The Role of DNA Sequence and Accessibility in Transcriptional Regulation

Tommy Kaplan
School of Computer Science and Engineering
The Hebrew University
The human body is built from many cells

- The body is composed of \(10^{14}\) cells
- Hundreds of cell types
- All originate from one cell
- Contain the same exact DNA sequence
Transcriptional regulation

• Regulatory regions control when and where genes are activated

• 1.5% of genome encodes proteins
• 3.5%-40% are regulatory

Understanding regulation is key to understanding our genome
Quantitative models of genome-wide binding during *Drosophila* development

Flies are better than humans

- Drosophila embryos as a model of gene regulation
- Easier (genetics, genomics)
Early development in fruit flies is transcriptionally regulated.

- 13 rapid divisions
- Gradients of transcription factors
- Segmentation & differentiation

Regulation via differential binding of transcription factors.
Genomic data on five core regulators

<table>
<thead>
<tr>
<th>Regulators</th>
<th>Protein-DNA specificities</th>
<th>Protein concentrations (in 3D)</th>
<th>Genome-wide maps of in vivo protein binding (from whole embryos)</th>
</tr>
</thead>
<tbody>
<tr>
<td>bicoid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>caudal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>giant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hunchback</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krüppel</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What mechanisms shape the binding landscape?
A (very simplistic) model of binding

Goal: Quantify each aspect!
Quantify importance via modular platform to predict binding

Cooperativity

Competition

Computational Model

Accessibility

BCD
Predicting binding from sequence using gHMM

- Modular
- Exact inference in linear time
- Probabilistic approach - binding as a continuum

Rabiner, 1989
Predicting binding from sequence using gHMM

• Modular
• Exact inference in linear time
• Probabilistic approach - binding as a continuum

Compare predictions to data: reverse peak fitting approach

Measure fit (e.g. using correlation)

Pr(binding)

Genome

Expected landscape

Measured

Capaldi, Kaplan et al, Nature Genetics, 2008
Protein-DNA specificities partially explain data

- 21 known early developmental genes
  - Train set (187 Kb: 6 genes + 3 genomic control regions)
  - Test set (501 Kb: 15 genes + 5 genomic control regions)
Protein-DNA specificities partially explain data

- 21 known early developmental genes
  - Train set (187 Kb: 6 genes + 3 genomic control regions)
  - Test set (501 Kb: 15 genes + 5 genomic control regions)
Modeling binding competition

• Many regulators show overlapping sites
• Incorporated all 5 factors into gHMM
Modeling binding competition cell-by-cell

Anterior

Posterior

Center

Whole embryo (averaged)
Competition with nucleosomes

DNA wrapped around nucleosomes

Prediction accuracy

<table>
<thead>
<tr>
<th>Test corr.</th>
<th>Sequence</th>
<th>+competition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.36</td>
<td></td>
<td>0.37</td>
</tr>
</tbody>
</table>

GT
CAD
BCD
HB
KR
nucleosome
(length = 141)
no binding
Model cooperative interactions using a full thermodynamic model

- Add cooperative interaction $\Delta \Delta G$ between adjacent sites
- Sample configurations from posterior
- Reweight by $\Delta \Delta G$

<table>
<thead>
<tr>
<th>Test corr.</th>
<th>Sequence</th>
<th>+competition</th>
<th>+cooperativity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.36</td>
<td>0.37</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Where do we go wrong?

With J. Stamatoyannopoulos, U. of Washington
Most false positives are inaccessible in vivo

With J. Stamatoyannopoulos, U. of Washington
Accessible regions = Higher prior probability of binding
DNA accessibility boosts accuracy

- Integrate DNA accessibility into model

Test loci (6/15) spanning 108Kb
DNA accessibility data boosts correlation between predictions and measured binding.

Test correlation:

- Sequence: 0.36
- +competition: 0.37
- +cooperativity: 0.40
- +accessibility: 0.67
Interim summary:

What determines where proteins bind?

- \( \frac{1}{3} \) sequence (=match to TF binding specificities)
- \( \frac{1}{3} \) DNA accessibility
- \( \frac{1}{3} \) other aspects
Zelda, an embryonic pioneer
- or -
How to kick-start a genome?

Harrison, Li, Kaplan et al, PLoS Genetics, 2011
The beginning of a life

• Fertilization is followed by 13 cycles of super-fast nuclear divisions
The same thing happened to me too!

1 cell

cycle 8

1 cell

cycle 14 (2-3 hours)

Drosophila melanogaster

1 cell

cycle 8

1 cell

cycle 12 (7-8 hours)

Xenopus laevis

Oocyte

1 cell

2 cell (24 hours)

Homo sapiens

maternal mRNA

zygotic transcription

maternal-to-zygotic transition (MZT)
CAGGTAG
a bioinformatician mystery

- CAGGTAG motif enriched near Sex-lethal (Sxl) and the first genes to be expressed in Drosophila embryos
- Why? How?
- Mutating a single CAGGTAG motif (near sc, sxl, or zen) is enough to postpone zygotic transcription and development
- Adding more CAGGTAGs advanced the onset of transcription

ten Bosch et al, *Development*, 2006
ten Bosch et al, *Development*, 2006
The protein Zelda (ZLD)

- CAGGTAG bound by C2H2 zinc-finger protein
- 1596 amino acid
- \textit{zld} mutants show severe developmental defects
- Expression patterns lack genes essential for cellularization, sex determination, and patterning.


Zelda binding during Maternal-to-Zygotic transition

- ChIP of early embryos is tricky

<table>
<thead>
<tr>
<th>Cycle</th>
<th>Nuclei</th>
<th>Fraction of Embryos</th>
<th>Fraction of Nuclei</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>128</td>
<td>10%</td>
<td>1%</td>
</tr>
<tr>
<td>9</td>
<td>256</td>
<td>60%</td>
<td>17%</td>
</tr>
<tr>
<td>10</td>
<td>512</td>
<td>10%</td>
<td>6%</td>
</tr>
<tr>
<td>11</td>
<td>1024</td>
<td>5%</td>
<td>6%</td>
</tr>
<tr>
<td>12</td>
<td>2048</td>
<td>5%</td>
<td>12%</td>
</tr>
<tr>
<td>13</td>
<td>4096</td>
<td>5%</td>
<td>23%</td>
</tr>
<tr>
<td>14</td>
<td>6144</td>
<td>5%</td>
<td>35%</td>
</tr>
</tbody>
</table>

cycles 8-10 early cycle 14 late cycle 14
Zelda binding during Maternal-to-Zygotic transition

- Hand-sorting embryos for stage-specific Zelda ChIP

Xiao-Yong Li
Zelda binds thousands of genes prior to genome activation

**scute**

**even skipped**
Genes with strong ZLD binding to promoters are down-regulated in zld mutants

Zelda binding is indicative of transcription

Zygotic expression

Promoter binding

Zygotic transcription from Lott et al, PLoS Biology, 2011
Most TF-bound regions were already bound by ZLD prior to MZT
Binding sites in genome poorly predict binding

except in the presence of Zelda
ZLD binding at mitotic cycle 8 predicts DNA accessibility at cycle 14
Binding sites in genome poorly predict binding

except in the presence of for Zelda
Zelda - a pioneer factor

- Zelda binds CAGGTag sites, prior to genome activation
- Maintains high accessibility
- Early transcribed genes = early and strong ZLD binding
- Most regulatory regions = bound by ZLD (before TFs)
- Other Zelda-like co-factors in early development?
Acknowledgements
CS/LS students wanted!

Epigenomics and Chromatin in Transcriptional Regulation and Human Diseases

Research Group in Computational Biology
tommy@cs.huji.ac.il