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 (Azotobacter 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.

Azotobacter 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), Azotobacter 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, 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 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 wherever 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 (1796 226 bp/1915 genes and 1796 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 GC content. Depending
Microbiology Comment

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.

<table>
<thead>
<tr>
<th>Name</th>
<th>Length</th>
<th>AT content (%)</th>
<th>No. of genes</th>
<th>tRNAs</th>
<th>rRNAs</th>
<th>Accession no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Azoarcus species EbN1 main chromosome</td>
<td>4 296 230</td>
<td>34-9</td>
<td>4133</td>
<td>58</td>
<td>4</td>
<td>CR555306</td>
</tr>
<tr>
<td>Azoarcus species EbN1 plasmid 1</td>
<td>207 355</td>
<td>42-4</td>
<td>274</td>
<td>0</td>
<td>0</td>
<td>CR555307</td>
</tr>
<tr>
<td>Azoarcus species EbN1 plasmid 2</td>
<td>223 670</td>
<td>36-9</td>
<td>196</td>
<td>0</td>
<td>0</td>
<td>CR555308</td>
</tr>
<tr>
<td>Borrelia garinii PBi main chromosome</td>
<td>904 246</td>
<td>71-7</td>
<td>832</td>
<td>33</td>
<td>1</td>
<td>CP000013</td>
</tr>
<tr>
<td>Borrelia garinii PBi plasmid lp54</td>
<td>55 560</td>
<td>73-5</td>
<td>74</td>
<td>0</td>
<td>0</td>
<td>CP000015</td>
</tr>
<tr>
<td>Borrelia garinii PBi plasmid cp26</td>
<td>27 108</td>
<td>74-4</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>CP000014</td>
</tr>
<tr>
<td>Streptococcus thermophilus LMG 18311</td>
<td>1 796 846</td>
<td>60-9</td>
<td>1 889</td>
<td>67</td>
<td>6</td>
<td>CP000023</td>
</tr>
<tr>
<td>Streptococcus thermophilus CNRZ 1066</td>
<td>1 796 226</td>
<td>60-9</td>
<td>1 915</td>
<td>67</td>
<td>6</td>
<td>CP000024</td>
</tr>
</tbody>
</table>

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 organis 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 mostly 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 (Shigelman et al., 1993), propeller twist (el Hassan & Calladine, 1996), DNA base 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) 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)


DOI 10.1099/mic.0.27811-0