Methods to Compare Microbial Genomes

Part 1: Introduction to DNA Atlases: Structure, Repeat, and Genome atlases

Dave Ussery
Comparative Microbial Genomics Workshop
BIOTEC building
Pathumthani, Thailand
Module 5, talk 1
3 June, 2008
Outline

• Structural Atlases
• DNA repeats and the "Repeat Atlas"
• The best of 3 levels: "Genome Atlas"
Ecoli_K-12_W3110_Main: Structural Profile

Distance from translation start

- AT content
- Position Preference
- Stacking Energy
- Intrinsic Curvature
- DNAase sensitivity
Comparative Microbial Genomics group
Center for Biological Sequence analysis
Department of Systems Biology, Technical University of Denmark

Promoter Structural profile

DNA curvature, flexibility important here

CDS

Cruciform

mRNA

-35 -10 +1

Rigid melts

β / β'}

α
2.10 "Refined" Junction Model

2.11 5' 3' 5' 3'
(B) Chromatin in form of “beads on a string”

Average length of DNA fragment ~200 bp per nucleosome

Light nuclease digestion
Comparative Microbial Genomics group

Center for Biological Sequence analysis

Department of Systems Biology, Technical University of Denmark

Trimmed

---

**Light nuclease digestion**

![Diagram showing light nuclease digestion with fragments of ~200 bp per nucleosome.]

---

**Heavier nuclease digestion**

![Diagram showing heavier nuclease digestion with fragments of ~145 bp per core particle.]

---

Average length of DNA fragment ~200 bp per nucleosome

Average length of core DNA ~145 bp per core particle
### "Travers" trinucleotide scale:

<table>
<thead>
<tr>
<th>Trinucleotide pair</th>
<th>% Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAT/ATT</td>
<td>-0.280</td>
</tr>
<tr>
<td>AAA/TTT</td>
<td>-0.274</td>
</tr>
<tr>
<td>CCA/TGG</td>
<td>-0.246</td>
</tr>
<tr>
<td>AAC/GTT</td>
<td>-0.205</td>
</tr>
<tr>
<td>ACT/AGT</td>
<td>-0.183</td>
</tr>
<tr>
<td>CCG/CGG</td>
<td>-0.136</td>
</tr>
<tr>
<td>ATC/GAT</td>
<td>-0.110</td>
</tr>
<tr>
<td>AAG/CTT</td>
<td>-0.081</td>
</tr>
<tr>
<td>CGC/GCG</td>
<td>-0.077</td>
</tr>
<tr>
<td>AGG/CCT</td>
<td>-0.057</td>
</tr>
<tr>
<td>GAA/TTC</td>
<td>-0.037</td>
</tr>
<tr>
<td>ACG/CGT</td>
<td>-0.033</td>
</tr>
<tr>
<td>ACC/GGT</td>
<td>-0.032</td>
</tr>
<tr>
<td>GAC/GTC</td>
<td>-0.013</td>
</tr>
<tr>
<td>CCC/GGG</td>
<td>-0.012</td>
</tr>
<tr>
<td>ACA/TGT</td>
<td>-0.006</td>
</tr>
<tr>
<td>CGA/TCG</td>
<td>-0.003</td>
</tr>
<tr>
<td>GGA/TCC</td>
<td>0.013</td>
</tr>
<tr>
<td>CAA/TTG</td>
<td>0.015</td>
</tr>
<tr>
<td>AGC/GCT</td>
<td>0.017</td>
</tr>
<tr>
<td>GTA/TAC</td>
<td>0.025</td>
</tr>
<tr>
<td>AGA/TCT</td>
<td>0.027</td>
</tr>
<tr>
<td>CTC/GAG</td>
<td>0.031</td>
</tr>
<tr>
<td>CAC/GTG</td>
<td>0.040</td>
</tr>
<tr>
<td>TAA/TTA</td>
<td>0.068</td>
</tr>
<tr>
<td>GCA/TGC</td>
<td>0.076</td>
</tr>
<tr>
<td>CTA/TAG</td>
<td>0.090</td>
</tr>
<tr>
<td>GCC/GGC</td>
<td>0.107</td>
</tr>
<tr>
<td>ATG/CAT</td>
<td>0.134</td>
</tr>
<tr>
<td>CAG/CTG</td>
<td>0.175</td>
</tr>
<tr>
<td>ATA/TAT</td>
<td>0.182</td>
</tr>
<tr>
<td>TCA/TGA</td>
<td>0.194</td>
</tr>
</tbody>
</table>


NOTE: we use a (slight) modification, in which the absolute value (magnitude) of the values is used to reflect trinucleotides which tend to exclude nucleosomes.

Eukaryotic genomes do not exist in vivo as naked DNA, but in complexes known as chromatin. Chromatin contains nucleosomes, short stretches of DNA tightly wrapped around a histone protein core, which exclude most DNA binding proteins and so act as repressors. A combined computational and experimental approach has been used to determine DNA sequence preferences of nucleosomes and to predict genome-wide nucleosome organization. The yeast genome encodes an intrinsic nucleosome organization that explains about half of the in vivo nucleosome positions. Highly conserved across eukaryotes, the code directs transcription factors to their binding sites and facilitates many other specific chromosome functions. An accompanying News and Views piece discusses the role of DNA sequence and other regulators in nucleosome positioning. The cover graphic represents a stretch of chromatin including several nucleosomes.

**NEWS AND VIEWS**

**Genomics: Predictable packaging**

Nuclear factors must access specific sites within genomic DNA to function, yet the DNA is bundled up into many nucleosomes. Is the DNA sequence sufficiently informative to predict where each nucleosome will be?

Timothy J. Richmond
doi:10.1038/442750a

**FULL TEXT | PDF (294K)**

**ARTICLE**

**A genomic code for nucleosome positioning**

Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thåström, Yair Field, Irene K. Moore, Ji-Ping Z. Wang and Jonathan Widom
doi:10.1038/nature04979

**Abstract | FULL TEXT | PDF (952K) | Supplementary information**
A genomic code for nucleosome positioning

Eran Segal¹, Yvonne Fondufe-Mittendorf², Lingyi Chen², AnnChristine Thåström², Yair Field¹, Irene K. Moore², Ji-Ping Z. Wang³ & Jonathan Widom²

Eukaryotic genomes are packaged into nucleosome particles that occlude the DNA from interacting with most DNA binding proteins. Nucleosomes have higher affinity for particular DNA sequences, reflecting the ability of the sequence to bend sharply, as required by the nucleosome structure. However, it is not known whether these sequence preferences have a significant influence on nucleosome position in vivo, and thus regulate the access of other proteins to DNA. Here we isolated nucleosome-bound sequences at high resolution from yeast and used these sequences in a new computational approach to construct and validate experimentally a nucleosome–DNA interaction model, and to predict the genome-wide organization of nucleosomes. Our results demonstrate that genomes encode an intrinsic nucleosome organization and that this intrinsic organization can explain ~50% of the in vivo nucleosome positions. This nucleosome positioning code may facilitate specific chromosome functions including transcription factor binding, transcription initiation, and even remodelling of the nucleosomes themselves.
Scientists Say They’ve Found a Code Beyond Genetics in DNA

By NICHOLAS WADE

Researchers believe they have found a second code in DNA in addition to the genetic code.

The genetic code specifies all the proteins that a cell makes. The second code, superimposed on the first, sets the placement of the nucleosomes, miniature protein spools around which the DNA is looped. The spools both protect and control access to the DNA itself.

The discovery, if confirmed, could open new insights into the higher order control of the genes, like the critical but still mysterious process by which each type of human cell is allowed to activate the genes it needs but cannot access the genes used by other types of cell.

The new code is described in the current issue of Nature by Eran Segal of the Weizmann Institute in Israel and Jonathan Widom of Northwestern University in Illinois and their colleagues.

There are about 30 million nucleosomes in each human cell. So many are needed because the DNA strand wraps around each one only 1.65 times, in a twist containing 147 of its units, and the DNA molecule in a single chromosome can be up to 225 million units in length.

Biologists have suspected for years that some positions on the DNA, notably those where it bends most easily, might be more favorable for nucleosomes than others, but no overall pattern was apparent. Drs. Segal and Widom analyzed the sequence at some 200 sites in the yeast genome where nucleosomes are known to bind, and discovered that there is indeed a hidden pattern.

Knowing the pattern, they were able to predict the placement of about 50 percent of the nucleosomes in other organisms.

The pattern is a combination of sequences that makes it easier for the DNA to bend itself and wrap tightly around a nucleosome. But the pattern requires only some of the sequences to be present in each nucleosome binding site, so it is not obvious. The looseness of its requirements is presumably the reason it does not conflict with the genetic code, which also has a little bit of redundancy or wiggle room built into it.

Having the sequence of units in DNA determine the placement of nucleosomes would explain a puzzling feature of transcription factors, the proteins that activate genes. The transcription factors recognize short sequences of DNA, about six to eight units in length, which lie just in front of the gene to be transcribed. If the nucleosome code was “a profound insight if true,” because it would explain many aspects of how the DNA is controlled.
Schizosaccharomyces pombe

All Three Chromosomes  11,896,623 bp total

A) Annotations:
- CDS +
- CDS -
- rRNA
- tRNA

I - B) Percent AT

D) Watson Repeats

Resolution: 1066

Schizosaccharomyces pombe

strain 972, chromosome I  5,570,797 bp
**Schizosaccharomyces pombe**

strain 972, chromosome 1  
5,570,797 bp
Intrinsic Curvature
dev: 0.11
avg: 0.23

Stacking Energy
dev: -9.45
avg: -7.66

Position Preference
dev: 0.13
avg: 0.17

Annotations:
- CDS +
- CDS -
- rRNA
- tRNA

DNase I Sensitivity
dev: -0.04
avg: -0.00

Propeller Twist
dev: -13.35
avg: -11.24

Protein Deformability
dev: 4.64
avg: 6.00

A+T Content
dev: 0.27
avg: 0.59

Resolution: 88

STRUCTURE ATLAS
Direct repeats form slipped strand structures

Inverted repeats form cruciforms
Identifying global repeats

Step 1: Take 100 nt and find best match along entire chromosome

Step 2: Place value for best match in position 51 (value 0-9), e.g., '6'

Step 3: Position window at +1 and repeat steps 1 and 2, new value e.g., '7'

By repeating the procedure, the chromosome receives values for each position but the first and last 50 nucleotides which receive '0'

Identifying local repeats

Step 1: Take 100 nt and find best match of up to 15 bp repeat within the window

Step 2: Place value for best match in position 15 (value 0-9), e.g., '3'

Step 3: Position window at +1 and repeat steps 1 and 2, new value e.g., '4'

By repeating the procedure, the chromosome receives values for each position but the first and last 50 nucleotides which receive '0'
Comparative Microbial Genomics group

- LOCAL DIRECT REPEATS
- LOCAL INVERTED REPEATS
- LOCAL MIRROR REPEATS
- LOCAL DIRECT REPEATS
Comparative Microbial Genomics group

Center for Biological Sequence analysis

Department of Systems Biology, Technical University of Denmark

0.2 0.4 0.6 0.8

10 20 30 40

% of local repeats in chromosome

AT content (fraction)

Observed in chromosomes
predicted for random DNA

- Observed in chromosomes
- Predicted for random DNA
Introducing the "Genome Atlas"
Comparative Microbial Genomics group

Center for Biological Sequence analysis
Department of Systems Biology, Technical University of Denmark

$P.\ thermopropionicum$ strain SI
3,025,375 bp

Annotations:
- CDS +
- CDS -
- rRNA
- tRNA

Intrinsic Curvature
- dev
- avg
0.14
0.23

Stacking Energy
- dev
- avg
-8.79
-7.46

Position Preference
- dev
- avg
0.14
0.16

Annotations:
- CDS +
- CDS -
- rRNA
- tRNA

Global Direct Repeats
- fix
- avg
5.00
7.50

Global Inverted Repeats
- fix
- avg
5.00
7.50

GC Skew
- fix
- avg
-0.08
0.08

Percent AT
- dev
- avg
0.37
0.57

Resolution: 1211

GENOME ATLAS
Figure 7. Structural cluster analysis. Distance tree showing the relative location of 11 gene clusters based on average structural measures. The number of genes in
Histogram of log(affy.ex[, i])

Highly Expressed Genes