Reverse engineering mammalian transcriptional regulatory circuits

Andrew D Smith    Pavel Sumazin

Zhang Lab, CSHL & Califano Lab, Columbia

ISMB 2007
Outline of part I

**Introduction**
- Background on regulatory networks
- Data available for analysis

**Analysis methods**
- Identifying gene modules
- Modeling regulatory elements
- Predicting binding sites
- Conservation of regulatory elements
- Motif discovery
- Cis-regulatory modules
Outline of part II

Analyzing sets of co-regulated genes
  An example gene module
  Identifying enriched known motifs
  Predicting functional binding sites

Analysis of transcription factor Localization data
  ChIP-chip data examples
  Identifying enriched known motifs
  Identifying co-factors
  Discovering motifs *de novo*
Part I

Part I: Lecture Format
Introduction

Background on regulatory networks
Data available for analysis

Analysis methods

Identifying gene modules
Modeling regulatory elements
Predicting binding sites
Conservation of regulatory elements
Motif discovery
Cis-regulatory modules
Introduction

Background on regulatory networks
Data available for analysis

Analysis methods
  Identifying gene modules
  Modeling regulatory elements
  Predicting binding sites
  Conservation of regulatory elements
  Motif discovery
  Cis-regulatory modules
Introduction

Background on regulatory networks

Data available for analysis

Analysis methods

Identifying gene modules
Modeling regulatory elements
Predicting binding sites
Conservation of regulatory elements
Motif discovery
Cis-regulatory modules
Assumed background

- Genes
- Promoters
- Transcription factors (TFs)
- Transcription factor binding sites

- Enhancers
- *cis*-Regulatory modules
- Gene expression microarrays
- ChIP-chip TF localization

(Levine & Tjian, 2003)
The goal: networks

- Identifying regulatory relationships between genes
- Understanding the underlying sequence-based mechanisms
- Deriving specific hypotheses about transcription
Components of regulatory networks

Nodes correspond to genes

- Networks include both regulators and targets
- Edges are regulatory relationships
- Most genes are only targets
- Interesting subnetworks composed of regulators
Components of regulatory networks

Nodes correspond to genes

- Networks include both regulators and targets
- Edges are regulatory relationships
- Most genes are only targets
- Interesting subnetworks composed of regulators
Introduction

Background on regulatory networks

Components of regulatory networks

Nodes correspond to genes

- Networks include both regulators and targets
- Edges are regulatory relationships
- Most genes are only targets
- Interesting subnetworks composed of regulators
Two kinds of regulatory networks

Direct networks
- Edges: physical interaction
- Interaction is specified in regulatory sequence of target
Two kinds of regulatory networks

Direct networks
- Edges: physical interaction
- Interaction is specified in regulatory sequence of target

Influence networks
- Edges: possibly indirect interaction
- Interaction may be mediated by another gene
Examples of what we can achieve

Understanding regulators

- Which TFs are most important in a given context?
- Do certain regulators appear to work together?
- Possibly infer novel regulators or functions
- Clues about regulatory mechanisms (e.g. TF binding specificity)

Understanding targets

- The set of TFs that appear to regulate some gene
- The condition-specific targets of particular TFs
- Sequence features are important to a gene’s transcription
- Do targets appear under control of the same set of regulators?
Introduction

Background on regulatory networks

Data available for analysis

Analysis methods

Identifying gene modules
Modeling regulatory elements
Predicting binding sites
Conservation of regulatory elements
Motif discovery
Cis-regulatory modules
Gene expression data

Sets of interesting genes

- Function in the context being examined
- Sets can be assembled gene-by-gene: very slow, but produces high-quality data

Microarray data

- Lots of expression data fast and easy
- Lower quality than sets of genes collected manually
- Ultimate test of understanding: Can we reliably predict high-throughput expression data?
**TF binding data**

**TF binding behavior**

- Several ways to examine binding (individually or high-throughput)
- Locations of binding sites tell much about TF function
- Not all sites that bind are involved in regulation

**ChIP-chip**

- Context-specific binding-sites genome wide and *in vivo*
- Familiar tradeoff: much more data, possibly low-quality
- ChIP-seq: emerging technology

Wang, Snyder & Gerstein (2007)
Genomic sequence and annotations

Raw sequence data

- High-quality genomes available
- Available from various sources: UCSC, NCBI, ENSEMBL

Genome alignments

- Describe cross-species conservation
  (important for sequence analysis in any context)
- Pre-computed alignments: easy to use, improve constantly

Genome annotations

- Locations of important genomic features
- Examples for transcription: TSS, CDS and repeat locations
- Increasing amount of annotation directly related to transcription
Other data about transcription

Databases of existing knowledge

- PUBMED: first place to check (know what’s already known)
- Databases about transcription (e.g. TRANSFAC, SCPD)
- Useful databases of characterized networks (hopefully soon)

Chromatin structure data

- Important for transcription, but less well understood
- Chromatin structure is an important regulatory mechanism
- Different modifications affect transcription differently
- Some modifications render genes “poised” for transcription
- Other modifications prevent transcription
- Emerging technology: Chromatin Capture (3C and 5C)
Introduction
  Background on regulatory networks
  Data available for analysis

Analysis methods
  Identifying gene modules
  Modeling regulatory elements
  Predicting binding sites
  Conservation of regulatory elements
  Motif discovery
  Cis-regulatory modules

Smith & Sumazin (CSHL & Columbia)  Transcriptional regulatory circuits  ISMB’07  18 / 109
Gene modules

What is a gene module?

• Many possible definitions, but let's keep it informal
• Usually a set of genes that function together
• Think: the genes whose regulation you want to understand
• Gene modules might have 10 genes, or 500 genes
The context

- Want to understand, for example
  - Expression in diseased cells
  - Cells from a developmental state
- Get expression from 2 conditions
  - Before and after some perturbation
  - Samples taken at different time-points
  - Different types of cells

Simplest gene modules

- Genes showing differential expression
- Maybe interested only in genes “over-expressed” or “under-expressed”
- Mann-Whitney U-test
Gene expression profiles

Gene expression matrix

- Columns ⇔ experiments
- Rows ⇔ genes
- $x_{i,j} \leftrightarrow$ level of gene $i$ in expmt $j$

$$
\begin{pmatrix}
  x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,m} \\
  x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,m} \\
  x_{3,1} & x_{3,2} & x_{3,3} & \cdots & x_{3,m} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  x_{n,1} & x_{n,2} & x_{n,3} & \cdots & x_{n,m}
\end{pmatrix}
$$
Gene expression profiles

Gene expression matrix

- Columns $\leftrightarrow$ experiments
- Rows $\leftrightarrow$ genes
- $x_{i,j} \leftrightarrow$ level of gene $i$ in expmt $j$

$$
\begin{pmatrix}
  x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,m} \\
  x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,m} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  x_{n,1} & x_{n,2} & x_{n,3} & \cdots & x_{n,m}
\end{pmatrix}
$$

Gene expression profile

- Each gene has a profile: a row of the matrix
- Statistical issues (e.g. normalization) outside current scope
- More experiments means more information in each profile
- Similar expression profiles suggest similar regulation
Analysis methods
Identifying gene modules

Using data from multiple experiments

Clustering genes expression profiles

- Get gene modules based on expression from multiple experiments
- Cluster genes with *similar* or *correlated* expression profiles
- Any clustering algorithm can be used (e.g. k-means, hierarchical)
- Best algorithm depends on data and analysis goals

Measuring profile similarity

- Examples: correlation, Euclidean distance, mutual information
- Again, best measure depends on data and analysis goals
Inferring influence networks

Obtaining the direction of a relationship

- Clusters suggest association, but not causation
- More interesting: infer which are regulators and which are targets
- Need sophisticated tools and the right kind/amount of data
- Examples of methods: Bayesian networks, ARACNE

How to use influence networks

- Influence networks can provide framework
- Connections can be annotated with direct information
Introduction

- Background on regulatory networks
- Data available for analysis

Analysis methods

- Identifying gene modules
- **Modeling regulatory elements**
- Predicting binding sites
- Conservation of regulatory elements
- Motif discovery
- Cis-regulatory modules
Binding sites

- Genomic sequences recognized and bound by binding domains of TFs
- Binding sites for same TF might be different from each other
- Often 8-12bp, but examples can be found from 5bp to \( \sim 30\text{bp} \)
What is a motif?

- Motifs are how we model the set of binding sites for a TF
- Should describe information important for binding
- Motifs ≠ binding sites
Consensus sequence representation

Consensus sequences

Pros: Easy to understand, easy to manipulate computationally
Cons: Does not express all important information
Analysis methods  Modeling regulatory elements

Consensus sequence representation

Degenerate nucleotides

\[
\begin{align*}
M & \Rightarrow \text{A or C} & V & \Rightarrow \text{A, C or G} \\
R & \Rightarrow \text{A or G} & H & \Rightarrow \text{A, C or T} \\
W & \Rightarrow \text{A or T} & D & \Rightarrow \text{A, G or T} \\
S & \Rightarrow \text{C or G} & B & \Rightarrow \text{C, G or T} \\
Y & \Rightarrow \text{C or T} & N & \Rightarrow \text{A, C, G or T} \\
K & \Rightarrow \text{G or T} & & \\
\end{align*}
\]

Degenerate consensus sequences

- IUPAC degenerate nucleotide codes
- Provides more flexible representation, but usually not enough
Matrix-based representation

What is the matrix representation?

- Matrix columns correspond to positions in sites
- Matrix rows correspond to nucleotides
- Entries correspond to base counts at the site
- Assumptions: independent positions, fixed with, no gaps
## Matrix-based representation

### Counts

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>15</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

### Probabilities (normalized counts)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.06</td>
<td>0.00</td>
<td>0.71</td>
<td>0.24</td>
</tr>
<tr>
<td>2</td>
<td>0.06</td>
<td>0.94</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>0.88</td>
<td>0.00</td>
<td>0.12</td>
</tr>
<tr>
<td>4</td>
<td>0.94</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>0.06</td>
<td>0.06</td>
<td>0.29</td>
</tr>
<tr>
<td>6</td>
<td>0.12</td>
<td>0.41</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>7</td>
<td>0.06</td>
<td>0.06</td>
<td>0.29</td>
<td>0.35</td>
</tr>
<tr>
<td>8</td>
<td>0.06</td>
<td>0.12</td>
<td>0.06</td>
<td>0.18</td>
</tr>
<tr>
<td>9</td>
<td>0.12</td>
<td>0.12</td>
<td>0.35</td>
<td>0.41</td>
</tr>
</tbody>
</table>

### Different kinds of matrices

- **Probability matrix**: columns are position-specific nucleotide distributions
- **Many names**: position-weight matrix (PWM), position-frequency matrix (PFM) profile, alignment matrix, etc.
- **We use PWM to refer to both count and probability matrices**
- **Only 3 different kinds of matrices** (we will see a **scoring matrix** later)
Sequence Logos

- Cartoon depiction of a motif
- Size of base is proportional to frequency in matrix
- Sometimes sizes are scaled by “information content” (not covered)
Sequence Logos

- Cartoon depiction of a motif
- Size of base is proportional to frequency in matrix
- Sometimes sizes are scaled by “information content” (not covered)
Motif Databases

- JASPAR (free) and TRANSFAC (BIOBASE)
- Hundreds of known motifs and binding sites
- Essential resources for regulatory sequence analysis
Introduction
- Background on regulatory networks
- Data available for analysis

Analysis methods
- Identifying gene modules
- Modeling regulatory elements
- Predicting binding sites
- Conservation of regulatory elements
- Motif discovery
- Cis-regulatory modules
Probability from a motif

Possible to compute probability of a sequence from a motif
 Multiply values corresponding to nucleotide at each position
 This works because we assume positions are independent
 In the example \( \Pr(TCTATGTTT) = 0.001419188 \)
Probability from a motif

Possible to compute probability of a sequence from a motif
Multiply values corresponding to nucleotide at each position
This works because we assume positions are independent
In the example \( \Pr(TCTATGTTT) = 0.001419188 \)
... but does that mean anything?
### Likelihood from motif vs base composition

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>G</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>T</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

- Sequence: T C T A T G T T T
- Likelihood from motif: \(0.00000152\)
- Base frequencies: 0.2x0.3x0.2x0.2x0.2x0.3x0.2x0.2x0.2 = 0.00000152

- Likelihood from motif was \(\approx 0.00142\)
- Assume each position sampled independently from base frequencies
- Ratio of the likelihoods: \(0.00142/0.00000152 \approx 934\)
- Match-score: obtained by taking log of this ratio
- Positive match-score \(\Rightarrow\) sequence more likely from motif
Making a scoring matrix

\[
\log \left( \frac{\text{probability from motif}}{\text{probability from base composition}} \right) = \log \left( \frac{0.94}{0.30} \right) = 1.6
\]
### Scanning a sequence

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-1.6</td>
<td>-1.6</td>
<td>-4.2</td>
<td>2.2</td>
<td>-4.2</td>
<td>-0.7</td>
<td>-1.6</td>
<td>-1.6</td>
<td>-0.7</td>
</tr>
<tr>
<td>C</td>
<td>-4.2</td>
<td>1.6</td>
<td>1.5</td>
<td>-2.1</td>
<td>-2.1</td>
<td>0.4</td>
<td>-2.1</td>
<td>-1.2</td>
<td>-1.2</td>
</tr>
<tr>
<td>G</td>
<td>1.2</td>
<td>-4.2</td>
<td>-4.2</td>
<td>-4.2</td>
<td>-2.1</td>
<td>0.0</td>
<td>-2.1</td>
<td>-0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>T</td>
<td>0.2</td>
<td>-4.2</td>
<td>-0.7</td>
<td>-4.2</td>
<td>2.1</td>
<td>-0.2</td>
<td>2.0</td>
<td>1.6</td>
<td>1.0</td>
</tr>
</tbody>
</table>

\[
0.2 + 1.6 - 0.7 + 2.2 + 2.1 + 0.0 + 2.0 + 1.6 + 1.0 = 10
\]

**AGTATCAC**TCTATGTTTGTGGCACA

### Basic steps

- Slide matrix along sequence
- Calculate score at each position
- Keep scores that meet some criteria (e.g. above a cutoff)
Remarks

About scoring matrices

- Match-scores are sensitive to the base composition assumed
- Also sensitive to pseudocount
- Several algorithms exist for calculating scores fast
- Statistical significance of matches can be measured multiple ways
Remarks

About scoring matrices

- Match-scores are sensitive to the base composition assumed
- Also sensitive to pseudocount
- Several algorithms exist for calculating scores fast
- Statistical significance of matches can be measured multiple ways

About predicted sites

- Provide mechanistic link between regulator and target in networks
- High false positive rate: match-scores only tell part of the story
- Should be combined with cross-species conservation (more later)
What does enrichment mean?

Three desirable properties

1. More total occurrences

VS.
What does enrichment mean?

Three desirable properties

1. More total occurrences
2. Stronger occurrences (i.e. higher scoring)
What does enrichment mean?

Three desirable properties

1. More total occurrences
2. Stronger occurrences (i.e. higher scoring)
3. More sequences containing an occurrence
What does enrichment mean?

Three desirable properties

1. More total occurrences
2. Stronger occurrences (i.e. higher scoring)
3. More sequences containing an occurrence

But different assumptions valid for different TFs/contexts
Enrichment based on likelihood

- Mixture models: rigorous statistical foundation for enrichment
- These models capture the 3 aspects of enrichment: each sequence is a *mixture* of sites and non-sites
- Likelihoods calculated for entire set of sequences
- Necessary calculations closely related to match-scores
Using a set of background sequences

Which motif is more enriched?

- Yellow motif occurs many times
- Blue motif also occurs many times (and in consistent location)
- Both may appear enriched
Using a set of background sequences

Why use a background set?

- Statistical models of “random” promoters don’t work
- Using a background can control many unknown variables
- Different backgrounds can be used to examine different questions
Selecting background sequences

Examples of desirable properties

- Similar to foreground in terms of primary sequence features (e.g. GC-content, CpG-content)
- Uniform length sequences (both FG and BG) can facilitate statistics
- Share similar biological properties (e.g. compare promoters to other promoters)

Common mistakes

- Compare promoters to exons (very bad)
- Comparing CpG-related promoters to non-CpG-related promoters
- Having different repeat composition in background
- Comparing sequences between species
- Using too few sequences (results in over-fitting)
Identifying enriched motifs

Why identify enriched motifs?

- Identify motifs that are important regulators of a gene module
- Obtain more information for connections in networks
- Identify candidates for site prediction

Significance of motif enrichment

- Enrichment scores more useful if $p$-values can be obtained
- Empirical $p$-values can be obtained in multiple ways:
  - shuffle sequences, permute sequence labels, permute matrix columns
- Correct for multiple testing if evaluating enrichment of multiple motifs
Introduction

Background on regulatory networks
Data available for analysis

Analysis methods

Identifying gene modules
Modeling regulatory elements
Predicting binding sites

Conservation of regulatory elements
Motif discovery
Cis-regulatory modules
Cross-species conservation

Why do we use it?

- Negative selection: things that are important will be conserved
- Helps distinguish functional from non-functional sites
How to use conservation

Conserved regions

- Search in pre-defined regions
- *e.g.* Ultraconserved regions
How to use conservation

Conserved regions

- Search in pre-defined regions
- *e.g.* Ultraconserved regions

Conservation profile

- Assign conservation score to each individual base
- *e.g.* phastCons scores
How to use conservation

Conserved regions
- Search in pre-defined regions
- \textit{e.g.} Ultraconserved regions

Conservation profile
- Assign conservation score to each individual base
- \textit{e.g.} phastCons scores

Use alignments directly
- Much information in alignments
- Requires more complex methods
Turnover and non-alignment methods

- Functionally analogous sites (in different species) that do not align
- Sites presumed to evolve under similar evolutionary constraints
- Importance of turnover still not clear, but some evidence exists
- Non-alignment methods in general less useful for predicting sites
- Can indicate important motifs (cross-species enrichment)
Things to consider

Which alignments to use

- Precomputed alignments: multiz17way (recently 28-way), mlagan
- Creating your own alignments raises many issues

Species to use

- Understand the network being investigated
- Make sure protein and function conserved in species compared
- Accounts for compensatory substitutions in sites
Introduction
  Background on regulatory networks
  Data available for analysis

Analysis methods
  Identifying gene modules
  Modeling regulatory elements
  Predicting binding sites
  Conservation of regulatory elements

Motif discovery
  Cis-regulatory modules
Motif discovery

What is motif discovery?

- Start with just sequences
- Identify strongly enriched motifs \textit{de novo}
- Algorithmically one of the most challenging analysis tasks
- Use it when you suspect important \textit{unknown} motifs in your data

Motif discovery methods

- Can be classified by motif representation
  - Word-based representation
  - Matrix-based representation
- Also by algorithmic strategy
  - Discrete optimization
  - General statistical algorithms
Motif discovery by word counting

Table of words and their occurrences

<table>
<thead>
<tr>
<th>Word</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAAA</td>
<td>521</td>
</tr>
<tr>
<td>AAAAC</td>
<td>534</td>
</tr>
<tr>
<td>AAAAG</td>
<td>243</td>
</tr>
<tr>
<td>AAAAT</td>
<td>847</td>
</tr>
<tr>
<td>AAACA</td>
<td>366</td>
</tr>
<tr>
<td>AGAC</td>
<td>501</td>
</tr>
<tr>
<td>GAGGT</td>
<td>622</td>
</tr>
<tr>
<td>GAGTA</td>
<td>718</td>
</tr>
<tr>
<td>GAGTC</td>
<td>???</td>
</tr>
<tr>
<td>GAGTG</td>
<td></td>
</tr>
<tr>
<td>GAGTT</td>
<td></td>
</tr>
<tr>
<td>TTTG</td>
<td></td>
</tr>
<tr>
<td>TTTTA</td>
<td></td>
</tr>
<tr>
<td>TTTTC</td>
<td></td>
</tr>
<tr>
<td>TTTTG</td>
<td></td>
</tr>
<tr>
<td>TTTTT</td>
<td></td>
</tr>
</tbody>
</table>

For each word of width k: count number of occurrences

Apply statistics to counts

Current word: GAGTC

Smith & Sumazin (CSHL & Columbia)
Gibbs Sampling

Start with a given motif and a set of occurrences

```
GCCATCTTT
GACATTITG
TCCATTTTG
TCTAGGTIT
GCTCCATTT
TCCATITTG
GCCATTTTG
GCCATGACA
ACCATGTCA
GCCATCTTG
TCCATGTG
```

```
A
G
C
T
```

```
1 1 0 11 0 1 1 0 2
0 11 10 1 1 3 0 2 0
7 0 0 0 1 5 1 1 5
4 0 2 0 10 3 10 9 5
```
Gibbs Sampling

Iterate these steps:
1) Sample a new occurrence from one sequence

Probability of selecting particular site related to strength of match to matrix

Smith & Sumazin (CSHL & Columbia)
Iterate these steps:
1) Sample a new occurrence from one sequence
2) Update the matrix based on new occurrence

Usually the changes will move matrix toward stronger motif
Other techniques

Expectation Maximization (EM)

- Instead of sampling sites with particular probability:
  - All possible sites contribute to the matrix
  - Contribution of each site related to probability (score)
- Iterate through motifs instead of sites
- Like deterministic version of Gibbs: no random choices after setting the starting point
Other techniques

Expectation Maximization (EM)

- Instead of sampling sites with particular probability:
  - All possible sites contribute to the matrix
  - Contribution of each site related to probability (score)
- Iterate through motifs instead of sites
- Like deterministic version of Gibbs: no random choices after setting the starting point

Variants of EM or Gibbs

- Gibbs Motif Sampler (Lawrence et al., 1993)
- MEME (Bailey & Elkan, 1995)
- AlignACE (Hughes et al., 2000)
- MDscan (Liu et al., 2002)

Good starting points are **critical** for Gibbs and EM
Things to consider

Current status

- Field starting to mature: many great algorithms exist!
- Probably none will be “perfect” for your application
- Try several algorithms, understand what they do
Things to consider

Current status

- Field starting to mature: many great algorithms exist!
- Probably none will be “perfect” for your application
- Try several algorithms, understand what they do

How to improve

- Combine best aspects of different algorithms
- Incorporate more biological knowledge
Things to consider

Current status

- Field starting to mature: many great algorithms exist!
- Probably none will be “perfect” for your application
- Try several algorithms, understand what they do

How to improve

- Combine best aspects of different algorithms
- Incorporate more biological knowledge

DME: Discriminating Motif Enumerator

- Enumerative search strategy, matrix-based motifs
- Smith, Sumazin & Zhang (PNAS, 2005)
Introduction
- Background on regulatory networks
- Data available for analysis

Analysis methods
- Identifying gene modules
- Modeling regulatory elements
- Predicting binding sites
- Conservation of regulatory elements
- Motif discovery

Cis-regulatory modules
What is a *cis*-regulatory module?

**The IFNβ Enhancer**

- Figure from Maniatis et al. (CSHL Symposium 1998)
- Critical property: sites that work together tend to cluster
What is a *cis*-regulatory module?

Sea Urchin Endo16 promoter

- Figure from Yuh et al. (2001)
- Promoter logic: CRMs are autonomous units encoding regulation
### Identifying cis-regulatory modules

#### PReMod (Blanchette et al, 2006)

<table>
<thead>
<tr>
<th>Matrix name</th>
<th>Total Unit Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z15 (M00715)</td>
<td>36</td>
</tr>
<tr>
<td>Pax-2 (M00039)</td>
<td>36</td>
</tr>
<tr>
<td>Mox-1 (M00334)</td>
<td>31</td>
</tr>
</tbody>
</table>

- Occurrences tightly clustered
- Far from gene
- Strong occurrences of known motifs
- Highly conserved region

---

**Analysis methods**

**Cis-regulatory modules**

**Smith & Sumazin (CSHL & Columbia)**

**Transcriptional regulatory circuits**

**ISMB’07** 54 / 109
Motif modules

What are they?

- A set of motifs for sites that frequently work together
- CRMs are the occurrences of motif modules
- Often can predict expression better than individual motifs
- Simplest kind: pair of sites for dimerizing TFs

Interesting properties

- Relative order: some motifs must be beside each other
- Total span and spacing of sites can be restricted
- Relative orientation sometimes important
- Weaker individual sites: combined affinity is important
Discovering motif modules

Library based

- Given a library of motifs construct modular motifs
- Many known motifs work have important interactions

De-novo discovery

- Discover modular motifs from sequence alone
- Currently no generally practical methods
- Anchoring strategy: almost de novo, and can be useful
- CisModule: one of the most sophisticated algorithms
Part II

Part II: Worked Examples
Analyzing sets of co-regulated genes
   An example gene module
   Identifying enriched known motifs
   Predicting functional binding sites

Analysis of transcription factor Localization data
   ChIP-chip data examples
   Identifying enriched known motifs
   Identifying co-factors
   Discovering motifs \textit{de novo}
Analyzing sets of co-regulated genes

An example gene module
Identifying enriched known motifs
Predicting functional binding sites

Analysis of transcription factor Localization data
ChIP-chip data examples
Identifying enriched known motifs
Identifying co-factors
Discovering motifs de novo
Analyzing sets of co-regulated genes
An example gene module
- Identifying enriched known motifs
- Predicting functional binding sites

Analysis of transcription factor Localization data
- ChIP-chip data examples
- Identifying enriched known motifs
- Identifying co-factors
- Discovering motifs *de novo*
Example gene module

LPS responsive genes

- Bacterial LPS (lipopolysaccharide) stimulates B-cell activation, proliferation, and differentiation
- Gene module compiled through individual experiments
- Ramirez-Carrozzi et al. (Genes & Dev, 2006) Selective and antagonistic functions of SWI/SNF and Mi-2b nucleosome remodeling complexes during an inflammatory response

Properties of the gene module

- The gene module comprises 35 genes
- Some are TFs (e.g. Irf1, Irf7, Junb, Fos, Nfkbiz, Egr1, Zfp369)
- Several known binding sites in promoters of these genes (e.g. IFNβ enhancer)
Analysis tasks

- Identify enriched known motifs
- Use known motifs to predict functional binding sites
Obtaining promoter sequences

Promoter databases

- Examples: EPD, DBTSS, CSHLmpd
- Use when promoter choice really matters (e.g. small data sets, many alternative promoters)

UCSC Table Browser to get promoters

- Start with set of RefSeq IDs for genes in module
- Select the appropriate table (refGene for mm8)
- Upload the RefSeq IDs
- Select sequence output format
- Select “upstream by 1000bp”
Analyzing sets of co-regulated genes
   An example gene module
Identifying enriched known motifs
Predicting functional binding sites

Analysis of transcription factor Localization data
   ChIP-chip data examples
   Identifying enriched known motifs
   Identifying co-factors
   Discovering motifs *de novo*
The \textit{motifclass} program

How it evaluates enrichment

- Compares set of foreground sequences to background sequences
- For a given motif, each sequence is assigned a score
- The score is the maximum match-score of any site in the sequence
- The scores are used to classify foreground and background sequences
- Sequences with higher scores are classified as foreground
- Better classification ability means greater enrichment
- $p$-values obtained by randomly permuting sequence labels
The `motifclass` program

- **Foreground sequences**

- **Background sequences**
Using **motifclass** to evaluate motif enrichment

**Sequence files**

- Foreground: the 35 proximal promoters
- Background: 1000 random mm8 RefSeq promoters
- Promoter sequences taken -1000 to -1 relative to the TSS
- Sequences given in FASTA format

**Motif library**

- Known motifs from the JASPAR database
- Total of 123 motifs (some redundancy)
- Motifs must be converted into CREAD motif format
The CREAD motif file format

AC: the accession

- Identifier for each motif
- Best to keep them unique
The CREAD motif file format

TY: the type of pattern

- Type of this pattern is “Motif”
- Just to tell programs what they are looking at
The CREAD motif file format

The matrix lines

- This is the actual PWM
- Transposed: one line per column
- Either counts or probabilities
The CREAD motif file format

AT: the attributes

- Annotate motifs with additional information
- Attribute=value pairs
- Usually optional
- Some programs require certain attributes
The CREAD motif file format

BS: the binding site lines
- To store sites for each motif
- More details on this later
Running **motifclass** on LPS-responsive promoters

- **-r**: use relative error as enrichment measure
- **-O**: find the score cutoff optimizing that enrichment
- **-P 1000**: report a $p$-value for each motif using 1000 shuffles
- **-v**: print progress information while running
Analyzing sets of co-regulated genes
Identifying enriched known motifs

What the output looks like

Attributes from **motifclass**
- Relative error rate
- Sensitivity and specificity
- Optimal score cutoff (Functional depth and threshold)
- \( p \)-value and rank (in set of motifs)
### Interpreting the results

<table>
<thead>
<tr>
<th>Name</th>
<th>Logo</th>
<th>Sn</th>
<th>Sp</th>
<th>Error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. NFKB1</td>
<td>GGGA</td>
<td>0.743</td>
<td>0.603</td>
<td>0.327</td>
<td>0</td>
</tr>
<tr>
<td>2. RELA</td>
<td>GGGG</td>
<td>0.686</td>
<td>0.655</td>
<td>0.33</td>
<td>0</td>
</tr>
<tr>
<td>3. NF-kappaB</td>
<td>GGGA</td>
<td>0.4</td>
<td>0.9</td>
<td>0.35</td>
<td>0.002</td>
</tr>
<tr>
<td>4. Dorsal_1</td>
<td>GGGG</td>
<td>0.886</td>
<td>0.413</td>
<td>0.351</td>
<td>0</td>
</tr>
<tr>
<td>5. REL</td>
<td>GGGA</td>
<td>0.314</td>
<td>0.956</td>
<td>0.365</td>
<td>0.008</td>
</tr>
<tr>
<td>6. En1</td>
<td>GGGA</td>
<td>0.686</td>
<td>0.584</td>
<td>0.365</td>
<td>0.009</td>
</tr>
<tr>
<td>7. IRF2</td>
<td>GGGG</td>
<td>0.371</td>
<td>0.872</td>
<td>0.378</td>
<td>0.015</td>
</tr>
<tr>
<td>8. TBP</td>
<td>GGGG</td>
<td>0.371</td>
<td>0.867</td>
<td>0.381</td>
<td>0.018</td>
</tr>
<tr>
<td>9. Dorsal_2</td>
<td>GGGG</td>
<td>0.429</td>
<td>0.798</td>
<td>0.387</td>
<td>0.032</td>
</tr>
<tr>
<td>10. ZNF42_5-13</td>
<td>GGGA</td>
<td>0.629</td>
<td>0.59</td>
<td>0.391</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Smith & Sumazin (CSHL & Columbia)  
Transcriptional regulatory circuits
Implications for the LPS network

NF-κB motif highly enriched

- Top 5 motifs all NF-κB family members
- Likely a master regulator
- Expected to have multiple direct targets (next task)

Other motifs and TFs

- IRF motif is important
- Could be Irf1, Irf7 or some other Irf family member
- Other IRF motifs ranked high
Analyzing sets of co-regulated genes
   An example gene module
   Identifying enriched known motifs
Predicting functional binding sites

Analysis of transcription factor Localization data
   ChIP-chip data examples
   Identifying enriched known motifs
   Identifying co-factors
   Discovering motifs de novo
Analyzing sets of co-regulated genes
Predicting functional binding sites

Predicting functional binding sites

1) Identify candidate sites
Scan sequences for sites scoring above the cutoff for each motif.

2) Filter by location
Eliminate candidate sites occurring inside these regions.

3) Filter by conservation
Eliminate candidate sites without desired conservation properties.

Predicted sites
Final set of predicted sites; to be evaluated experimentally

Motif library
Known or novel motifs whose sites we want to identify

Sequences
Where we will search (e.g. promoters)

Genomeic Regions
Where functional sites are not likely (e.g. inside CDS)

Alignments
for conservation in sequences searched
Identifying candidate sites

About this step

- Goal: identify sites that strongly match our motifs
- Sequences: 1000bp promoters of the 35 LPS-responsive genes
- Motif library: the JASPAR motifs
- We will use the storm program for finding sites

The STORM program

- Select a \( p \)-value cutoff
- Indicate that the cutoff is a match-score \( p \)-value
- Often difficult to select this
Using the storm program

- **-C**: give a base composition (used to build scoring matrices)
- **-t**: specify the score threshold for sites
- **-p**: indicate that the threshold is a match-score $p$-value
- **-v**: print progress information while running

```
[15:27][elite3 sites]$ ls
browser_header.txt  conserved_header.txt  most_conserved_header.txt
candidates_header.txt  jaspar_core.mat  README.txt
commands.txt  lps_plk.fa
[15:28][elite3 sites]$ storm -C 0.27,0.23,0.23,0.27 \  >  -vpt 1e-04 -o lps_plk_jasp_m4.mat  -s lps_plk.fa  jaspar_core.mat
base comp: 0.27,0.23,0.23,0.27
 calculating score cutoff   100%
 file lps_plk.fa 100%  (100% total)
 removing duplicate sites  100%
 adding new sites to motifs  done
[15:29][elite3 sites]$ 
```
The set of candidate binding sites

- Figure: candidate binding sites in part of a storm output file
- 1674 candidates identified for 123 motifs (13.6 sites/motif)
- Additional candidates identified in larger -10K to -1001 region
- Vast majority are false-positives, and must be filtered
Excluding less important regions

About this step

- **Goal:** eliminate candidates less likely to be functional
- **Regions to exclude:** CDS and Repeat Masker repeats
- **Functional sites are less likely in those regions**
- **Program:** *sitesifter* from CREAD

The sitesifter program

- Filters set of sites based on location
- Identifies sites contained in, or excluded from, a set of regions
- Can also filter set of sites based on scores (above/below some cutoff)
Using the sitesifter program

- Running the program is straight-forward
- Figure: filtered 402 sites contained in repeat regions
Filtering based on conservation

About this step

- Goal: identify remaining candidate sites that appear conserved
- Alignments: precomputed UCSC multiz17way alignments
- Species: all vertebrates species in the alignment
- We will use the multistorm program to evaluate site conservation

What multistorm does

- Takes a set of candidate sites for some motifs
- Evaluates the aligned sites in other species using same motif
- Given a cutoff score, count species scoring greater at aligned sites
- Final score is number of species scoring above cutoff at the site
Using the multistorm program

- **-C**: give a base composition (same as we used to get candidates)
- **-c**: specify score $p$-value cutoff (also same value as for candidates)
- **-v**: print progress information while running
- Others params specify input (i.e. alignment) and output files
- Used **sitesifter** to get the 317 sites conserved in 4 species
The set of predicted functional sites

Properties of the predicted sites

1. Each is a strong match to a known binding site motif
2. None appear in CDS or repeats regions
3. Each is conserved through multiple species

What did we find?

- 317 total sites
- Includes overlapping sites, and sites for redundant motifs
- 26 unique high-confidence predicted sites for NFkB
- 9 unique high-confidence predicted sites for IRFs
Analysis of transcription factor localization data

Analyzing sets of co-regulated genes
An example gene module
Identifying enriched known motifs
Predicting functional binding sites

Analysis of transcription factor Localization data
ChIP-chip data examples
Identifying enriched known motifs
Identifying co-factors
Discovering motifs \textit{de novo}
Analyzing sets of co-regulated genes
- An example gene module
- Identifying enriched known motifs
- Predicting functional binding sites

Analysis of transcription factor Localization data
ChIP-chip data examples
- Identifying enriched known motifs
- Identifying co-factors
- Discovering motifs *de novo*
ChIP arrays

Promoter arrays

- Use long probes to cover proximal promoters
- Probe coverage is sparse
- Transcription factor localization evidence from few probes

Tiling arrays

- Dense covering of proximal promoters, possibly including distal regions or even whole genome coverage
- Varying coverage density
- Transcription factor localization evidence from a set of probes
E2F4 localization in primary human fibroblasts

E2F4 background

- The E2F family of transcription factors is essential for cell cycle activity
- E2F transcription factors are known to bind proximally to the TSS
- E2F4 is known to regulate the G2/M phase

Our data

- A set of probed promoters
- A subset composed of promoters found to be localized with E2F4

Our task

- Identify enriched motifs in the set of E2F4-localized promoters
CTCF localization in primary human fibroblasts

CTCF background

- CTCF is an 11-zink finger vertebrate nuclear insulator
- CTCF binds far from transcription start sites
- CTCF localization appears to be independent of cell type

Our data

- A set of regions that were identified to be localized with CTCF

Our task

- Discover and identify enriched motifs in the set of CTCF-localized regions
Obtaining sequence sets

E2F foreground and background sequences

- Over 10K probed promoters to form the foreground
- Segments lengths from 700 to 1000 have to be normalized
- 236 E2F4-localized promoters
- Background selected by sampling from the remaining promoters

CTCF foreground and background sequences

- Over 15K CTCF-localized segments to form the foreground
- Segments lengths from 350 to 5150 have to be normalized
- We analyze a sample – 500 is plenty
- Background constructed by either
  - shuffling the foreground to preserve base composition or dinucleotide composition
  - using non-overlapping same-size flanking regions
Analyzing sets of co-regulated genes
  An example gene module
  Identifying enriched known motifs
  Predicting functional binding sites

Analysis of transcription factor Localization data
  ChIP-chip data examples
  Identifying enriched known motifs
  Identifying co-factors
  Discovering motifs \textit{de novo}
Selecting foreground and background – CTCF

- Sample 500 sequences from CTCF-localized segments
- Identify non-overlapping flanking regions
Selecting foreground and background – CTCF

- Shuffling to preserve base composition and dinucleotide composition

```
~/Data/ismb07tutorial_ChIP/CTCF/sequences>ll
total 1560
-rw-r--r-- 1 ps ps 1044754 Jul 18 16:53 flanking.fa
-rw-r--r-- 1 ps ps 528 Jul 18 18:43 README.txt
-rw-r--r-- 1 ps ps 522302 Jul 18 16:53 sample.fa
~/Data/ismb07tutorial_ChIP/CTCF/sequences>shufflet 1 1 < sample.fa > shuffle.fa
~/Data/ismb07tutorial_ChIP/CTCF/sequences>shufflet 1 2 < sample.fa > \
> dinucShuffle.fa
~/Data/ismb07tutorial_ChIP/CTCF/sequences>l
```
Analysis of transcription factor localization data
Identifying enriched known motifs

Running **motifclass**

```
~/Data/ismb07tutorial_ChIP/CTCF/jasparMotifs>ls
total 32
-rw-r----- 1 ps ps 431 Jul 19 11:37 dinucShuffle-commands.txt
-rw-r----- 1 ps ps 407 Jul 19 11:40 flanking-commands.txt
-rw-r--r-- 1 ps ps 319 Jul 18 18:30 README.txt
-rw-r----- 1 ps ps 396 Jul 19 08:37 shuffle-commands.txt
~/Data/ismb07tutorial_ChIP/CTCF/jasparMotifs>motifclass -Or -P 1000 -f \
> ../sequences/sample.fa -b ../sequences/flanking.fa -o flanking.mats \
> ../..jaspar_core.mat
~/Data/ismb07tutorial_ChIP/CTCF/jasparMotifs>
```

- **-r**: use relative error as enrichment measure
- **-O**: find the score cutoff optimizing that enrichment
- **-P 1000**: report a $p$-value for each motif using 1000 shuffles
uniqmotifs and matcompare

Programs to compare motifs

- Consider all legal alignments specified using max overhang for the smaller matrix (-h)
- Require that the average K-L divergence per aligned column is no greater than specified (-t)
- uniqmotifs clusters a sorted list of similar motifs so that lower ranking motifs are listed below similar higher ranking motifs
- matcompare queries a motif library to identify similar motifs to those in the input list
Sorting and pruning

- Sort by relative error rate
- Cluster similar motifs
## CTCF – enrichment against shuffled and flanking

<table>
<thead>
<tr>
<th>Acc</th>
<th>TF</th>
<th>Err</th>
<th>Sens</th>
<th>Spec</th>
<th>p-val</th>
<th>FD</th>
<th>Logo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 MA0045</td>
<td>HMG-IY</td>
<td>0.348</td>
<td>0.60</td>
<td>0.70</td>
<td>0.000</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>MA0120</td>
<td>ID1</td>
<td>0.416</td>
<td>0.43</td>
<td>0.74</td>
<td>0.000</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>2 MA0041</td>
<td>Foxd3</td>
<td>0.355</td>
<td>0.72</td>
<td>0.57</td>
<td>0.000</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>MA0042</td>
<td>FOXI1</td>
<td>0.386</td>
<td>0.76</td>
<td>0.47</td>
<td>0.000</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>3 MA0013</td>
<td>Broad-complex4</td>
<td>0.381</td>
<td>0.61</td>
<td>0.63</td>
<td>0.000</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>4 MA0010</td>
<td>Broad-complex1</td>
<td>0.383</td>
<td>0.67</td>
<td>0.56</td>
<td>0.000</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>5 MA0082</td>
<td>SQUA</td>
<td>0.385</td>
<td>0.61</td>
<td>0.62</td>
<td>0.000</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>1 MA0123</td>
<td>ABI4</td>
<td>0.435</td>
<td>0.32</td>
<td>0.81</td>
<td>0.000</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>2 MA0003</td>
<td>TFAP2A</td>
<td>0.443</td>
<td>0.50</td>
<td>0.61</td>
<td>0.000</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>3 MA0048</td>
<td>NHLH1</td>
<td>0.445</td>
<td>0.62</td>
<td>0.49</td>
<td>0.000</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>4 MA0117</td>
<td>MafB</td>
<td>0.448</td>
<td>0.45</td>
<td>0.66</td>
<td>0.000</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>5 MA0028</td>
<td>ELK1</td>
<td>0.449</td>
<td>0.27</td>
<td>0.83</td>
<td>0.000</td>
<td>0.91</td>
<td></td>
</tr>
</tbody>
</table>
Selecting foreground and background – E2F

- Sample 500 sequences from non-positive promoters
Running **motifclass**

- **-r**: use relative error as enrichment measure
- **-O**: find the score cutoff optimizing that enrichment
- **-P 1000**: report a *p*-value for each motif using 1000 shuffles
## E2F enrichment

<table>
<thead>
<tr>
<th>Acc</th>
<th>TF</th>
<th>Err</th>
<th>Sens</th>
<th>Spec</th>
<th>p-val</th>
<th>FD</th>
<th>Logo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MA0060</td>
<td>NF-Y</td>
<td>0.327</td>
<td>0.64</td>
<td>0.71</td>
<td>0.000</td>
<td>0.86</td>
</tr>
<tr>
<td>2</td>
<td>MA0024</td>
<td>E2F1</td>
<td>0.350</td>
<td>0.72</td>
<td>0.57</td>
<td>0.000</td>
<td>0.86</td>
</tr>
<tr>
<td>3</td>
<td>MA0080</td>
<td>SPI1</td>
<td>0.359</td>
<td>0.61</td>
<td>0.68</td>
<td>0.000</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>MA0028</td>
<td>ELK1</td>
<td>0.364</td>
<td>0.58</td>
<td>0.69</td>
<td>0.000</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>MA0076</td>
<td>ELK4</td>
<td>0.373</td>
<td>0.62</td>
<td>0.64</td>
<td>0.000</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>MA0026</td>
<td>E74A</td>
<td>0.399</td>
<td>0.42</td>
<td>0.78</td>
<td>0.000</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>MA0062</td>
<td>GABPA</td>
<td>0.412</td>
<td>0.64</td>
<td>0.54</td>
<td>0.000</td>
<td>0.85</td>
</tr>
<tr>
<td>4</td>
<td>MA0021</td>
<td>Dof3</td>
<td>0.375</td>
<td>0.45</td>
<td>0.80</td>
<td>0.000</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>MA0020</td>
<td>Dof2</td>
<td>0.449</td>
<td>0.91</td>
<td>0.19</td>
<td>0.006</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>MA0053</td>
<td>MNB1A</td>
<td>0.449</td>
<td>0.91</td>
<td>0.19</td>
<td>0.000</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>MA0064</td>
<td>PBF</td>
<td>0.449</td>
<td>0.91</td>
<td>0.19</td>
<td>0.003</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>MA0123</td>
<td>ABI4</td>
<td>0.398</td>
<td>0.81</td>
<td>0.40</td>
<td>0.000</td>
<td>0.91</td>
</tr>
<tr>
<td>6</td>
<td>MA0018</td>
<td>CREB1</td>
<td>0.406</td>
<td>0.43</td>
<td>0.76</td>
<td>0.000</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>MA0096</td>
<td>bZIP910</td>
<td>0.433</td>
<td>0.45</td>
<td>0.68</td>
<td>0.000</td>
<td>0.86</td>
</tr>
<tr>
<td>7</td>
<td>MA0034</td>
<td>GAMYB</td>
<td>0.420</td>
<td>0.59</td>
<td>0.57</td>
<td>0.000</td>
<td>0.90</td>
</tr>
</tbody>
</table>
Testing CpG-island influence

- The positive set is highly CpG enriched and the analysis may be biased – identifying patterns common to special or just active promoters
- We compare foreground CpG-island promoters to background CpG-island promoters to eliminate this potential bias
## E2F CpG-conditional enrichment

<table>
<thead>
<tr>
<th>Acc</th>
<th>TF</th>
<th>Err</th>
<th>Sens</th>
<th>Spec</th>
<th>p-val</th>
<th>FD</th>
<th>Logo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MA0060</td>
<td>NF-Y</td>
<td>0.321</td>
<td>0.64</td>
<td>0.71</td>
<td>0.000</td>
<td>0.86</td>
</tr>
<tr>
<td>2</td>
<td>MA0024</td>
<td>E2F1</td>
<td>0.344</td>
<td>0.73</td>
<td>0.58</td>
<td>0.000</td>
<td>0.86</td>
</tr>
<tr>
<td>3</td>
<td>MA0080</td>
<td>SPI1</td>
<td>0.347</td>
<td>0.63</td>
<td>0.68</td>
<td>0.000</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>MA0028</td>
<td>ELK1</td>
<td>0.367</td>
<td>0.58</td>
<td>0.69</td>
<td>0.000</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>MA0076</td>
<td>ELK4</td>
<td>0.373</td>
<td>0.54</td>
<td>0.72</td>
<td>0.000</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>MA0062</td>
<td>GABPA</td>
<td>0.406</td>
<td>0.66</td>
<td>0.53</td>
<td>0.000</td>
<td>0.85</td>
</tr>
<tr>
<td>4</td>
<td>MA0021</td>
<td>Dof3</td>
<td>0.374</td>
<td>0.46</td>
<td>0.80</td>
<td>0.000</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>MA0053</td>
<td>MNB1A</td>
<td>0.448</td>
<td>0.91</td>
<td>0.19</td>
<td>0.000</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>MA0064</td>
<td>PBF</td>
<td>0.448</td>
<td>0.91</td>
<td>0.19</td>
<td>0.000</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>MA0123</td>
<td>ABI4</td>
<td>0.386</td>
<td>0.83</td>
<td>0.39</td>
<td>0.000</td>
<td>0.91</td>
</tr>
<tr>
<td>6</td>
<td>MA0018</td>
<td>CREB1</td>
<td>0.396</td>
<td>0.45</td>
<td>0.76</td>
<td>0.000</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>MA0096</td>
<td>bZIP910</td>
<td>0.434</td>
<td>0.30</td>
<td>0.83</td>
<td>0.002</td>
<td>0.88</td>
</tr>
<tr>
<td>7</td>
<td>MA0034</td>
<td>GAMYB</td>
<td>0.407</td>
<td>0.62</td>
<td>0.57</td>
<td>0.000</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>MA0100</td>
<td>Myb</td>
<td>0.424</td>
<td>0.57</td>
<td>0.58</td>
<td>0.000</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Analyzing sets of co-regulated genes
   An example gene module
   Identifying enriched known motifs
   Predicting functional binding sites

Analysis of transcription factor Localization data
   ChIP-chip data examples
   Identifying enriched known motifs
   Identifying co-factors
   Discovering motifs *de novo*
Identifying co-factors

- MA0060 (NF-Y) and MA0024 (E2F1) are the best localization predictors
- To identify possible cofactors we
  - identify putative sites for the two motifs
  - get flanking regions to search for co-factor sites
  - identify enriched motifs in flanking regions
  - we search only in CpG-island promoters to eliminate bias
Evaluating putative co-factors

~/Data/ismb07tutorial_ChIP/E2F/anchoring>
ll
total 44
-rw-r--r-- 1 ps ps 849 Jul 19 13:49 MA0024-commands.txt
-rw-r--r-- 1 ps ps 1079 Jul 18 21:02 MA0024.mat
-rw-r--r-- 1 ps ps 849 Jul 19 13:48 MA0060-commands.txt
-rw-r--r-- 1 ps ps 5673 Jul 18 21:02 MA0060.mat
-rw-r--r-- 1 ps ps 877 Jul 18 22:39 README.txt
~/Data/ismb07tutorial_ChIP/E2F/anchoring>
cat ../sequences/cpg.fa \ > ../sequences/negCpGsampler.fa > ../sequences/cpgAndNegCpGsampler.fa
~/Data/ismb07tutorial_ChIP/E2F/anchoring>storm -l RELATIVE_ERRORTIMATE_THRESHOLD \ > -C `basecomp ../sequences/cpgAndNegCpGsampler.fa` MA0024.mat -s \ > ../sequences/cpg.fa -o MA0024.pos
~/Data/ismb07tutorial_ChIP/E2F/anchoring>storm -l RELATIVE_ERRORTIMATE_THRESHOLD \ > -C `basecomp ../sequences/cpgAndNegCpGsampler.fa` MA0024.mat -s \ > ../sequences/negCpGsampler.fa -o MA0024.neg
~/Data/ismb07tutorial_ChIP/E2F/anchoring>getcontext -f ../sequences/cpg.fa \ > -o MA0024pos.fa -c 200 -m MA0024.pos
~/Data/ismb07tutorial_ChIP/E2F/anchoring>getcontext -f ../sequences/cpg.fa \ > -o MA0024neg.fa -c 200 -m MA0024.neg
~/Data/ismb07tutorial_ChIP/E2F/anchoring>motifclass -Or -P 1000 -f MA0024pos.fa\ > -b MA0024neg.fa -o MA0024.mats ../jasper_core.mat
## Enrichment in proximity to MA0024 sites

<table>
<thead>
<tr>
<th>Acc</th>
<th>TF</th>
<th>Err</th>
<th>Sens</th>
<th>Spec</th>
<th>p-val</th>
<th>FD</th>
<th>Logo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 MA0060</td>
<td>NF-Y</td>
<td>0.390</td>
<td>0.37</td>
<td>0.85</td>
<td>0.000</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>2 MA0037</td>
<td>GATA3</td>
<td>0.417</td>
<td>0.65</td>
<td>0.52</td>
<td>0.000</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GATA2</td>
<td>0.433</td>
<td>0.61</td>
<td>0.53</td>
<td>0.002</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pbx</td>
<td>0.437</td>
<td>0.70</td>
<td>0.42</td>
<td>0.016</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ubx</td>
<td>0.438</td>
<td>0.81</td>
<td>0.31</td>
<td>0.002</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>3 MA0011</td>
<td>Broad-complex2</td>
<td>0.419</td>
<td>0.89</td>
<td>0.27</td>
<td>0.000</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SQUA</td>
<td>0.435</td>
<td>0.81</td>
<td>0.32</td>
<td>0.011</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>4 MA0110</td>
<td>ATHB5</td>
<td>0.421</td>
<td>0.70</td>
<td>0.45</td>
<td>0.000</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prrx2</td>
<td>0.428</td>
<td>0.77</td>
<td>0.38</td>
<td>0.000</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Athb-1</td>
<td>0.430</td>
<td>0.91</td>
<td>0.23</td>
<td>0.001</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>5 MA0096</td>
<td>bZIP910</td>
<td>0.423</td>
<td>0.41</td>
<td>0.75</td>
<td>0.000</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CREB1</td>
<td>0.428</td>
<td>0.72</td>
<td>0.43</td>
<td>0.001</td>
<td>0.78</td>
<td></td>
</tr>
</tbody>
</table>
Analyzing sets of co-regulated genes
   An example gene module
   Identifying enriched known motifs
   Predicting functional binding sites

Analysis of transcription factor Localization data
   ChIP-chip data examples
   Identifying enriched known motifs
   Identifying co-factors
   Discovering motifs *de novo*
Running DME

Overview

- DME enumerates through a set of matrices to identify those with the greatest number of potential sites in the foreground relative to the background
- DME restricts the type of matrices it evaluates
  - it evaluates matrices with width specified using `-w`
  - it evaluates only those matrices that have a minimum average information per column specified using `-i`
  - the number of matrices it reports is set using `-n`
  - it evaluates matrices corresponding to degenerate words with the level of degeneracy optionally specified using `-g`
  - it uses a 2-iteration scheme, refining discovered motifs to a higher degeneracy optionally specified using `-r`
Running DME

```
~/Data/ismb07tutorial_ChIP/CTCF/deNovoMotifs>ll
total 32
-rw-r--r-- 1 ps ps 1084 Jul 18 18:01 dinucShuffle-commands.txt
-rw-r--r-- 1 ps ps 1005 Jul 20 03:04 flanking-commands.txt
-rw-r--r-- 1 ps ps 1156 Jul 18 18:41 README.txt
-rw-r--r-- 1 ps ps 984 Jul 18 18:01 shuffle-commands.txt
~/Data/ismb07tutorial_ChIP/CTCF/deNovoMotifs>dme -n 50 -w 8 -i 1.8 -r 0.125 \
> -a DME-8-1.80- -o flanking-DME-8-1.80 -b ../sequences/flanking.fa \
> ../sequences/sample.fa
~/Data/ismb07tutorial_ChIP/CTCF/deNovoMotifs>dme -n 50 -w 10 -i 1.6 -r 0.125 \
> -a DME-10-1.60- -o flanking-DME-10-1.60 -b ../sequences/flanking.fa \
> ../sequences/sample.fa
~/Data/ismb07tutorial_ChIP/CTCF/deNovoMotifs>cat flanking-DME-8-1.80 \
> flanking-DME-10-1.60 > flanking.mats
~/Data/ismb07tutorial_ChIP/CTCF/deNovoMotifs>
```
Analysis of transcription factor localization data

Discovering motifs de novo

Evaluating motif enrichment

```
~/Data/ismb07tutorial_ChIP/CTCF/deNovoMotifs>
```

```
total 3352
-rw-r--r-- 1 ps ps 1084 Jul 18 18:01 dinucShuffle-commands.txt
-rw-r--r-- 1 ps ps 1004 Jul 18 17:28 flanking-commands.txt
-rw-r--r-- 1 ps ps 900688 Jul 18 16:57 flanking-DME-10-1.60
-rw-r--r-- 1 ps ps 581454 Jul 18 16:57 flanking-DME-8-1.80
-rw-r--r-- 1 ps ps 1891760 Jul 18 17:10 flanking.mats
-rw-r--r-- 1 ps ps 1156 Jul 18 18:41 README.txt
-rw-r--r-- 1 ps ps 984 Jul 18 18:01 shuffle-commands.txt

~/Data/ismb07tutorial_ChIP/CTCF/deNovoMotifs>
```

```
motifclass -Or -f
> ../sequences/sample.fa -b ../sequences/flanking.fa
> -o flanking.mats flanking
```

```
~/Data/ismb07tutorial_ChIP/CTCF/deNovoMotifs>
```

```
sortmotifs -ank
> RELATIVE_ERRORRATE flanking.mats -o flanking
```

```
~/Data/ismb07tutorial_ChIP/CTCF/deNovoMotifs>
```

```
uniqmotifs -kh 1 -t 1.0
> -r RELATIVE_ERRORRATE flanking -o flanking.mats
```

```
~/Data/ismb07tutorial_ChIP/CTCF/deNovoMotifs>
```

Smith & Sumazin (CSHL & Columbia)
## Motif enrichment

<table>
<thead>
<tr>
<th>Acc</th>
<th>Err</th>
<th>Sens</th>
<th>Spec</th>
<th>FD</th>
<th>Logo</th>
</tr>
</thead>
</table>
| DME-10-1.60-6 | 0.372| 0.42 | 0.83 | 0.90 | TC
|              | TGCC| AC   | CT   | GG  | C    |
| DME-10-1.60-10| 0.373| 0.42 | 0.83 | 0.98 | CC
|              | AGCC| GA   | TGCC| CT   | GG  |
| DME-10-1.60-11| 0.375| 0.49 | 0.76 | 0.97 | AG
|              | GGAT| GC   | C    | GA   | GG  |
| DME-10-1.60-28| 0.422| 0.60 | 0.56 | 0.90 | CT
|              | CGTC| AC   | GA   | GG   | C   |
| DME-10-1.60-26| 0.423| 0.54 | 0.61 | 0.90 | GG
|              | AGGA| G    | CG   | GC   | AG  |
| DME-10-1.60-39| 0.423| 0.42 | 0.73 | 0.90 | GC
|              | CTGC| AT   | GC   | AG   | GC  |
| DME-10-1.60-27| 0.426| 0.38 | 0.77 | 0.98 | GA
|              | TATG| GC   | CAG  | CA   | TG  |
| DME-10-1.60-23| 0.427| 0.54 | 0.61 | 0.90 | CG
|              | TGCC| TT   | GC   | GC   | GC  |
| DME-10-1.60-35| 0.427| 0.32 | 0.83 | 0.96 | GG
|              | AGAG| GC   | G    | AG   | GA  |
| DME-10-1.60-13| 0.429| 0.36 | 0.78 | 0.97 | AG
|              | AGGA| AG   | G    | AG   | AG  |
Summary – gene-module and ChIP-chip examples

The good news

- ChIP-chip data can be used to describe binding affinity of sequence-specific transcription factors
- Good tools exist to discover and evaluate motifs for their ability to predict expression and binding
- Some tools exist for identifying co-factor binding affinity

Careful analysis is paramount

- Select negative control carefully
- Try to make certain that you are detecting DNA patterns associated with the phenomena under investigation
- Reverse engineering regulatory circuits using sequence analysis is always a detective story – tooling is important but experience shows that each case is special and requires specialized analysis