

Three views of microbial genomes

Lars Juhl Jensen, Carsten Friis, David W. Ussery*

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

Abstract — We describe here GenomeAtlases as a method for visualising three different aspects of complete microbial chromosomes: repeats, DNA structural characteristics, and base composition. We have applied this method to all publicly available genomes, and find a general strand preference of global repeats. The atlas for the *Mycoplasma genitalium* genome is presented as an example, and results from all three views are consistent with known characteristics of the genome. © 1999 Éditions scientifiques et médicales Elsevier SAS

complete genome / repetitive DNA / DNA structure / base composition / *Mycoplasma genitalium*

1. Introduction

It is inherently difficult to get an overview of a complete chromosome simply because of its size. We present here a method for visualising complete microbial chromosomes so that repetitive sequences and anomalies in base composition or DNA structure become visible. A number of parameters are calculated for the DNA double helix based on the nucleotide sequence. These parameters belong to three categories: repeats, structural parameters, and parameters directly related to the base composition. An atlas in which these parameters are visualised as coloured circles is made for each of these three categories; in addition a combined atlas summarising the most informative parameters is constructed (see *figure 1*).

2. Materials and methods

2.1. Repeats and symmetry elements

Repeats are multiple copies of the same sequence at different locations on a piece of DNA.

We divide repeats into three major categories: simple repeats, symmetry elements (also termed local repeats), and global repeats. All the different types of repeats are found using variations of the same basic algorithm. This algorithm finds the highest degree of homology for an R bp repeat within a window of length W .

2.1.1. Simple repeats

A simple repeat is a region consisting only of a repeated oligonucleotide, and can therefore be thought of as an extension to microsatellites. Simple repeats are found by looking for local repeats of length R within a $2R$ -bp window. By using the values 12, 14, 15, 16, and 18 for R , all simple repeats of lengths 1 through 9 covering at least 24 bp are detected.

2.1.2. The four types of symmetry elements

There are four kinds of symmetry that a repeated sequence can possess, since two copies of a sequence can be on either the same or opposite strands and independently of this pointing in the same or opposite directions (*table 1*).

Table 1. The four types of symmetry elements.

	Same direction	Opposite directions
Same strand	Direct repeat	Mirror repeat
Opposite strands	Everted repeat	Inverted repeat

* Correspondence and reprints
Tel.: +45 45 25 24 88; fax: +45 45 93 15 85
dave@clos.dtu.dk

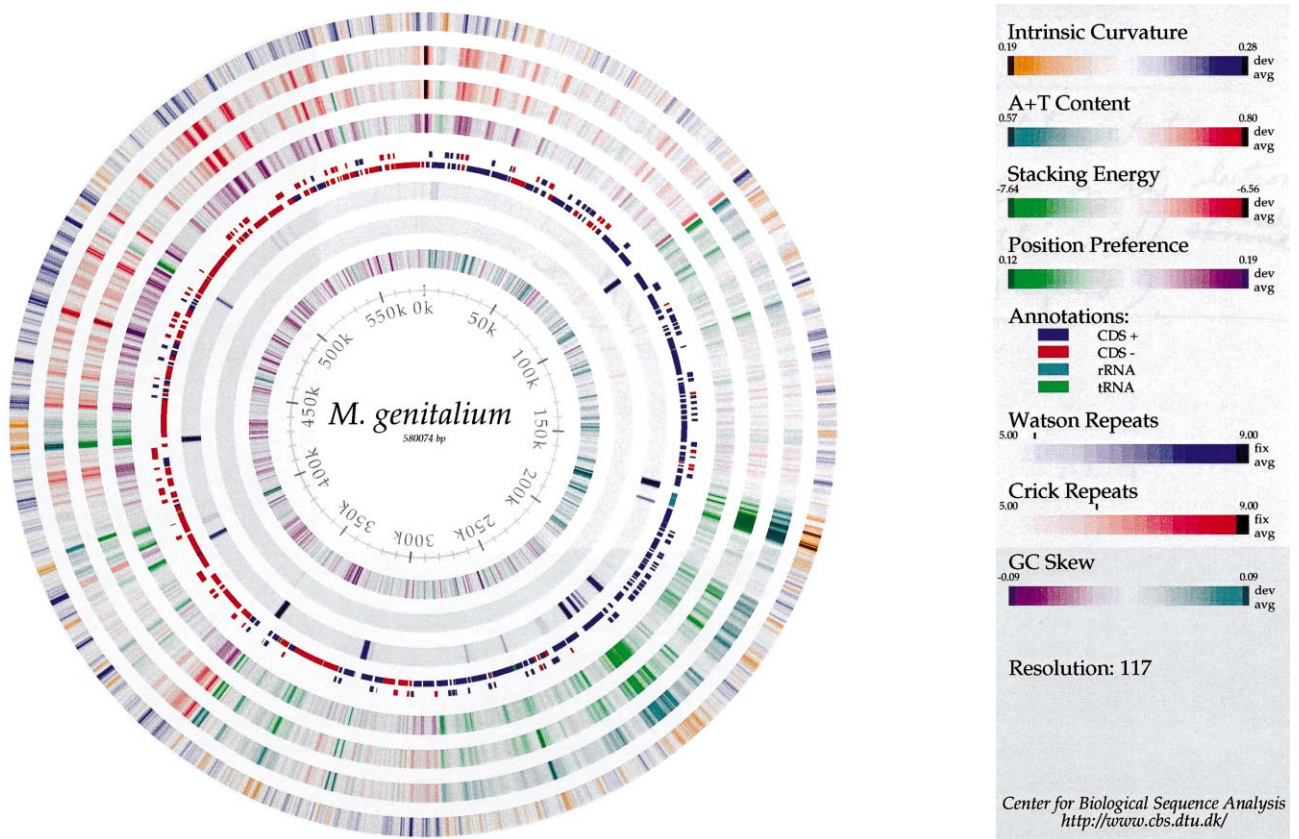


Figure 1. The GenomeAtlas summarising repeats, structural properties, and base composition of the complete *M. genitalium* genome. Similar atlases are available for all sequenced microbial genomes.

Local repeats are occurrences of symmetry elements within a small region (on the order of 100 bp). All four kinds of local repeats can be involved in the formation of special DNA structures – the best known of which is the cruciform which can be formed by inverted repeats.

Mirror repeats can form a very different structure: intramolecular triple-stranded DNA, also known as triplex DNA and H-DNA [8]. In addition to a triple-stranded part, triplex DNA also consists of a single-stranded part, which can hybridise to another piece of DNA; this is thought to be a possible mechanism for homologous recombination [21].

Special DNA structures associated with direct repeats and everted repeats exist as well. Direct repeats can form so-called slipped-strand DNA (S-DNA), which can cause frame shift muta-

tions. Everted repeats can form parallel-stranded DNA structures, in which the normal antiparallel 5′–3′ and 3′–5′ base pairing between the strands is substituted by 5′–3′ and 5′–3′ base pairing [18].

2.1.3. Global repeats

On a global scale, a direct repeat is a sequence that is present in at least two copies on the same strand, whereas two copies located on opposite strands will give rise to an inverted repeat. To avoid confusion with local repeats, we use the terms Watson and Crick repeats for global direct and global inverted repeats respectively.

Both duplicated and homologous genes will give rise to either Watson repeats or Crick repeats depending on the relative orientation of the genes – possible sites for homologous recom-

bination can therefore be identified by searching for global repeats. Other sources of global repeats include insertion elements and the repetitive sequences dispersed in most eukaryotic genomes.

2.2 Structural parameters

A number of measures for the local structure of DNA have been devised, most of which are based on simple lookup tables of either dinucleotide or trinucleotide values that have been obtained by fitting either experimental results or theoretical estimates.

2.2.1. *Intrinsic curvature*

Intrinsic curvature is a property of DNA that is closely related to anomalous gel mobility, as DNA fragments with high intrinsic curvature will migrate more slowly on polyacrylamide gels than markers with the same length. In this work we have used the CURVATURE programme [17], which is based on a wedge model [19, 20], for prediction of intrinsic curvature. From a set of dinucleotide values for the twist, wedge, and direction angles the three-dimensional path of a 21-bp fragment is calculated. Curvature profiles for longer sequences can thus be calculated using a 21-bp running window. Other theoretical models for DNA curvature exist – however, these give very similar predictions [7].

2.2.2. *Parameters for helix rigidity*

Three of the structural parameters that we have examined are related to helix rigidity and stability. As should be expected, these are strongly correlated with each other [4].

One, the stacking energy, is a measure of the interaction energy between adjacent basepairs in the DNA double helix. The total stacking energy of a DNA segment can be estimated from the set of dinucleotide values determined by quantum mechanical calculations on crystal structures [14]. All stacking energies are negative since base stacking is an energetically favourable interaction that serves to stabilise the double helix. This means that regions with large stacking energies are strongly stabilised and

therefore less likely to destack or melt than regions with less negative stacking energies.

Propeller twist angles can also be used as a measure for the rigidity of DNA, since the two have been shown to be inversely related [2]. Several sets of dinucleotide values for propeller twist exist [2, 5, 13], but they are all similar. The nine values determined by el Hassan and Caladine from crystallographic data of DNA oligomers were used in this work, and a theoretical estimate by Gorin et al. was used for the remaining TA step [2, 5].

Protein-induced deformability is a dinucleotide model for how easily DNA is deformed by proteins. The values have been determined by comparison of crystal structures of more than a hundred DNA/protein complexes and crystal structures of pure DNA [13]. The protein-induced deformability is a measure of the size of the conformational space covered by DNA in protein complexes compared to that of pure DNA.

2.2.3. *Flexibility measures*

DNase I is most often used for mapping the footprint of transcription factors or the location of nucleosomes. But because DNase I has low preference for cutting at specific sequences and requires the DNA to be bent before cutting, the DNase I sensitivity of naked DNA can be used as a measure for the bendability or anisotropic flexibility. By fitting a trinucleotide model to DNaseI experiments, a set of parameters for prediction of DNaseI sensitivity has been obtained [1]. On this scale a higher value corresponds to a more flexible sequence.

Another measure of flexibility is based on a set of 32 trinucleotide values giving the log-odds of the minor groove facing outwards when wrapped around a nucleosome core [16]. On this scale a value of zero represents no preference of the trinucleotide for specific positions in the nucleosomes, while large absolute values mean that the trinucleotide has strong preference. Because large absolute values thereby imply that the sequence is inflexible, a measure of flexibility is obtained by removing the sign

from the original trinucleotide values [15]. On that scale low values correspond to high bendability.

2.3. Base composition

The trivial way to parameterise the base composition is to simply use the G, A, T, and C content. A drawback of this representation is that the four parameters are mutually correlated as they sum to 1. An alternative parameterisation for the base composition is $A + T$, $A - T$, and $G - C$. In addition to being mutually independent measures, they also have the advantage of being easier to interpret in a biological context. In the base composition atlas both parameterisations are shown.

The $A + T$ content is strongly correlated with the structural parameters described above - especially the stacking energy. $A + T$ -rich regions usually destack more readily, have a higher intrinsic curvature, and are less flexible. Since the parameters $A - T$ and $G - C$ have almost no correlation with the structural properties of DNA, the $A + T$ content contains all the structural information arising from the mononucleotide composition.

$G - C$ and $A - T$ are strongly related to a set of measures known as skews, which are useful for locating the origin and terminus of replication in bacteria [9]. A number of different skews can be calculated for DNA sequences: GC skew ($(G - C)/(G + C)$), AT skew ($(A - T)/(A + T)$), purine skew ($(G + A - T - C)/N$), and keto skew ($(G + T - A - C)/N$) [11]. The keto and purine skews are correlated with each other as well as with the GC and AT skew.

The variance of the GC and AT skew is dependent on the $A + T$ content, for which reason strong fluctuations in GC skew are likely to occur in AT-rich regions; the opposite is true for the AT skew. Therefore, we prefer to use $G - C$ and $A - T$ instead of the corresponding skews. Because $G - C$ is often more useful than $A - T$ for locating the origin and terminus of replication, only $G - C$ and the $A + T$ content are included in the composite atlas.

3. Results and discussion

3.1. The *M. genitalium* atlas

To illustrate the method, we present an atlas of the *M. genitalium* genome (figure 1). A somewhat surprising feature, considering the size of the genome, is the presence of global repeats [6]. These repeats are clearly visible in the atlas, which also reveals a clear strand preference for the repeats, in the sense that the copies that constitute a repeat usually occur on the same strand in the DNA double helix. A similar, strong strand bias of global repeats was also found for *M. pneumoniae* and to a lesser degree for most other bacterial genomes (data not shown).

In figure 1, the rRNA operon stands out by having very strong stacking interactions and an unusually low $A + T$ content. This is similar to what is observed for *H. influenzae* [3], *T. maritima* [12], as well as other organisms (data not shown). By careful inspection of the *M. genitalium* atlas we found the same to be true for tRNA genes.

The $A + T$ content also varies at the global scale. Looking at the $A + T$ circle in figure 1, we find that the lower right part has higher GC content (approx. 64% $A + T$) than average for the genome whereas the upper left part is more AT-rich (approx. 71% $A + T$). We have found no biological explanation for this observation.

With the exception of the GC-rich region, the innermost circle in the *M. genitalium* atlas ($G - C$) resembles that of most other bacteria. While the origin of replication is easily identified, the precise location of the terminus is hard to predict - if a precise location in fact exists. Unlike other bacteria, the gene orientation is also strongly correlated with the direction of replication in *M. genitalium* [10].

3.2. Atlases for other organisms

GenomeAtlases are not limited to the study of small genomes such as *M. genitalium*. We have also applied the method to other microbial genome, in particular *E. coli*, where regions such

as the *rhs* elements light up by having unusual structural properties (D. Ussery et al., unpublished).

We have created GenomeAtlases for all the fully sequenced microbial chromosomes that are publicly available. These atlases are available on the internet at <http://www.cbs.dtu.dk/services/GenomeAtlas/>.

Acknowledgments

This work was supported by a grant from the Danish National Research Foundation.

References

- [1] Brukner I., Sanchez R., Suck D., Pongor S., Sequence-dependent bending propensity of DNA as revealed by DNase I: Parameters for trinucleotides, *EMBO J.* 14 (1995) 1812–1818.
- [2] ElHassan M., Calladine C., Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA, *J. Mol. Biol.* 259 (1996) 95–103.
- [3] Fleischmann R.D. et al., Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, *Science* 269 (1995) 496–512.
- [4] Gabrielian A., Pongor S., Correlation of intrinsic DNA curvature with DNA property periodicity, *FEBS Lett.* 393 (1996) 65–68.
- [5] Gorin A., Zhurkin V., Olson W., B-DNA twisting correlates with base-pair morphology, *J. Mol. Biol.* 247 (1995) 34–48.
- [6] Hancock J., Simple sequences in a 'minimal' genome, *Nature Genetics* 14 (1996) 14–15.
- [7] Haran T., Kahn J., Crothers D., Sequences elements responsible for DNA curvature, *J. Mol. Biol.* 225 (1994) 729–738.
- [8] Htun H., Dahlberg J., Single strands, triple strands, and kinks in H-DNA, *Science* 241 (1988) 1791–1796.
- [9] Lobry J., Asymmetric substitution patterns in the two DNA strands of bacteria, *Mol. Biol. Evol.* 13 (1996) 660–665.
- [10] Lobry J., Origin of replication of *Mycoplasma genitalium*, *Science* 272 (1996) 745–746.
- [11] McLean M., Wolfe K., Devine K., Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes, *J. Mol. Evol.* 47 (1998) 691–696.
- [12] Nelson K.E. et al., Evidence of lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*, *Nature* 399 (1999) 323–329.
- [13] Olson W., Gorin A., Lu X., Hock L., Zhurkin V., DNA sequence-dependent deformability deduced from protein-DNA crystal complexes, *Proc. Natl. Acad. Sci. USA* 95 (1998) 11163–11168.
- [14] Ornstein R., Rein R., Breen D., Macelroy R., An optimized potential function for the calculation of nucleic acid interaction energies, I, Base stacking, *Biopolymers* 17 (1978) 2341–2360.
- [15] Pedersen A., Baldi P., Chauvin Y., Brunak S., DNA structure in human RNA polymerase II promoters, *J. Mol. Biol.* 281 (1998) 663–673.
- [16] Satchwell S., Drew H., Travers A., Sequence periodicities in chicken nucleosome core DNA, *J. Mol. Biol.* 191 (1986) 659–675.
- [17] Shpigelman E., Trifonov E., Bolshoy A., CURVATURE: Software for the analysis of curved DNA, *CABIOS* 9 (1993) 435–444.
- [18] Sinden R., Pearson C., Potaman V., Ussery D., DNA: Structure and function, *Advances in Genome Biology* 5A (1998) 1–141.
- [19] Trifonov E., Sussman J., The pitch of chromatin DNA is reflected in its nucleotide sequence, *Proc. Natl. Acad. Sci. USA* 77 (1980) 3816–3820.
- [20] Ulanovsky L., Bodner M., Trifonov E., Curved DNA: Design, synthesis, and circularization, *Proc. Natl. Acad. Sci. USA* 83 (1986) 862–866.
- [21] Wells R., Collier D., Hanvey J., Shimizu M., Wohlrab F., The chemistry and biology of unusual DNA structures adopted by oligopyrimidine, oligopyrimidine sequences, *FASEB J.* 2 (1988) 2939–2949.