Canadian Bioinformatics Workshops

www.bioinformatics.ca

Contains material by Wyeth Wasserman, William Noble, Michael Hoffman, and Tim Bailey
Module 3
Transcription Factor Regulatory Analysis

Michael M. Hoffman
Pathway and Network Analysis of -omics Data
June 4-6, 2014

Overview

Part 1: Overview of transcription

Part 2: Prediction of transcription factor binding sites using binding profiles (“Discrimination”)

Part 3: Detection of novel motifs (TFBS) over-represented in regulatory regions

Part 4: Interrogation of sets of co-expressed genes or ChIP-seq regions to identify mediating transcription factors

Part 5: Gene regulatory networks
Part 1
Introduction to transcription in eukaryotic cells

Transcription over-simplified

Three-step Process:
1. TF binds to TFBS (DNA)
2. TF catalyzes recruitment of polymerase II complex
3. Production of RNA from transcription start site (TSS)
Anatomy of transcriptional regulation

WARNING: Terms vary widely in meaning between scientists

- Core promoter – Sufficient for initiation of transcription; orientation dependent
  - TSS – transcription start site
    - Often really a transcription start region
- TFBS – single transcription factor binding site
- Regulatory regions
  - Proximal/Distal – vague reference to distance from TSS
  - May be positive (enhancing) or negative (repressing)
  - Orientation independent (generally)
  - Modules – Sets of TFBS within a region that function together
- Transcriptional unit
  - DNA sequence transcribed as a single polycistronic mRNA

Complexity in transcription
Laboratory data on regulatory regions

- **Promoters**
  - RNA 5' ends (CAGE)
  - Epigenetic marks (ChIP-seq)
  - Polymerase complex (ChIP-seq)
  - RNA (RNA-seq)

- **TFBSs**
  - TFs (ChIP-seq)

- **Regulatory regions**
  - Co-activators (ChIP-seq)
  - Epigenetic marks (ChIP-seq)
  - Enhancer RNA (RNA-seq)

Accessing laboratory data

- **UCSC Genome Browser**
  - [http://genome.ucsc.edu](http://genome.ucsc.edu)

- **Gene Expression Omnibus (GEO)**

- **ENCODE Project**
  - [http://encodeproject.org/](http://encodeproject.org/)

- **Roadmap Epigenomics**
  - [http://www.roadmapepigenomics.org/](http://www.roadmapepigenomics.org/)

- **oRegAnno**
  - [http://www.oreganno.org](http://www.oreganno.org)
Overview

Part 1: Overview of transcription

Part 2: Prediction of transcription factor binding sites using binding profiles (“Discrimination”)

Part 3: Detection of novel motifs (TFBS) over-represented in regulatory regions

Part 4: Interrogation of sets of co-expressed genes or ChIP-seq regions to identify mediating transcription factors

Part 5: Gene regulatory networks

Part 2
Prediction of TF binding sites

Teaching a computer to find TFBS...
Representing binding sites for a TF

- Single site
  - AAGTTAATGA

- Set of sites, represented as a consensus
  - VDRTWRWWSHD (IUPAC degenerate DNA)

- Set of sites, represented as a position frequency matrix (PFM)

```
A 14 16 4 0 1 19 20 1 4 13 4 4 13 12 3
C 3 0 0 0 0 0 0 7 3 1 0 3 1 12
G 4 3 17 0 0 2 0 9 1 3 0 5 2 2
T 0 2 0 21 20 0 1 20 1 4 13 17 8 6 4
```

Logo – A graphical representation of frequency matrix. Y-axis is information content, which reflects the strength of the pattern in each column of the matrix.

Conversion of PFM to position specific scoring matrices (PSSM)

Add the following features to the matrix profile:
1. Correct for nucleotide frequencies in genome
2. Weight for the confidence (depth) in the pattern
3. Convert to log-scale probability for easy arithmetic

```
PFM
A 5 0 1 0 0
C 0 2 2 4 0
G 0 3 1 0 4
T 0 0 1 1 1
```

```
PSSM
A 1.6 -1.7 -0.2 -1.7 -1.7
C -1.7 0.5 0.5 1.3 -1.7
G -1.7 1.0 -0.2 -1.7 1.3
T -1.7 -1.7 -0.2 -0.2 -0.2
```

TGCTG = 0.9
**Detecting binding sites in a single sequence**

### Raw scores

```
ACCCCTCCAGGGCCGGGGGCGGTGGCCAGGACGGTAGCTCC
```

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-0.2284</td>
<td>0.4368</td>
<td>-1.5</td>
<td>-1.5</td>
<td>-1.5</td>
<td>0.4368</td>
<td>-1.5</td>
<td>-1.5</td>
<td>-0.2284</td>
</tr>
<tr>
<td>C</td>
<td>-0.2284</td>
<td>-0.2284</td>
<td>-1.5</td>
<td>-1.5</td>
<td>1.5128</td>
<td>-1.5</td>
<td>-0.2284</td>
<td>-1.5</td>
<td>-0.2284</td>
</tr>
<tr>
<td>G</td>
<td>1.2348</td>
<td>1.2348</td>
<td>2.1222</td>
<td>2.1222</td>
<td>0.4368</td>
<td>1.2348</td>
<td>1.5128</td>
<td>1.7457</td>
<td>1.7457</td>
</tr>
<tr>
<td>T</td>
<td>0.4368</td>
<td>-0.2284</td>
<td>-1.5</td>
<td>-1.5</td>
<td>-0.2284</td>
<td>0.4368</td>
<td>0.4368</td>
<td>0.4368</td>
<td>-1.5</td>
</tr>
</tbody>
</table>
```

**Abs_score = 13.4** (sum of column scores)

### Relative scores

```
ACCCCTCCAGGGCCGGGGGCGGTGGCCAGGACGGTAGCTCC
```

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-0.2284</td>
<td>0.4368</td>
<td>-1.5</td>
<td>-1.5</td>
<td>-1.5</td>
<td>0.4368</td>
<td>-1.5</td>
<td>-1.5</td>
<td>-0.2284</td>
</tr>
<tr>
<td>C</td>
<td>-0.2284</td>
<td>-0.2284</td>
<td>-1.5</td>
<td>-1.5</td>
<td>1.5128</td>
<td>-1.5</td>
<td>-0.2284</td>
<td>-1.5</td>
<td>-0.2284</td>
</tr>
<tr>
<td>G</td>
<td>1.2348</td>
<td>1.2348</td>
<td>2.1222</td>
<td>2.1222</td>
<td>0.4368</td>
<td>1.2348</td>
<td>1.5128</td>
<td>1.7457</td>
<td>1.7457</td>
</tr>
<tr>
<td>T</td>
<td>0.4368</td>
<td>-0.2284</td>
<td>-1.5</td>
<td>-1.5</td>
<td>-0.2284</td>
<td>0.4368</td>
<td>0.4368</td>
<td>0.4368</td>
<td>-1.5</td>
</tr>
</tbody>
</table>
```

**Max_score = 15.2** (sum of highest column scores)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-0.2284</td>
<td>0.4368</td>
<td>-1.5</td>
<td>-1.5</td>
<td>-1.5</td>
<td>0.4368</td>
<td>-1.5</td>
<td>-1.5</td>
<td>-0.2284</td>
</tr>
<tr>
<td>C</td>
<td>-0.2284</td>
<td>-0.2284</td>
<td>-1.5</td>
<td>-1.5</td>
<td>1.5128</td>
<td>-1.5</td>
<td>-0.2284</td>
<td>-1.5</td>
<td>-0.2284</td>
</tr>
<tr>
<td>G</td>
<td>1.2348</td>
<td>1.2348</td>
<td>2.1222</td>
<td>2.1222</td>
<td>0.4368</td>
<td>1.2348</td>
<td>1.5128</td>
<td>1.7457</td>
<td>1.7457</td>
</tr>
<tr>
<td>T</td>
<td>0.4368</td>
<td>-0.2284</td>
<td>-1.5</td>
<td>-1.5</td>
<td>-0.2284</td>
<td>0.4368</td>
<td>0.4368</td>
<td>0.4368</td>
<td>-1.5</td>
</tr>
</tbody>
</table>
```

**Min_score = -10.3** (sum of lowest column scores)

**Rel_score = Abs_score - Min_score**

= 13.4 - (-10.3)

= 23.7% (100%)

= 15.2 - (-10.3) = 25.5%

**Empirical p-value Scores**

```
0.0 0.2 0.4 0.6 0.8 1.0
```

**JASPAR:**

An open-access database of TF binding profiles

**http://jaspar.genereg.net**
The Good...

- Tronche (1997) tested 50 predicted HNF1 TFBS using an in vitro binding test and found that 96% of the predicted sites were bound!

- Stormo and Fields (1998) found in detailed biochemical studies that the best weight matrices produce scores highly correlated with in vitro binding energy

...the Bad...

- Fickett (1995) found that a profile for the MyoD TF made predictions at a rate of 1 per ~500bp of human DNA sequence
  - This corresponds to an average of 20 sites / gene (assuming 10,000 bp as average gene size)
...and the Ugly!

Human Cardiac α-Actin gene analyzed with a set of profiles
(each line represents a TFBS prediction)

Futility conjecture:
TFBS predictions are almost always wrong

Red boxes are protein coding exons - TFBS predictions excluded in this analysis

More stringency doesn’t help

- Counter to intuition, the ratio of true positives to predictions fails to improve for “stringent” thresholds
  - For most predictive models this ratio would increase
- Why?
  - True binding sites are defined by properties not incorporated into the profile scores - above some threshold all sites could be bound if present in the right setting
Section 2

What have we learned?

• PSSMs accurately reflect *in vitro* binding properties of DNA binding proteins

• Suitable binding sites occur at a rate far too frequent to reflect *in vivo* function

• Bioinformatics methods that use PSSMs for binding site studies must incorporate additional information to enhance specificity
  – Unfiltered predictions are too noisy for most applications
  – Organisms with short regulatory sequences are less problematic (e.g. yeast and E.coli)

Overview

**Part 1:** Overview of transcription

**Part 2:** Prediction of transcription factor binding sites using binding profiles (“Discrimination”)

**Part 3:** Detection of novel motifs (TFBS) over-represented in regulatory regions

**Part 4:** Interrogation of sets of co-expressed genes or ChIP-seq regions to identify mediating transcription factors

**Part 5:** Gene regulatory networks
Part 3: de novo Discovery of TF Binding Sites

Motif discovery problem

- Given sequences

\[
\begin{align*}
&\text{seq. 1} \\
&\text{seq. 2} \\
&\text{seq. 3}
\end{align*}
\]

- Find motif

\[
\begin{align*}
&\text{IGRGGFGEVY at position 515} \\
&\text{LGECFGQVV at position 430} \\
&\text{VGSGFGQVY at position 682}
\end{align*}
\]
Motif discovery problem

• Given:
  – a sequence or family of sequences.

• Find:
  – the number of motifs
  – the width of each motif
  – the locations of motif occurrences

Why is this hard?

• Input sequences are long (thousands or millions of residues).
• Motif may be subtle
  – Instances are short.
  – Instances are only slightly similar.
TFBS motif discovery example

We are given a set of promoters from co-regulated genes.

An unknown transcription factor binds to positions unknown to us, on either DNA strand.
TFBS motif discovery example

The DNA binding motif of the transcription factor can be described by a position-specific scoring matrix (PSSM).

The sequence motif discovery problem is to discover the sites (or the motif) given just the sequences.

Module 3  bioinformatics.ca
Gibbs sampling

Alternating approach

1. Guess an initial weight matrix
2. Use weight matrix to predict instances in the input sequences
3. Use instances to predict a weight matrix
4. Repeat 2 & 3 until satisfied.
Initialization

- Randomly guess an instance $s_i$ from each of $t$ input sequences $\{S_1, ..., S_t\}$.

Gibbs sampler

- Initially: randomly guess an instance $s_i$ from each of $t$ input sequences $\{S_1, ..., S_t\}$.
- Steps 2 & 3 (search):
  - Throw away an instance $s_i$: remaining ($t - 1$) instances define weight matrix.
  - Weight matrix defines instance probability at each position of input string $S_i$
  - Pick new $s_i$ according to probability distribution
- Return highest-scoring motif seen
Sampler step illustration:

<table>
<thead>
<tr>
<th>ACAGTGT</th>
<th>TAGCCGT</th>
<th>ACACCGT</th>
<th>????????</th>
<th>CAGGTTT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>.45</td>
<td>.45</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>C</td>
<td>.25</td>
<td>.45</td>
<td>.05</td>
<td>.45</td>
</tr>
<tr>
<td>G</td>
<td>.05</td>
<td>.05</td>
<td>.45</td>
<td>.65</td>
</tr>
<tr>
<td>T</td>
<td>.25</td>
<td>.05</td>
<td>.05</td>
<td>.45</td>
</tr>
</tbody>
</table>

ACAGTGT
TAGCCGT
ACACCGT
ACGCCGT
CAGGTTT

sequence 4

ACGCCGT: 20%
ACGCGCT: 52%

TOMTOM predicts which protein(s) may bind a DNA motif

Query motif

Alignment to Matching Motif

• TOMTOM compares the query motif against all motifs in databases of known motifs (e.g., Transfac).
• TOMTOM reports all statistically significant matches.
Overview

Part 1: Overview of transcription

Part 2: Prediction of transcription factor binding sites using binding profiles (“Discrimination”)

Part 3: Detection of novel motifs (TFBS) over-represented in regulatory regions

Part 4: Interrogation of sets of co-expressed genes or ChIP-seq regions to identify mediating transcription factors

Part 5: Gene regulatory networks

Part 4:
Inferring regulating TFs for sets of co-expressed genes
Deciphering regulation of co-expressed genes

TFBS enrichment

- Akin to methods for GO term enrichment/over-representation analysis, we seek to determine if a set of co-expressed genes contains an over-abundance of predicted binding sites for a known TF
Two examples of TFBS enrichment

- More Genes with TFBS
- More Total TFBS

Statistical methods for identifying enriched TFBS

- Binomial test (Z scores)
  - Based on the number of occurrences of the TFBS relative to background
  - Normalized for sequence length
  - Simple binomial distribution model

- Fisher’s exact test probability scores
  - Based on the number of genes containing the TFBS relative to background
  - Hypergeometric probability distribution
Validation using reference gene sets

### A. Muscle-specific (23 input; 16 analyzed)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Z-score</th>
<th>Fisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRF</td>
<td>1</td>
<td>21.41</td>
</tr>
<tr>
<td>1.18e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEF2</td>
<td>2</td>
<td>18.12</td>
</tr>
<tr>
<td>8.05e-04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c-MYB_1</td>
<td>3</td>
<td>14.41</td>
</tr>
<tr>
<td>1.25e-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myf</td>
<td>4</td>
<td>13.54</td>
</tr>
<tr>
<td>3.83e-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TEF-1</td>
<td>5</td>
<td>11.22</td>
</tr>
<tr>
<td>2.87e-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>deltaEF1</td>
<td>6</td>
<td>10.88</td>
</tr>
<tr>
<td>1.09e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S8</td>
<td>7</td>
<td>5.874</td>
</tr>
<tr>
<td>2.93e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Irf-1</td>
<td>8</td>
<td>5.245</td>
</tr>
<tr>
<td>2.63e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thing1-E47</td>
<td>9</td>
<td>4.485</td>
</tr>
<tr>
<td>4.97e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HNF-1</td>
<td>10</td>
<td>3.353</td>
</tr>
<tr>
<td>2.93e-01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### B. Liver-specific (20 input; 12 analyzed)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Z-score</th>
<th>Fisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRF</td>
<td>1</td>
<td>38.21</td>
</tr>
<tr>
<td>8.83e-08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HNF-1</td>
<td>2</td>
<td>11.00</td>
</tr>
<tr>
<td>9.50e-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sox-5</td>
<td>3</td>
<td>9.822</td>
</tr>
<tr>
<td>1.22e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FREAC-4</td>
<td>4</td>
<td>7.101</td>
</tr>
<tr>
<td>1.60e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HNF-3beta</td>
<td>5</td>
<td>4.494</td>
</tr>
<tr>
<td>4.66e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOX17</td>
<td>6</td>
<td>4.229</td>
</tr>
<tr>
<td>4.20e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yin-Yang</td>
<td>7</td>
<td>4.070</td>
</tr>
<tr>
<td>1.16e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S8</td>
<td>8</td>
<td>3.821</td>
</tr>
<tr>
<td>1.61e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S8</td>
<td>8</td>
<td>3.821</td>
</tr>
<tr>
<td>1.61e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HNF-1</td>
<td>10</td>
<td>3.286</td>
</tr>
<tr>
<td>2.97e-01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TFs with experimentally-verified sites in the reference sets.
Structurally-related TFs with Indistinguishable TFBS

- Most structurally related TFs bind to highly similar DNA motifs
  - Zn-finger family is a big exception
EXAMPLE

Ets Family

- EG232974
- EG432800
- Ehf
- Elf1
- Elf2
- Elf3
- Elf4
- Elf5
- Elk1
- Elk3
- Elk4
- Erf
- Erg
- Ets1
- Ets2
- Etv1
- Etv2
- Etv3
- Etv3l
- Etv4
- Etv5
- Etv6
- Fev
- Fli1
- Gabpa
- LOC100
- LOC100
- factor)
- LOC634494
- Sfpi1
- Spdef
- Spib
- Spic

- How to pick which one?
  - At this stage there are TF catalogs coming that will be coupled to characteristics.
- Candidate gene prioritization software can be used (if not tightly coupled to chromosomal region) such as TOPPGENE

Overview

Part 1: Overview of transcription

Part 2: Prediction of transcription factor binding sites using binding profiles (“Discrimination”)

Part 3: Detection of novel motifs (TFBS) over-represented in regulatory regions

Part 4: Interrogation of sets of co-expressed genes or ChIP-seq regions to identify mediating transcription factors

Part 5: Gene regulatory networks
Part 5
“Gene Regulation Networks”

Composite scoring using multiple data classes

• Goal: Predict active regulatory regions in a given cell or tissue based on the integrated analysis of diverse genome-scale data such as epigenomic marks, chromatin accessibility, TF binding.

Example:
• Segway [http://pmgenomics.ca/proj/segway/](http://pmgenomics.ca/proj/segway/)
How does Segway work?

• Goal is to segment the genome into sub-classes. In each case, the tools identify certain data properties that subsets of the genome. Based on the groupings, the presence of known features (e.g. transcription start regions) is scored to suggest the functional meaning of the classes established.

• Require specific data for a cell type or tissue, making them well-suited for ENCODE target cells. Unclear how long until large-scale data available beyond ENCODE cells.

GREAT

• GREAT predicts functions of cis-regulatory regions.
• Many coding genes are well annotated with their biological functions. Non-coding regions typically lack such annotation. GREAT assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes. Thus, it is particularly useful in studying functions of sets of non-coding genomic regions.
GREAT input

• Takes as input a BED file indicating regions of genome

• Gives as output multiple enrichment measures of attributes associated with either the genomic regions themselves, or the proximal genes

Sample GREAT output
TF interactions

- TF-TF interactions provide greater diversity of potential patterns of expression.
- Much work has focused on determining which sets of TFBS tend to co-occur (e.g. oPOSSUM Anchor Site Analysis)
- The precise spatial patterns between TFBS may reveal more reliable relationships
  - SPAMO within the MEME Suite provides this function
**SPAMO spacing analysis**

Spacings of "UP000017 (Sox8 secondary)" relative to "Stat1"

Spacings of "MA00854.2 (H3P)" relative to "Stat1"

### Big challenges ahead

- Understanding all TFs across a developing organism
- Genetic variation in TFBS
- Integration of data sources
- Transition from matrices to HMMs or energy models
Reflections

- **Part 1**
  - Futility conjecture – Essentially predictions of individual TFBS have no relationship to an *in vivo* function
  - Successful bioinformatics methods for site discrimination incorporate additional information (clusters, conservation)
- **Part 2**
  - TFBS enrichment is a powerful new means to identify TFs likely to contribute to observed patterns of co-expression
- **Part 3**
  - Pattern discovery methods are severely restricted by the signal-to-noise problem
  - Successful methods for pattern discovery will have to incorporate additional information (ChIP-seq, conservation, structural constraints on TFs)
We are on a Coffee Break & Networking Session