction, average number of base pairs per gene, number of tRNAs and number of 5S–16S–23S rRNA operons. We have also included some base-skew information: AT skew and GC skew. With regard to these traits, the two *S. thermophilus* strains appear to be most correlated with another *Streptococcus* species, *S. agalactiae*, which is a commensal bacterium that colonizes the intestinal tract. From the cluster plot, it is evident that the absolute values for these two species are entirely different, but they share the same trend of, for example, having a relatively higher average number of base pairs per gene compared to the AT content.

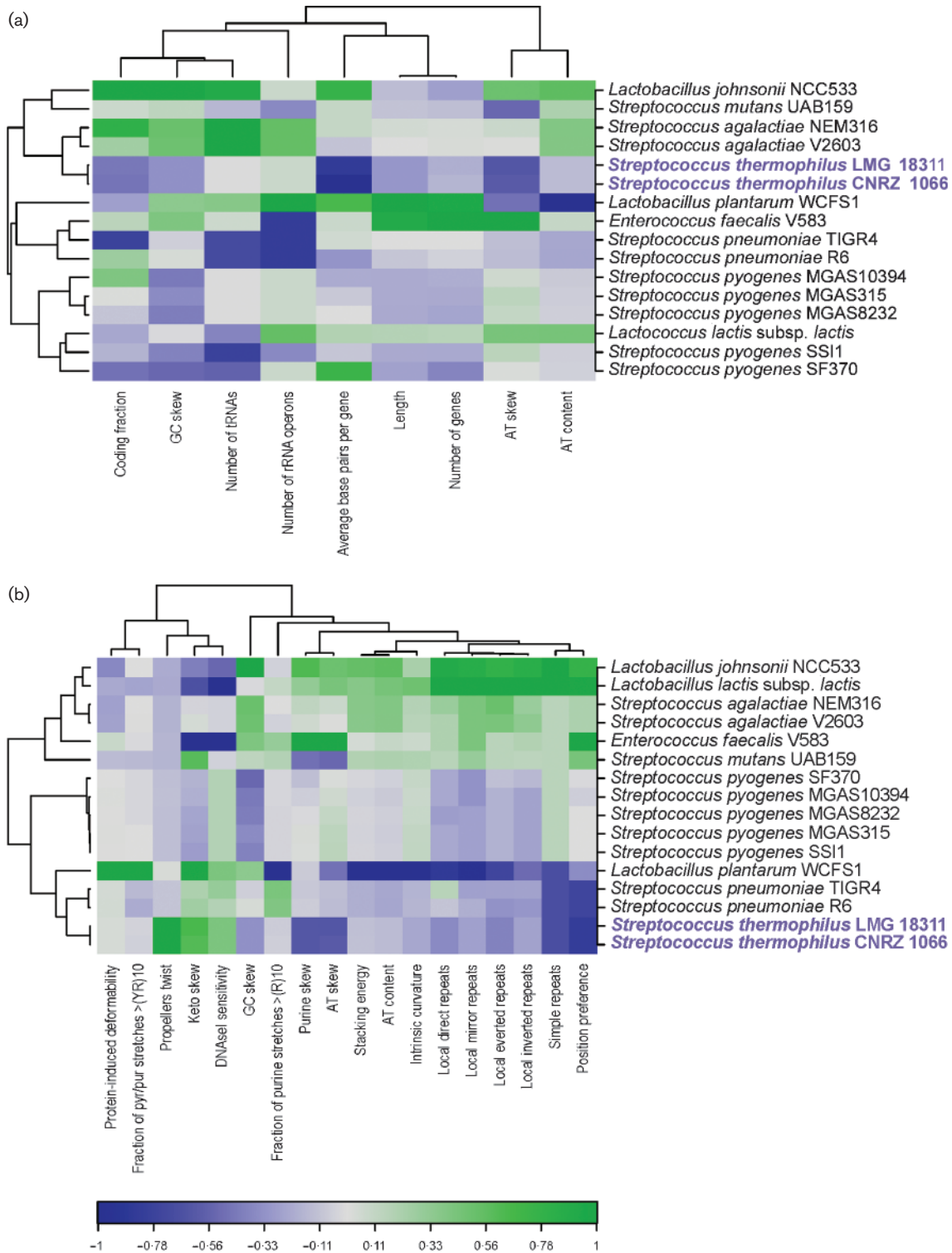
In Fig. 1(b), we illustrate the clustering of some structural parameters. Intrinsic curvature (Shpigelman *et al.*, 1993), propeller twist (el Hassan & Calladine, 1996), DNA base stacking energy (Ornstein *et al.*, 1978), position preference

**Fig. 1.** Hierarchical cluster analysis of some *Streptococcus* and dairy bacteria. (a) General properties of the genomes: coding fraction, number of tRNAs, average number of base pairs per gene, number of rRNA operons, AT skew, AT content, GC skew, genome length and number of annotated genes. (b) Various genomic properties affecting the DNA structure (horizontal): intrinsic curvature (Shpigelman *et al.*, 1993), propellers twist (el Hassan & Calladine, 1996), stacking energy (Ornstein *et al.*, 1978), position preference (Satchwell *et al.*, 1986), protein-induced deformability (Olson *et al.*, 1998) and DNase I sensitivity (Brukner *et al.*, 1995). Simple repeats, local direct repeats, local everted repeats, local inverted repeats and local mirror repeats (Jensen *et al.*, 1999). Fraction of purine stretches, fraction of pyr/pur stretches and purine skew (Ussery *et al.*, 2002). AT content, GC skew, AT skew and keto skew. For both plots, the values in each column (for each feature) were centred by subtracting the column mean and scaled by dividing the resulting values with the maximum absolute values. This way values in each column are between  $-1$  and  $1$ , with at least one of these extreme values present as a dark-blue or dark-green spot.

(Satchwell *et al.*, 1986), protein-induced deformability (Olson *et al.*, 1998) and DNase I sensitivity (Brukner *et al.*, 1995) are all direct measures of DNA curvature, flexibility and bendability. Simple repeats, local direct repeats, local everted repeats, local mirror repeats and local inverted

repeats also influence DNA secondary structure (Jensen *et al.*, 1999). Moreover, the fraction of purine stretches, fraction of pyr/pur stretches and the purine skew may also influence DNA topology (Ussery *et al.*, 2002). Finally, AT content, GC skew, AT skew and keto skew were included.

For *S. thermophilus*, these structural features were mostly correlated with those for strains of *S. pneumoniae*, a pathogen of the same genus although the R6 strain is avirulent (Hoskins *et al.*, 2001). From the cluster plot, it is evident that especially the average propeller twist angle was much



higher than for the other organisms included in the plot. On the other hand, position preference was very low for both *S. thermophilus* and *S. pneumoniae*.

From Fig. 1, it is also possible to see how traits are related. As expected, length and number of genes form a very tight cluster in Fig. 1(a). In Fig. 1(b), it can be seen that AT content and stacking energy are very much correlated. This is consistent with the fact that AT-rich regions tend to destack more readily and thus have less negative stacking energy (Ornstein *et al.*, 1978).

The statistical software package R was used to construct the cluster plots. A fully automated version of this package has been implemented as a web-based service in the Genome Atlas Database (<http://www.cbs.dtu.dk/services/GenomeAtlas/>). The user can make a 'cluster within search' for selected organisms and selected traits from the database.

#### Supplemental web pages

Web pages containing material related to this article can be accessed from the following url: <http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/GenUp013/>

#### Acknowledgements

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

**Hanni Willenbrock, Tim T. Binnewies, Peter F. Hallin and David W. Ussery**

Center for Biological Sequence Analysis, Department of Biotechnology, Building 208, The Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

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

**Bolotin, A., Quinquis, B., Renault, P. & 20 other authors (2004).** Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*. *Nat Biotechnol* **22**, 1554–1558.

**Brukner, I., Sanchez, R., Suck, D. & Pongor, S. (1995).** Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J* **14**, 1812–1818.

**el Hassan, M. A. & Calladine, C. R. (1996).** Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J Mol Biol* **259**, 95–103.

**Glockner, G., Lehmann, R., Romualdi, A., Pradella, S., Schulte-Spechtel, U., Schilhabel, M., Wilske, B., Suhnel, J. & Platzer, M. (2004).** Comparative analysis of the *Borrelia garinii* genome. *Nucleic Acids Res* **32**, 6038–6046.

**Hoskins, J., Alborn, W. E., Jr, Arnold, J. & 39 other authors (2001).** Genome of the bacterium *Streptococcus*

*pneumoniae* strain R6. *J Bacteriol* **183**, 5709–5717.

**Jensen, L. J., Friis, C. & Ussery, D. W. (1999).** Three views of microbial genomes. *Res Microbiol* **150**, 773–777.

**Jimenez, J. I., Minambres, B., Garcia, J. L. & Diaz, E. (2002).** Genomic analysis of the aromatic catabolic pathways from *Pseudomonas putida* KT2440. *Environ Microbiol* **4**, 824–841.

**Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M. & Zhurkin, V. B. (1998).** DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc Natl Acad Sci U S A* **95**, 11163–11168.

**Ornstein, R., Rein, R., Breen, D. & MacElroy, R. (1978).** An optimized potential function for the calculation of nucleic acid interaction energies. *Biopolymers* **17**, 2341–2360.

**Rabus, R., Kube, M., Heider, J., Beck, A., Heitmann, K., Widdel, F. & Reinhardt, R. (2005).** The genome sequence of an anaerobic aromatic-degrading denitrifying bacterium, strain EbN1. *Arch Microbiol* **183**, 27–36.

**Satchwell, S. C., Drew, H. R. & Travers, A. A. (1986).** Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* **191**, 659–675.

**Shpigelman, E. S., Trifonov, E. N. & Bolshoy, A. (1993).** CURVATURE: software for the analysis of curved DNA. *Comput Appl Biosci* **9**, 435–440.

**Ussery, D., Soumpasis, D. M., Brunak, S., Stærfeldt, H. H., Worning, P. & Krogh, A. (2002).** Bias of purine stretches in sequenced chromosomes. *Comput Chem* **26**, 531–541.

DOI 10.1099/mic.0.27811-0