Statistical mechanics of nucleosome positioning

Alexandre V. Morozov
Department of Physics & Astronomy and the BioMaPS Institute for Quantitative Biology, Rutgers University
morozov@physics.rutgers.edu

RosettaCON, August 2009
Chromatin length scales

Felsenfeld & Groudine, Nature 2003
Chromatin under the microscope

Electron micrograph of *D. Melanogaster* chromatin: arrays of regularly spaced nucleosomes, each ~80 Å across.

http://www.accessexcellence.org/AB/GG/chromosome.html
Biophysical picture of gene transcription

A

promoter (cis-regulatory region) | transcription unit

module (enhancer) | basal (core) promoter | exon | intron | UTR

B

chromatin remodeling complex

transcription co-factors

transcription factors

TAFs

pol II holoenzyme

transcription start site

TATA box

TATA-binding protein

looping factors

chromatin


Molecular Biology and Evolution
Nucleosome core particle

Left-handed superhelix: (1.84 turns, 147 bp, R = 41.9 Å, P = 25.9 Å)
PDB code: 1kx5

What factors determine nucleosome positioning on genomic DNA?
DNA elastic energy explains histone-DNA interactions

- Nucleosome affinity depends on the presence and spacing of key dinucleotide motifs (e.g. TA, CA)
- Nucleosome affinity can be explained by sequence-dependent DNA flexibility
Adding key dinucleotide motifs increases nucleosome affinity
Deleting dinucleotide motifs or disrupting their spacing decreases affinity

Nucleosome formation is sequence-dependent

Segal et al., Nature 2006
Nucleosomes interfere with DNA function
Nucleosomes form an approx. 1D liquid of particles of size L=147 bp (~80% of yeast genome is nucleosome-covered) – regular arrays are created simply by steric exclusion
Nucleosomes affect transcription factor binding to DNA

Thermodynamic competition for binding sites:

**Nucleosome**

**Transcription factors**
Nucleosome occupancy is affected by other DNA-binding factors

Nucleosome-free site

Nucleosome-occluded site

Nucleosome is displaced through TF binding
Figure 1. One potential scenario by which MeCP2 and the SWI/SNF complex might cooperate in establishment or maintenance of transcriptional repression at a methylated locus.

Inferring nucleosome positioning and energetics from high-throughput *in vitro* datasets

Kevin Struhl, Harvard
From nucleosome energies to probabilities and occupancies: 1D liquid of core particles

Use a **recursive algorithm** to find the grand canonical partition function and thus nucleosome probabilities and occupancies:

\[ P_i = \frac{Z_{i-1} e^{-E(i)/Z_{i+L}}}{Z} \]

---

Morozov et al., Nucl. Acids Res. 2009

---

**Nucleosome energy**

**Nucleosome Probability & Occupancy**
Sequence read profiles for nucleosomes positioned *in vitro* by salt dialysis

*S. cerevisiae*

Solexa/Illumina sequencer

Collaboration with Kevin Struhl, Harvard
Translational and rotational positioning of \textit{in vitro} nucleosomes

Auto-correlation function of sequence read profiles
Infer nucleosome energies directly from the occupancy profiles:

\[ E_i - \mu = \log \frac{P_i + P^L_i}{P_i} + \sum_{j=i}^{i+L-1} \log \frac{P^L_j}{P_j + P^L_j} \]

Explain the energies in terms of sequence features through a linear model fit to the energies of individual words:

\[ E_i - \mu = \sum_{j=\text{words}} n^j_{\text{words}} \epsilon^j_{\text{words}} \]
Fitting nucleosome energies to sequence features

DNA sequence

TTAAGG.................CTCAGC

E_i → GC
    → AAAAAA

T → ...
Nucleosomes are not well-positioned with respect to TSS \textit{in vitro} ...

Mavrich \textit{et al}, Genome Res. 2008

Zhang \textit{et al}, to appear in NSMB
Nonetheless we can extract sequence-specific positioning signals
Table of correlation coefficients between various experimental & theoretical occupancy profiles:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1.0000</td>
<td>0.6542</td>
<td>0.4446</td>
<td>0.5903</td>
<td>0.2558</td>
<td>0.3607</td>
<td>0.4522</td>
<td>0.5993</td>
<td>0.3771</td>
</tr>
<tr>
<td>B</td>
<td>0.6542</td>
<td>1.0000</td>
<td>0.5705</td>
<td>0.7695</td>
<td>0.2400</td>
<td>0.4781</td>
<td>0.5801</td>
<td>0.7437</td>
<td>0.4489</td>
</tr>
<tr>
<td>C</td>
<td>0.4446</td>
<td>0.5705</td>
<td>1.0000</td>
<td>0.5354</td>
<td>0.1866</td>
<td>0.7879</td>
<td>0.9914</td>
<td>0.5364</td>
<td>0.6045</td>
</tr>
<tr>
<td>D</td>
<td>0.5903</td>
<td>0.7695</td>
<td>0.5354</td>
<td>1.0000</td>
<td>0.2681</td>
<td>0.3827</td>
<td>0.5614</td>
<td>0.8826</td>
<td>0.4972</td>
</tr>
<tr>
<td>E</td>
<td>0.2558</td>
<td>0.2400</td>
<td>0.1866</td>
<td>0.2681</td>
<td>1.0000</td>
<td>0.1817</td>
<td>0.1937</td>
<td>0.3141</td>
<td>0.2022</td>
</tr>
<tr>
<td>F</td>
<td>0.3607</td>
<td>0.4781</td>
<td>0.7879</td>
<td>0.3827</td>
<td>0.1817</td>
<td>1.0000</td>
<td>0.7259</td>
<td>0.4339</td>
<td>0.5073</td>
</tr>
<tr>
<td>G</td>
<td>0.4522</td>
<td>0.5801</td>
<td>0.9914</td>
<td>0.5614</td>
<td>0.1937</td>
<td>0.7259</td>
<td>1.0000</td>
<td>0.5532</td>
<td>0.6182</td>
</tr>
<tr>
<td>H</td>
<td>0.5993</td>
<td>0.7437</td>
<td>0.5364</td>
<td>0.8826</td>
<td>0.3141</td>
<td>0.4339</td>
<td>0.5532</td>
<td>1.0000</td>
<td>0.5004</td>
</tr>
<tr>
<td>I</td>
<td>0.3771</td>
<td>0.4489</td>
<td>0.6045</td>
<td>0.4972</td>
<td>0.2022</td>
<td>0.5073</td>
<td>0.6182</td>
<td>0.5004</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

- A = Struhl's *in vitro* data
- B = Widom's *in vitro* data
- C = Widom's *in vivo* data (combined YPD)
- D = Segal's model
- E = Struhl's MNase data
- F = Widom's *in vivo* data (CL, YPD)
- G = Widom's *in vivo* data (no CL, YPD)
- H = Percus energy model
- I = Widom's 454 *in vivo* data (YPD)
Sequence determinants of nucleosome positioning and energetics

Dinucleotide frequencies for *in vitro* nucleosomes

<table>
<thead>
<tr>
<th>Words</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>20</td>
</tr>
<tr>
<td>AT</td>
<td>40</td>
</tr>
<tr>
<td>TA</td>
<td>60</td>
</tr>
<tr>
<td>TT</td>
<td>80</td>
</tr>
<tr>
<td>AC</td>
<td>100</td>
</tr>
<tr>
<td>AG</td>
<td>120</td>
</tr>
<tr>
<td>CA</td>
<td>140</td>
</tr>
<tr>
<td>CC</td>
<td>160</td>
</tr>
<tr>
<td>CG</td>
<td>180</td>
</tr>
<tr>
<td>CT</td>
<td>20</td>
</tr>
<tr>
<td>GA</td>
<td>40</td>
</tr>
<tr>
<td>GC</td>
<td>60</td>
</tr>
<tr>
<td>GG</td>
<td>80</td>
</tr>
<tr>
<td>GT</td>
<td>100</td>
</tr>
<tr>
<td>TC</td>
<td>120</td>
</tr>
<tr>
<td>TG</td>
<td>140</td>
</tr>
</tbody>
</table>

$\times 10^{-7}$
Lab members:

- Dr. Denis Tolkunov
- Allan Haldane
- George Locke
- Julia Tsitron

Funding:

- NIH R01 HG004708
- Alfred P. Sloan Fellowship
- BioMaPS Institute for Quantitative Biology
- National Institutes of Health
- Alfred P. Sloan Foundation
- BioMaPS
- James R. Broach
  Karl Zawadzki
  (Princeton)
- Kevin Struhl
  (Harvard)
- Eric Siggia
  (Rockefeller)
- Jon Widom
  (Northwestern)
Inter-nucleosomal interactions and "quantized" linker lengths