Visualization of Pathogenicity Regions in Bacteria

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

Center for Biological Sequence Analysis

Department of Biotechnology

Building 208

The Technical University of Denmark

DK-2800 Lyngby, Denmark

Abstract - We show here how pathogenicity islands can be analysed using GenomeAtlases, which is a method for visualising repeats, DNA structural characteristics, and base composition of chromosomes and plasmids. We have applied this method to the *E. coli* plasmid pO157, and the *Y. pestis* plasmid pPCP1. In both cases pathogenic genes were shown to differ in $A + T$ content and structural properties. Furthermore, examination of an antibiotic resistance gene cluster from *S. typhimurium* showed that the same was true for genes encoding antibiotic resistance.
Base composition , *Escherichia coli* , **Pathogenicity islands** , *Salmonella typhimurium DT104* , *Yersenia pestis*

Introduction

With the availability of DNA sequences from pathogenic organisms, computational analysis of these organisms and their toxicity has become feasible. We present here a method for visualizing pathogenicity islands so that repetitive sequences and anomalies in base composition or DNA structure become visible. The GenomeAtlas is a wheelplot summarizing such different properties of DNA [Jensen *et al.*, 1999]. This method can be applied to both small plasmids and gene clusters as seen in this work, but also to complete microbial chromosomes. 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/.

To illustrate the usefulness of the GenomeAtlas for finding genes responsible for pathogenicity, three GenBank entries from different organisms were examined. pO157, a 92077 bp plasmid from the pathogenic *E. coli* strain O157:H7 (GenBank accession number AF074613), was chosen because it is in part responsible for the hosts pathogenicity and is believed to encode at least one toxic protein [Burland *et al.*, 1998]. pPCP1, a plasmid from *Y. pestis* which carries two known virulence genes, was also selected. Whereas infection of humans with strains of *Y. enterocolitica* and *Y. pseudotuberculosis* typically result in diarrhea and abdominal pains, *Y. pestis* with the plasmid pPCP1 is the cause of the bubonic plague [Hu *et al.*, 1998]. Finally, a cluster of five antibiotic resistance genes from *S. typhimurium* DT104 was selected (GenBank accession number AF071555)[Briggs & Fratamico, 1999]. A recent outbreak of this multidrug resistant *Salmonella* in Denmark resulted in the death of two people [Molbak *et al.*, 1999]. Multidrug resistant *Salmonella* are becoming increasingly difficult to treat with antibiotics, and may become a major health concern in the future.

Methods

To generate wheelplots, 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 visualized as colored circles is made for each of these three categories; in addition the combined GenomeAtlas summarizing the most informative parameters is constructed.

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 dinucleotide or trinucleotide values that have been obtained by fitting either experimental results or theoretical estimates[1].

---

* To whom correspondence should be addressed.
Tel: (+45) 45 25 24 88; Fax: (+45) 45 93 15 85; email: dave@cbs.dtu.dk

---

[1]Pedersen *et al.*, "A DNA structural atlas of *E. coli*", manuscript submitted to J. Mol. Biol.

Intrinsic curvature is a property of DNA that is closely related to anomalous gel mobility, as DNA fragments with high intrinsic curvature will migrate slower on polyacrylamide gels than markers with the same length. In this work we have used the CURVATURE programme [Shpigelman *et al.*, 1993], which is based on a wedge model [Trifonov & Sussman, 1980, Ulanovsky *et al.*, 1986], 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. Curves are often encountered upstream of highly expressed genes such as the toxin genes in pathogenicity islands [Bracco *et al.*, 1989].

Next is the stacking energy, which relates to 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 [Ornstein *et al.*, 1978]. 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.

The position preference is a measure of helix flexibility based on a set of 32 trinucleotide values giving the log-odds of the minor groove facing outwards when wrapped around a nucleosome core [Satchwell *et al.*, 1986]. On this scale a value of zero represents no preference of the trinucleotide for specific positions in the nucleosomes, while large absolute values means that the trinucleotide has strong preference. Because large absolute values thereby implies that the sequence is inflexible, a measure of flexibility is obtained by removing the sign from the original trinucleotide values [Pedersen *et al.*, 1998]. On that scale low values correspond to high bendability.

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.

The $A + T$ content is strongly correlated to 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 to the structural properties of DNA, the $A + T$ content contains nearly all the structural information arising from the mono-nucleotide composition. Pathogenic islands can often be detected from their different $A + T$ content alone, which also influence almost all the structural properties of DNA.

Repeat 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 the same basic algorithm which finds the highest degree of homology for an R bp repeat within a window of length W.

Here we will only focus on repeats on the global scale, where a direct repeat is a sequence that is present in at least two copies on the same strand, whilst two copies located on opposite strands will give rise to an inverted repeat. Global repeats can arise from duplicated genes as well as from IS elements which are often seen flanking pathogenic islands [Hacker *et al.*, 1997].

Results and discussion

The plasmid pO157

The atlas of pO157 (figure 1) reveals two strongly curved $A + T$ rich areas which destack more readily than the surroundings. One consists primarily of the four *EHEC-hly genes*, the other is composed almost entirely of the gene *L7095*. Although no experimental evidence exists of the function of the proteins encoded by these genes, all four show remarkable resemblance (99% identity) to a hemolysin toxin protein (*EHEC-hlyA*) and three hemolysin transport proteins (*hlyB*, *hlyC*, and *hlyD*) [Burland *et al.*, 1998].

Not far from *L7095* an inverted repeat can be observed with the two copies located upstream and downstream of the gene itself. Considering also the aforementioned difference in base composition observed for the gene, it can be suggested that the repeats are in fact the boundaries of a foreign transposon which has been integrated into the plasmid. This hypothesis is supported by the presence of *L7093*, *L7094* and *L7096*, located on either side of *L7095*, which are believed to code for transposases. The function of the very large (3170 aa) protein coded for by *L7095* is not entirely known, but the protein has some similarity (approx. 22% identity) with a known cytotoxin from *Clostridium sordelii* [Burland *et al.*, 1998].

The plasmid pPCP1

The associated genes *pst* and *pim*, encoding the pesticin and pesticin immunity protein respectively are readily apparent on the atlas from their high $A + T$ content (figure 2). The

strongest structural signal is seen for the *pim* gene, which has several highly curved regions and appears to destack more readily. Although the *pst* gene itself shows no significant structural motifs, the promoter region is highly curved, suggesting high levels of expression [Bracco *et al.*, 1989]. A similar albeit weaker signal can be seen for *pla*, coding for the plasminogen activator [Hu *et al.*, 1998].

S. typhimurium DT104

As can be seen in table I, *S. typhimurium* DT104 is resistant to five antibiotics. The genes coding this resistance are clustered together within a small region of the DT104 chromosome. A GenomeAtlas of this region is shown in figure 3. The atlas reveals a high $A + T$ content for the β-lactamase gene, which also destacks readily and contains several peaks in intrinsic curvature. This signal continues into the repeated region and also encompasses the *ammB* gene. However, the remaining antibiotic resistance genes show no appreciable difference in any of the measurements.

Another striking feature of figure 3 are the long direct repeats located around the ends of the sequenced island. The presence of nearby integrase genes suggests that parts of the cluster might be foreign DNA integrated into the chromosome itself. Furthermore the regions surrounding the β-*lac* and *strep* genes are identical. Possibly only one of these regions were originally in DT104, while the other was (part of) a plasmid which was then inserted into the genome by cross-over between the homologous regions on the chromosome and the plasmid.

Transposons and other foreign regions are often involved in an organisms pathogenesis. We have shown that the GenomeAtlas is an effective tool for revealing these features based on their differences in base composition, structural properties, and the occurrence of nearby repeats.

Acknowledgements

References

[Bracco *et al.*, 1989] Bracco, L., Kotlarz, D., Kolb, A., Diekmann, S. & Buc, H. (1989). Synthetic curved DNA sequences can act as transcriptional activators in *Escherichia coli*. *EMBO J.*, **8**, 4289–4296.

[Briggs & Fratamico, 1999] Briggs, C. E. & Fratamico, P. M. (1999). Molecular characterization of an antibiotic resistance gene cluster of *Salmonella typhimurium* dt104. *Antimicrobial Agents and Chemotherapy*, **43**, 846–849.

[Burland *et al.*, 1998] Burland, V., Shao, Y., Perna, N. T., Plunkett, G., Sofia, H. J. & Blattner, F. R. (1998). The complete dna sequence and analysis of the large virulence plasmid of *Escherichia coli* o157:h7. *Nucleic Acids Research*, **26**, 4196–4204.

[Hacker *et al.*, 1997] Hacker, J., Blum-Oehler, G., Muhldorfer, I. & Tschape, H. (1997). Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol. Microbiol*, **23**, 1089–97.

[Hu *et al.*, 1998] Hu, P., Elliot, J., McCready, P., Skowronski, E., Garnes, J., Kobayashi, A., Brubaker, R. R. & Garcia, E. (1998). Structural organization of virulence-associated plasmids of *Yersinia pestis*. *Journal of Bacteriology*, **180**, 5192–5202.

[Jensen *et al.*, 1999] Jensen, L., Friis, C. & Ussery, D. (1999). Three views of microbial genomes. *Research in Microbiology*, **150**, 773–777.

[Molbak *et al.*, 1999] Molbak, K., Baggesen, D. L., Aarestrup, F. M., Ebbesen, J. M., Engberg, J., Frydendahl, K., Gerner-Smidt, P., Petersen, A. M. & Wegener, H. C. (1999). An outbreak of mutlidrug-resistant, quinolone-resistant salmonella enterica serotype typhimurium dt104. *The New England Journal of Medicine*, **341**, 1420–1425.

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

[Pedersen *et al.*, 1998] Pedersen, A., Baldi, P., Chauvin, Y. & Brunak, S. (1998). DNA structure in human RNA polymerase II promoters. *J. Mol. Biol.*, **281**, 663–673.

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

[Shpigelman *et al.*, 1993] Shpigelman, E., Trifonov, E. & Bolshoy, A. (1993). CURVATURE: Software for the analysis of curved DNA. *CABIOS*, **9**, 435–444.

[Trifonov & Sussman, 1980] Trifonov, E. & Sussman, J. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl. Acad. Sci. USA*, **77**, 3816–3820.

[Ulanovsky *et al.*, 1986] Ulanovsky, L., Bodner, M. & Trifonov, E. (1986). Curved DNA: Design, synthesis, and circularization. *Proc. Natl. Acad. Sci. USA*, **83**, 862–866.

Fig. 1. The GenomeAtlas of *E. coli* plasmid pO157.

Fig. 2. The GenomeAtlas of *Y. pestis* pPCP1

Gene	Product
IntA	Integrase
Strep	Res. to Streptomycin/spectinomycin
AmmA	Res. to Quaternary ammonium compounds (disinfectants)
Chlor	Res. to Chloramphenicol
RegT	Regulation of Tetracycline resistance
Tetra	Res. to Tetracycline
IntB	GroEL-like/integrase fusion protein
β-lac	Res. to β-Lactams (ampicillin etc.)
AmmB	Res. to Quaternary ammonium compounds (disinfectants)
Sulph	Res. to Sulphonamides

Table I. Genes present in the *S. typhimurium* DT104 island
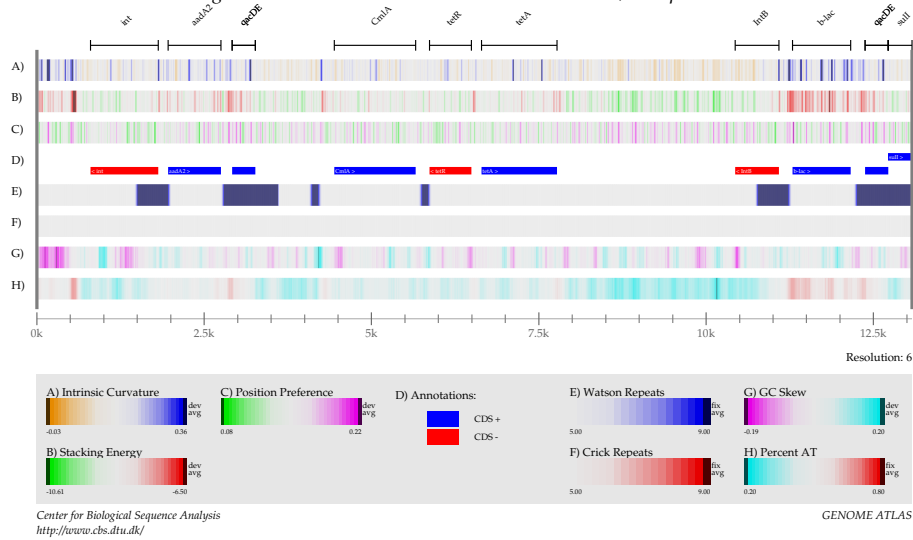
Fig. 3. The GenomeAtlas of *S. typhimurium* DT104