

The Atlas Visualisation of Genome-wide Information

Marie Skovgaard, Lars Juhl Jensen, Carsten Friis, Hans Henrik Stærfeldt,
Peder Worning, Søren Brunak, and David Ussery*

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

January 4, 2002

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

Introduction

The wealth of information contained in a microbial genome is not easy to comprehend at all scales. Even after the genome of an organism has been sequenced, the problem of gaining an overview of the newly acquired data still remains. One way to get an overview is to visualise positional features at the chromosome level; we have developed a method, Atlases, for showing correlations between position dependent information in sequenced chromosomes.

The DNA sequence is not only hard to comprehend because of its size but also because the genomic sequence is not linked in a simple way to the biology of the organism. For example, examining the AT content of genomes, a property often reported in genome sequencing papers, considerable variation is observed. Figure 1 shows the percent AT of 25 different proteobacterial genomes, ranging from 33% to 74% AT. The AT content does not appear to correlate to the proteobacterial subdivisions.

The percent AT of a chromosome reflects only an *average* property of the chromosome. However, the AT content is not homogeneously distributed throughout the DNA. Often there are clusters of AT-rich and AT-poor regions; for example most promoter regions are more AT-rich than the average coding sequences (Ozoline *et al.*, 1999; Pedersen *et al.*, 2000). In many cases the variations between regions will tell more than the average value, as exemplified in Figure 2 where an AT-rich region is found to contain genes involved in pathogenesis.

The AT content within a region of a chromosome is a very simple property to calculate from the nucleotide sequence. More complex features like DNA curvature or major groove compressibility, that reflect structural properties of a given region, can be estimated directly from the sequence and give biological insight. Additional information can be accessed by looking at the genes encoded by the chromosome. Once the location of the genes is known, it is possible to visualise both experimentally determined expression levels and RNA sequence features. By translating the RNA sequence to protein sequence, it is also possible to visualise properties of the proteome, such as protein function.

1 Construction of the visualisation software

In order to be able to visualise such diverse data, a flexible software tool is needed. We have developed a computer programme, GeneWiz, which enables us to visualise a complete chromosome compactly. The programme creates an either circular or linear graphical representation of the entire chromosome or of a specified subsection. Sufficiently large regions that display significant variation from the rest of the chromosome can be readily found — in order to be able to see deviations of smaller regions a zoomed Atlas must be made.

Each feature, such as AT content or gene expression level, is represented as a separate lane in the atlas. The value is at each position colour coded according to a user specified colour scale. We generally use colour scales where regions of extreme values are highlighted (this can be one-ended or two-ended scales) whereas typical values are grey. If wanted the plot can be smoothed by a running average.

The properties to be visualised must be present in the form of one value per basepair in the chromosome. For simple sequence features like the AT content this is the natural format, whereas for data such as gene expression levels, the value for each gene must be mapped onto the corresponding range of basepairs. In addition to the data series, the annotations from a GenBank file can be displayed using a series of icons with user-defined colours. This allows for the identification of short or long annotated regions of interest.

GeneWiz is solely a visualisation program, and is not capable of calculating the data used in the different atlases. All data must be calculated and properly formatted. While this obviously adds to the work of creating an atlas, it gives great flexibility as these data can come from any source. In this paper we use simple measures generated from lookup tables, publicly available programs like BLAST (Altschul *et al.*, 1997), methods developed in-house like ProtFun (Jensen *et al.*, 2002) as well as experimentally determined expression data.

2 Use of Atlases to Visualise DNA Information

2.1 Genome Atlases

The GenomeAtlas is a general atlas made for all the fully sequenced microbial chromosomes found in public databases (Pedersen *et al.*, 2000; Jensen *et al.*, 1999). The GenomeAtlas is a

combination of some generally informative parameters and can be used as an offset for identifying unique regions or special features for the given chromosome. The GenomeAtlas for all public available sequenced chromosomes can be found at <http://www.cbs.dtu.dk/services/GenomeAtlas/>.

Introducing the parameters

To generate GenomeAtlas plots, 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. These three categories are combined into a common atlas where the parameters are visualized, giving the values of the parameters as the intensity of the colour (Jensen *et al.*, 1999).

2.1.1 Structural parameters

A number of measures for the local structure of DNA have been devised, most of which are based dinucleotide or trinucleotide models that have been obtained by fitting either experimental results or theoretical estimates (Pedersen *et al.*, 1998; Pedersen *et al.*, 2000).

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 (Bracco *et al.*, 1989).

Stacking energy 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 histone octamer (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.

2.1.2 Base composition

The trivial way to parameterise the base composition is to simply use the G-, A-, T-, and C-contents. 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 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. The parameter G-C, known as the GC skew (McLean *et al.*, 1998) reflects a general bias of purines towards the leading strand of DNA replication (Tillier & Collins, 2000). Since the GC skew has almost no correlation to the structural properties of DNA, the A+T content contains nearly all the structural information arising from the mononucleotide composition.

2.1.3 Repeat elements

Repeats are multiple copies of the same sequence at different locations on a piece of DNA. The repeats can be found either by a very accurate method using a basic algorithm which finds the highest degree of homology for an R bp long repeat within a window of length W (Jensen *et al.*, 1999), or by cutting the sequence up in fragments and using the heuristic alignment algorithm BLAST (Altschul *et al.*, 1997) to find the homologous regions with the length R . The basic algorithm is more accurate than BLAST but it is also computationally demanding, therefore BLAST

is used on large sequences. There are two kinds of repeats, 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.

2.2 GenomeAtlases of Pathogenicity Plasmids

A GenomeAtlas can give a quick overview of a given chromosome and thereby be the reason for further analysis of a given organism or a more specific search for a given feature can be made by looking through a collection of atlases. The latter was the case when a study of pathogenicity islands in bacterial plasmids was based on the knowledge of the correlation between pathogenicity islands and variation in AT content, such as the toxin genes in plasmid pO157 from pathogenic *E. coli* strains (Friis *et al.*, 2000). Another example of the correlation between pathogenicity islands and changes in AT content can also be found in the large virulence plasmid of *Shigella flexneri* (GenBank accession number AF348706) (Venkatesan *et al.*, 2001).

The atlas of the *Shigella flexneri* 5a virulence plasmid pWR501 (Figure 2) reveals an A+T rich area, which is strongly curved, will destack or melt more readily than the rest of the plasmid, and is more rigid. This region encodes a locus of genes (*ipa-mxi-spa*) involved in the pathogenic invasion of mammalian cells, and includes a type III secretion pathway (Schuch *et al.*, 1999; Page *et al.*, 2001).

Variations in AT content are obviously not always correlated with the presence of toxic genes; another indication of potential pathogenic regions can be the localisation of multiple repeats (especially Insertion Sequence (IS) elements) (Hacker *et al.*, 1997; Hacker & Kaper, 2000). Large numbers of direct and inverted repeats can be seen in Figure 2. Typically, global direct repeats account for around three percent or less of most bacterial chromosomes (data not shown, but “GenomeAtlases” for all sequenced genomes can be found on our web page). Many of the repeats (especially the global inverted repeats) are reflective of IS elements. Note that the A+T rich *ipa-mxi-spa* region is the largest region free of repeats in the plasmid.

Another example of a plasmid with many repeats is the plasmid pBtoxis¹ from the spore-forming bacteria *Bacillus thuringiensis* subsp. *israelensis*. Like pWR501, the repeats in pBtoxis are scattered all over the plasmid (Figure 3); a search was made for genes from transposable

¹This sequence data were produced by the Microbial Genomes Sequencing Group at the Sanger Institute and can be obtained from <ftp://ftp.sanger.ac.uk/pub/pathogens/bti/>

elements like transposases and integrases and by doing a simple BLAST search against SWISS-PROT these genes were located and they were indeed found to be associated with the repeats. In the case of this plasmid the presence of transposable elements was known long before the plasmid was sequenced (Mahillon *et al.*, 1994), but the GenomeAtlas can be used as an easy method for localisation of transposons and IS elements.

B. thuringiensis is used in agriculture as an alternative to synthetic chemical pesticides. It produces parasporal crystals that have insecticidal activity, and the genes that are believed to be responsible for this activity are marked in Figure 3 (Schnepf *et al.*, 1998). The transposable elements in pBtoxis are at least partly responsible for the high degree of genetic plasticity that makes *B. thuringiensis* adaptable to a variety of environments. However, it should also lead to caution in the use, since *B. thuringiensis*, based on genetic evidence, is from the same species as *Bacillus anthracis* and *Bacillus cereus*, both human pathogens (Helgason *et al.*, 2000).

Some of the repeats that are not associated with genes from transposable elements seem to be copies of *cry*, the gene for the pesticide crystal protein. The three *cyt* genes produce cytolytic delta-endotoxins; the absence of repeats in this area indicates that they are not similar to each other at the nucleotide level.

In the case of the two plasmids presented here the atlas was used as a method to screen large plasmids for signs that indicated the presence of toxic genes; many pathogenic regions within plasmids might not be found in this way, but the atlas serves as a very strong method for initial examination of the sequences (Friis *et al.*, 2000).

2.2.1 A custom made DNA Atlas

The GenomeAtlas is our “standard” atlas, which can capture interesting features of a chromosome. As an example, consider chromosome 1 from the protozoan *Leishmania major*, an intracellular pathogen of the immune system. This chromosome has an unusual organisation of its genes, with the 79 protein coding genes being in two large clusters. The first 29 genes are coded on one strand whilst the last 50 genes are on the other strand (Mylers *et al.*, 1999). From the GenomeAtlas² a correlation between intergenic regions and global repeats can be observed. In order to further investigate the possible relationship between other structural parameters and intergenic regions,

²The atlas can be seen at <http://www.cbs.dtu.dk/services/GenomeAtlas/Eukaryotes/Leishmania/major/>

DNA Property	Length (bp)	% Direct	% Inverted	% (Y)10	% (YR)5	% AT
Coding	140,229	4.5	0.0	1.2	2.6	34.6
Non coding	128,755	6.0	4.6	11.2	6.9	39.4
Whole chromosome	268,984	5.2	2.2	6.0	4.7	36.9

Table 1: Characteristics of coding and noncoding sequences in *Leishmania major* chromosome 1

we constructed a custom atlas (see Figure 4).

Several properties of the chromosome are revealed by the base composition parameters (AT content and GC-skew). The telomeres and a region around 80 kbp have a much higher AT content than the rest of the chromosome. Also a shift in the GC-skew is observed around 80 kbp, correlating with the unusual gene organisation. This is in agreement with the region being proposed as the origin of replication (McDonagh *et al.*, 2000).

More direct than inverted repeats are found in this chromosome. Some of these arise from gene duplications — the most obvious example is the two genes around position 240 kbp. Even though gene duplications are observed, the direct repeats still exhibit a slight preference for non-coding regions. This preference is much stronger for inverted repeats, which occur exclusively in intergenic regions, as shown in Table 1.

The exclusive localisation of inverted repeats in intergenic regions prompted an interest in whether other DNA structural elements might also be preferentially positioned within non-coding regions. Runs of purines (or pyrimidines) as well as alternating pyrimidine/purine stretches occur more often than would be expected from the base-composition of *Leishmania major* (Ussery *et al.*, 2002). The location of these regions was visualised by plotting the location of all such stretches of at least 10 bp. Many purine stretches can adopt an A-DNA conformation, whereas pyrimidine/purine stretches that are GC-rich can adopt a Z-DNA conformation. There is a strong preference (about 10-fold, see (Y)10 column in table 1) for purine stretches in the intergenic regions, while the pyr/pur regions are less strongly correlated with the non-coding DNA.

3 Atlases for Visualising genome-wide RNA expression

It has often been said that even non-coding DNA is far from a random string of bases. Since helix structure is a function of that same string of bases, this statement must apply to structural features as well. These structural features are suspected of affecting not just the termination of transcription as mentioned earlier, but also the rate of transcription itself.

Almost unheard of five years ago, genome-wide mRNA expression analysis has become mainstream in most major microbiological laboratories. With this technology it is feasible to examine the transcription levels for an entire microbial genome under a broad range of different circumstances, and to some degree reverse engineer regulatory pathways (Spellman *et al.*, 1998).

By visualizing measured levels of transcription in an atlas it becomes possible to examine if correlations exist between the mRNA expression levels and DNA structural properties or base composition. Such correlations could be expected due to chromatin packing (Ussery *et al.*, 2001). Also the relationship between the level of transcription and chromosomal location may reveal interesting aspects (Hughes *et al.*, 2000b).

3.1 ExpressionAtlas

The strength of genome-wide RNA expression analysis lies in the ability to simultaneously measure the expression levels of an entire genome. When analysing several arrays with thousands of genes one is faced with much the same problem as when analysing whole genome sequences: the sheer amount of data makes it hard to get an overview. The ExpressionAtlas is a way of visualising expression experiments taking into account chromosomal position and other factors suspected of being involved in transcription, such as DNA structure and repeats.

In the example shown, the average intensities from cDNA arrays (Cho *et al.*, 1998) were used as an estimate of the constitutive expression levels of genes in *Saccharomyces cerevisiae*. Alternatively log-fold changes could be plotted to highlight regulated genes. We chose average intensities to ensure comparability to the predicted expression levels also displayed on the atlas.

Neural networks trained on average expression values from *E. coli* microarray experiments predicted the expression level of each gene. The predicted levels of expression were normalized to a range from 0 to 1. As input to the neural networks the trinucleotide frequencies of the coding regions were used. These 64 frequencies were calculated without taking the reading frame into

account. This representation was chosen because the majority of DNA structural properties can be captured at the trinucleotide level. In this way we can capture possible correlation between the structural properties of the coding DNA and the expression levels.

Both the experimentally measured and the predicted expression levels are displayed in the atlas in Figure 5, together with position preference, global repeats and AT content. The AT content and the global repeats were included to give a general view of the composition of the chromosome, whereas the position preference, being a measure of the flexibility of the double helix, is expected to be correlated with the expression levels (Pedersen *et al.*, 2000). The inverted repeats clearly mark the telomeric regions.

Comparing the actual expression levels with the levels predicted by neural networks reveals a strong correlation between the two. A similar, albeit weaker, correlation is observed between expression levels and the position preference measure. The fact that neural networks trained on the prokaryote *E. coli* data can predict highly expressed genes in a eukaryote implies the existence of universal DNA properties that influence transcription. The correlation with the position preference measure suggests that helix flexibility plays a part in this. Speculations on such generic features of expressed genes have been proposed before (Sharp & Li, 1987).

4 Atlases for Visualizing Global Prediction of Protein Function

By looking at the ExpressionAtlas it is possible to identify regions with genes that are highly expressed (and possibly regulated) under one or more experimental conditions. It is at this point obvious to ask what the function of these genes might be.

Unfortunately the function of a large fraction genes remains unknown in most fully sequenced chromosomes. Of the 30,000 to 50,000 genes believed to be present in the human genome no more 40-60% can be assigned a functional role based on homology to known proteins. Even though the situation is a bit more favourable when looking at simpler model organisms like *S. cerevisiae* and *C. elegans*, the function of more than 30% of the predicted protein sequences still remains unknown.

In newly sequenced chromosomes most of the functional annotation of genes is based on homology inference. Using methods such as BLAST (Altschul *et al.*, 1997) homologous proteins are identified by sequence similarity and the function is inferred from the knowledge about the

homologs. However it is usually the case that somewhere from 30–50% of the proteins give no matches to proteins of known function. These are known as “orphan” proteins.

Traditionally, protein function has been viewed as something directly related to the conformation of the polypeptide chain. However, as the three-dimensional structure currently is quite hard to calculate from the sequence (Lesk *et al.*, 2001), a computational strategy for the elucidation of orphan protein function may benefit also from the prediction of functional attributes which are more directly related to the linear sequence of amino acids.

Our approach to function prediction is based on the fact that a protein is not alone when performing its biological task. As it will have to operate using the same cellular machinery for modification and sorting as all the other proteins do, one can expect some conservation of essential types of post-translational modifications (PTMs). Because reasonably precise methods for prediction of PTMs from sequence exist today, our prediction method which integrates such relevant features to assign orphan protein to functional class, can be applied to all proteins where the sequence is known (Jensen *et al.*, 2002; Gupta *et al.*, 2002). This is in contrast to methods that rely on clustering of co-expressed genes (Eisen *et al.*, 1998), prediction of gene fusions and/or phylogenetic profiles (Marcotte *et al.*, 1999a; Marcotte *et al.*, 1999b; Hughes *et al.*, 2000a; Pellegrini *et al.*, 1999).

For any function prediction method, the ability to correctly assign the relationship depends strongly on the function classification scheme used. We predict a scheme of twelve cellular functions that is closely related to the fourteen class classification originally proposed by Riley for the *E. coli* genome (Riley, 1998). The system consists of an ensemble of neural networks for each functional category, each neural network having a different combination predicted protein features as its input. The networks were trained exclusively on human protein sequences, but perform well on a wide selection eukaryotes (including *S. cerevisiae*). For each protein sequence the outputs of these neural networks are subsequently combined into a probability for each category.

We have applied this software to all predicted protein sequences from *S. cerevisiae* chromosome VIII. Based on our performance estimates of the method on *S. cerevisiae* sequences, we have selected a subset of eight categories out of the original twelve category system. The probabilistic scores of each protein sequence were mapped onto the position in the chromosome of the corresponding gene. Figure 6 shows the resulting FunctionAtlas along with the actual expression levels

also shown in the ExpressionAtlas.

One feature that is visible from a FunctionAtlas is clusters of genes with related functions. Examples of this include the regions 10k–50k and 250k–260k that contain very large numbers of predicted transport and binding proteins. The regions 50k–60k, 335k–350k and 410k–420k that are predicted to contain a large number of genes involved in replication or transcription, several of which are likely to serve a regulatory role according to our predictions. Since genes of related function are known to often cluster (although the extent varies from organism to organism) predicted functional clusters can be trusted more than individual predictions. If the function of some of the genes within a cluster is known and in agreement with the prediction — as is the case for several of the mentioned clusters — this obviously adds to the evidence.

Another possibility is to correlate the predicted protein function to expression data. Close inspection of Figure 6 reveals that many of the constitutively highly expressed genes are predicted to be involved in energy metabolism although this is overall a quite rare category. A hypergeometric test of the underlying data verifies that this correlation is indeed significant at a 95% confidence level. A large number of highly expressed transcripts for proteins involved in replication and transcription can also be identified although no correlation between function and expression level is found in this case.

Concluding remarks

In summary, we have shown several different applications of Atlases for visualising different type of information, within the context of the whole plasmid or chromosome. Essentially, any type of information concerning the DNA, RNA or protein can be plotted along the chromosome, allowing for rapid analysis of global properties in a serendipitous manner. The atlas gives the researcher the option to view the calculated or experimentally measured data in a position dependent way and thereby see correlations between a feature and its position or variation in a feature within the chromosome.

The atlas can be used to spot variation in different features within a region but not all the information can be viewed at the same time. For a bacterial chromosome of the same size as *E. coli* only features with approximately the same size as a gene can be observed, this means that some of the features showed for *L. major* with repeats in intergenic regions would not be visible

in the *E. coli* genome. If variation in a smaller scale is to be seen for a large chromosome a shorter region should be visualized.

Acknowledgements

The authors would like to acknowledge and thank the people at CBS for their help, in particular Ramneek Gupta, Steen Knudsen, Anders Krogh, and Anders Gorm Pedersen. This work was supported by a grant from the Danish National Research Foundation.

References

- Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- 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.
- 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.
- Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D. & Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
- Eisen, M., Spellman, P., Brown, P. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**, 14863–14868.
- Friis, C., Jensen, L. & Ussery, D. (2000). Visualization of pathogenicity regions in bacteria. *Genetica*, **108**, 47–51.
- Gupta, R., Jensen, L. & Brunak, S. (2002). Orphan protein function and its relation to glycosylation. In Mewes, H., Weiss, B. & Seidel, H., (eds.) *Ernst Schering Research Foundation Proceedings*. Springer-Verlag, Berlin, p. Chapter 13.
- 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**, 1089–1097.
- Hacker, J. & Kaper, J. (2000). Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.*, **54**, 641–679.
- Helgason, E., Økstad, O., Caugant, D., Johansen, H., Fouet, A., Mock, M., Hegna, I. & Kolstø, A.

- (2000). *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis* - one species on the basis of genetic evidence. *Appl Environ Microbiol*, **66**, 2627–30.
- Hughes, T., Marton, M., Jones, A., Roberts, C., Stoughton, R., Armour, C., Bennett, H., Coffey, E., Dai, H., He, Y., Kidd, K., King, A., Meyer, M., Slade, D., Lum, P., Stepaniants, S., Shoemaker, D., Gachotte, D., Chakraburttu, K., Simon, J., Bard, M. & Friend, S. (2000a). Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
- Hughes, T., Roberts, C., Dai, H., Jones, A., Meyer, M., Slade, D., Burchard, J., Dow, S., Ward, T., Kidd, M., Friend, S. & Marton, M. (2000b). Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat. Genet.*, **25**, 333–337.
- Jensen, L., Friis, C. & Ussery, D. (1999). Three views of the *E. coli* genome. *Research in Microbiology*, **150**, 773–777.
- Jensen, L., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Stærfeldt, H., Rapacki, K., Workman, C., Andersen, C., Knudsen, S., Krogh, A., Valencia, A. & Brunak, S. (2002). *Ab initio* prediction of human orphan protein function from post-translational modifications and localization features. *JMB*.
- Lesk, A. M., Conte, L. & Hubbard, T. (2001). Assessment of novel fold targets in CASP4: Predictions of three-dimensional structures, secondary structures and interresidue contacts. *Proteins*, to appear.
- Mahillon, J., Rezsóhazy, R., Hallet, B. & Delcour, J. (1994). IS231 and other *Bacillus thuringiensis* transposable elements: a review. *Genetica*, **93**, 13–26.
- Makino, K., Ishii, K., Yasunaga, T., Hattori, M., Yokoyama, K., Yutsudo, C., Kubota, Y., Yamaichi, Y., Iida, T., Yamamoto, K., Honda, T., Han, C., Ohtsubo, E., Kasamatsu, M., Hayashi, T., Kuhara, S. & Shinagawa, H. (1998). Complete nucleotide sequences of 93-kb and 3.3-kb plasmids of an enterohemorrhagic *Escherichia coli* O157:H7 derived from sakai outbreak. *DNA Res.*, **5**, 1–9.
- Marcotte, E., Pellegrini, M., Ng, H., Rice, D. W., Yeates, T. & Eisenberg, E. (1999a). Detecting

- protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.
- Marcotte, E., Pellegrini, M., Thompson, M., Yeates, T. & Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
- McDonagh, P., Myler, P. & Stuart, K. (2000). The unusual gene organization of *Leishmania major* friedlin chromosome 1 may reflect novel transcription processes. *Nucl. Acids Res.*, **28**, 2800–2803.
- McLean, M., Wolfe, K. & Devine, K. (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryotic genomes. *J. Mol. Evol.*, **47**, 691–696.
- Myler, P., Audleman, L., de Vos, T., Hixson, G., Kiser, P., Magness, C., Rickel, E., Sisk, E., Sunkin, S., Swartzell, S., Westlake, T., Bastein, P., Fu, G., Ivens, A. & Stuart, K. (1999). *Leishmania major* friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proc. Natl. Acad. Sci. USA*, **96**, 2902–2906.
- 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.
- Ozoline, O., Deev, A., Arkhipova, M., Chasov, V. & Travers, A. (1999). Proximal transcribed regions of bacterial promoters have a non-random distribution of A/T tracts. *Nucleic Acids Res.*, **27**, 4768–4774.
- Page, A., Fromont-Racine, M., Sansonetti, P., Legrain, P. & Parsot, C. (2001). Characterization of the interaction partners of secreted proteins and chaperones of *Shigella flexneri*. *Mol. Microbiol.*, **42**, 1133–1145.
- Pedersen, A., Baldi, P., Chauvin, Y. & Brunak, S. (1998). DNA structure in human RNA polymerase II promoters. *J. Mol. Biol.*, **281**, 663–673.
- Pedersen, A., Jensen, L., Stærfeldt, H., Brunak, S. & Ussery, D. (2000). A DNA structural atlas of *E. coli*. *J. Mol. Biol.*, **299**, 907–930.

- Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D. & Yeates, T. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, **96**, 4285–4288.
- Riley, M. (1998). Systems for categorizing functions of gene products. *Curr. Opin. Struct. Biol.*, **8**, 388–392.
- Satchwell, S., Drew, H. & Travers, A. (1986). Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, **191**, 659–675.
- Schnepf, E., Crickmore, N., Rie, J. V., Lereclus, D., Baum, J., Feitelson, J., Zeigler, D. & Dean, D. (1998). *Bacillus thuringiensis* and its pesticidal crystal proteins. *Microbiol Mol Biol Rev*, **62**, 775–806.
- Schuch, R., Sandlin, R. & Maurelli, A. (1999). A system for identifying post-invasion functions of invasion genes: requirements for the *Mxi-Spa* type III secretion pathway of *Shigella flexneri* in intercellular dissemination. *Mol. Micro.*, **34**, 675–689.
- Sharp, P. M. & Li, W. H. (1987). The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
- Shpigelman, E., Trifonov, E. & Bolshoy, A. (1993). CURVATURE: Software for the analysis of curved DNA. *CABIOS*, **9**, 435–444.
- Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *S. cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
- Tillier, E. & Collins, R. (2000). The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.*, **50**, 249–257.
- 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, L., Bodner, M. & Trifonov, E. (1986). Curved DNA: Design, synthesis, and circularization. *Proc. Natl. Acad. Sci. USA*, **83**, 862–866.

- Ussery, D., Soumpasis, D., Brunak, S., Stærfeldt, H., Worning, P. & Krogh, A. (2002). Bias of purine stretches in sequenced genomes. *Computers in Chemistry*, **26**, in press.
- Ussery, D. W., Larsen, T., Wilkes, K., Friis, C., Worning, P., Krogh, A. & Brunak, S. (2001). Genome organisation and chromatin structure in *Escherichia coli*. *Biochimie*, **83**, 201–212.
- Venkatesan, M., Goldberg, M., Rose, D., Grotbeck, E., Burland, V. & Blattner, F. (2001). Complete DNA sequence and analysis of the large virulence plasmid of *Shigella flexneri*. *Infetc. Immun.*, **69**, 3271–3285.

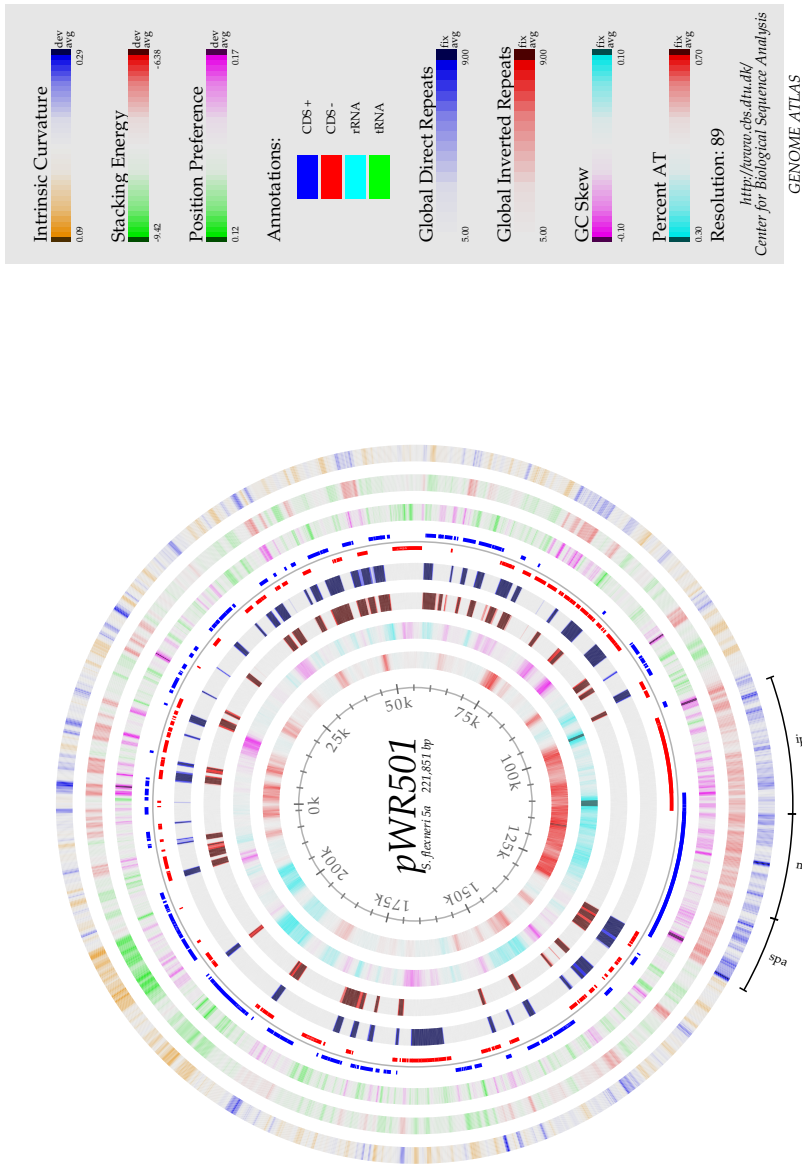


Figure 2: GenomeAtlas for *Shigella flexneri* 5a virulence plasmid pWR501. The marked regions contain three loci which codes for a total of 34 virulence-related genes.

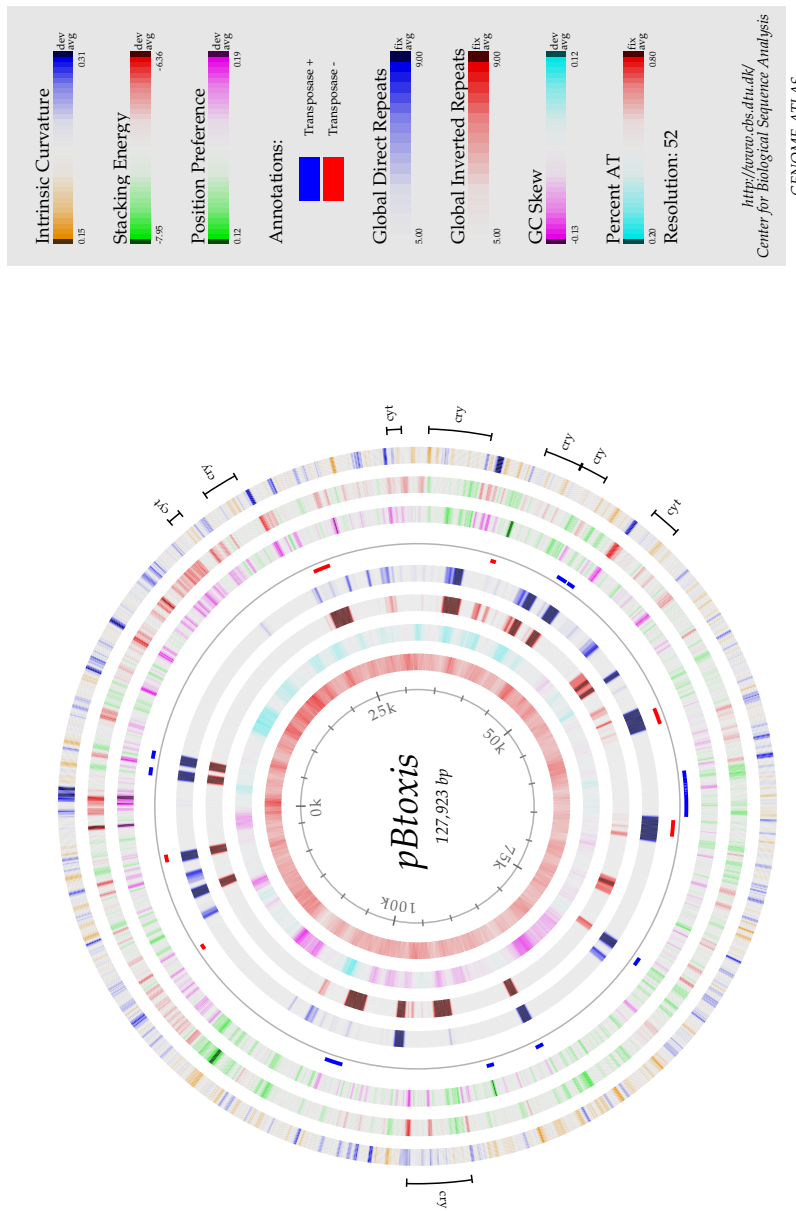


Figure 3: GenomeAtlas for *Bacillus thuringiensis* pBtoxis. The insecticide activity comes from the marked *cry* and *cyt* genes.

Leishmania major

Freidlin Chromosome 1 268,984 bp

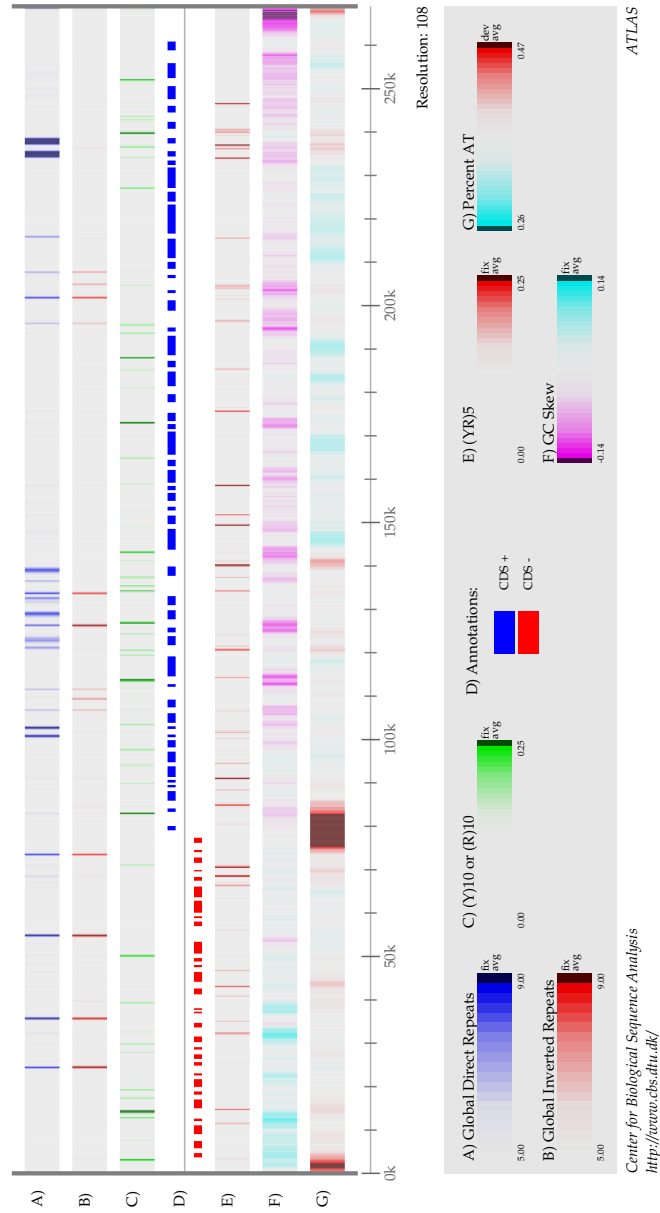


Figure 4: Specialised Atlas for *Leishmania major* chromosome 1.

Saccharomyces cerevisiae

Chromosome VIII 562,639 bp total

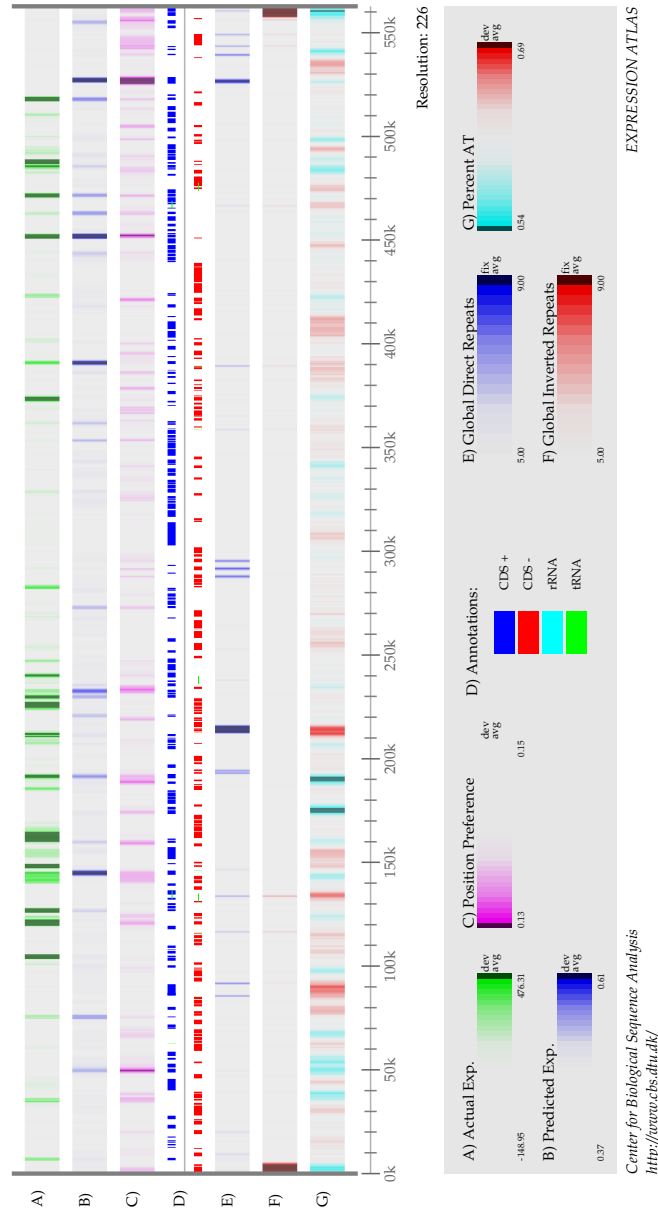


Figure 5: The ExpressionAtlas of *S. cerevisiae* chromosome VIII. Lane A shows the average intensities from cDNA arrays, indicating the constitutively expressed genes, whilst lane B is the predicted expression.

S. cerevisiae VIII

562,638 bp

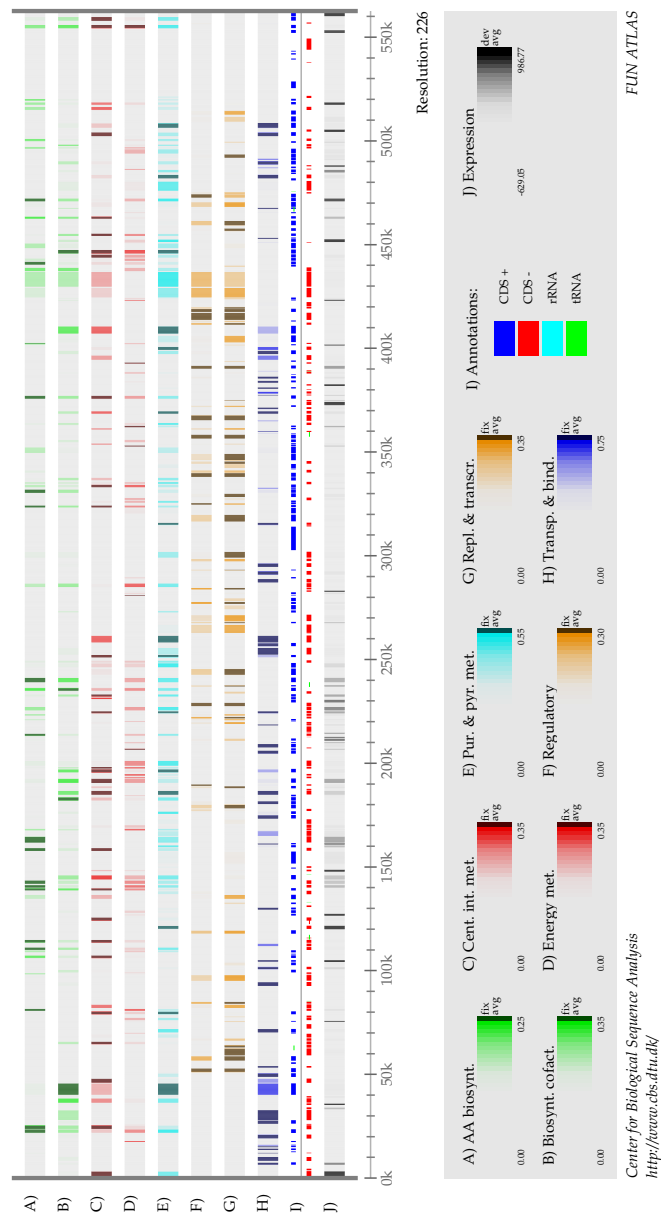


Figure 6: The FunctionAtlas of *S. cerevisiae* chromosome VIII. Lane A: Amino acid biosynthesis, Lane B: Biosynthesis of co-factors, Lane C: Central intermediary metabolism, Lane D: Energy metabolism, Lane E: Purine and pyrimidine metabolism, Lane F: Regulatory function, Lane G: Replication and transcription, Lane H: Transport and binding, Lane J: Average intensity from cDNA experiments, (the same as in Figure 5).