Cis-Regulatory module network

Swathi Ashok Kumar

-Hardison Lab

05/01/2007
Gene Regulation

http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/P/Promoter.html
Types of regulatory elements

Figure 1
Schematic of a typical gene regulatory region. The promoter, which is

Human Gamma Globin gene
Cis regulatory elements
**Genes Co-expressed in Late Erythroid Maturation**

G1E-ER cells: proerythroblast line lacking the transcription factor GATA-1.
Can rescue by expressing an estrogen-responsive form of GATA-1
*Rylski et al., Mol Cell Biol. 2003*

---

**Transcriptome**

<table>
<thead>
<tr>
<th>Up</th>
<th>Down</th>
</tr>
</thead>
<tbody>
<tr>
<td>175 features</td>
<td>303 features</td>
</tr>
</tbody>
</table>

**Verify by RT-PCR**

<table>
<thead>
<tr>
<th>Days of induction</th>
<th>Hbb-b1</th>
<th>Alas2</th>
<th>Btg2</th>
<th>Vav2</th>
<th>Hist1h1c</th>
<th>Hipk2</th>
<th>Hep2</th>
<th>Zfpm1</th>
<th>Gata2</th>
<th>Gapdh</th>
</tr>
</thead>
</table>
Conservation of predicted binding sites for transcription factors

Binding site for GATA-1

<table>
<thead>
<tr>
<th>Conserved Consensus</th>
<th>Conserved NonConsensus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WGATAR</strong>&lt;br&gt;M TTAGATAAACCC&lt;br&gt;Rt CTAGATAACT&lt;br&gt;H GCAGATAGTT&lt;br&gt;Cp GCAGATAGTC&lt;br&gt;D ATTGATAACT</td>
<td><strong>WGATGR</strong>&lt;br&gt;M TTAGATGACC&lt;br&gt;Rt CTAGATGACT&lt;br&gt;H GCAGATGGTT&lt;br&gt;Cp GCAGATGGTC&lt;br&gt;D ATTGATGACT</td>
</tr>
</tbody>
</table>

cc | cnc
Regulatory Potential (RP) to distinguish functional classes

- A 3-way alignment has 124 types of columns. Collapse these to a smaller alphabet with characters $s$ (for example, 1-9).

- Train two order $t$ Markov models for the probability that $t$ alignment columns are followed by a particular column in training sets:
  - positive (alignments in known regulatory regions)
  - negative (alignments in ancestral repeats, a model for neutral DNA)
  - E.g. Frequency that 3 4 is followed by 5:
    - 0.001 in regulatory regions
    - 0.0001 in ancestral repeats

- RP of any 3-way alignment is the sum of the log likelihood ratios of finding the strings of alignment characters in known regulatory regions vs. ancestral repeats.

\[
RP = \sum_{a \text{ in segment}} \log \left( \frac{P_{\text{REG}}(S_a | S_{a-1} \ldots S_{a-t})}{P_{\text{AR}}(S_a | S_{a-1} \ldots S_{a-t})} \right)
\]
Chip-Chip
Putative transcriptional regulatory regions = pTRRs

- High likelihood hits for ChIP-chip
  - 5% false discovery rate
- Set of all sequence-specific factor hits that pass the threshold
  - 1332 pTRRs total
  - 996 pTRRs are supported by at least 2 ChIP-chip datasets
preCRMs with conserved consensus GATA-1 BS tend to be active on transfected plasmids
preCRMs with conserved consensus GATA-1 BS tend to be active after integration into a chromosome

Site-directed stable integration into MEL cells

Log2 (Fold Change relative to parent)

Day5 of induction

cc  cnc  neutral
preCRMs with High RP and Conserved Consensus GATA-1 Tend To Be Validated
Categories of Tested DNA Segments

Segments of mouse DNA with RP scores (align with rat, human, chimp, dog)

Positive RP
- preCRMcc
  - conserved consensus
  - GATA-1BS
  - 44 segments

- preCRMnc
  - conserved nonconsensus
  - GATA-1BS
  - 19 segments

Negative RP
- ccGATA1
  - 6 segments

- No conserved consensus GATA-1 binding site
  - preNeutral
  - 17 segments
Network description of the cis-Regulatory Modules

- Bipartite: \{cTFBS\} → \{pCRM\}
- Similarity Based: \{pCRM-1\} → \{pCRM-2\}

If they share a given number of cTFBS
Bi-partite
Similarity-based network
Positive pCRMs -- cTFBS

Blue Nodes: cTFBS
White Nodes: pCRM
Negative pCRMS—cTFBS graph

Red Nodes: cTFBS
Similarity Based Network (threshold = 4)

Node: pCRM
Edge: If two pCRMs share more than 4 cTFBS
Degree Distribution

Positive pCRM and cTFBS

Negative pCRM and cTFBS

Positive pCRM (threshold=4)

Negative pCRM
## Summary of the degree distribution

<table>
<thead>
<tr>
<th></th>
<th>Positive pCRM and CTFBS</th>
<th>Negative pCRM and cTFBS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Regression Coefficient</strong></td>
<td>0.911</td>
<td>0.856</td>
</tr>
<tr>
<td><strong>Power Law</strong></td>
<td>0.975</td>
<td>0.968</td>
</tr>
</tbody>
</table>
Future plans

- **Incorporate** Gene Expression data
- Use statistically enriched conserved motifs instead of cTFBS
Acknowledgements

- Hardison Lab members
- Nimblegen Corp
- Haussler Lab, UCSC
- Miller Lab, PSU
Stage 1: Reduced representations

A. Stage 1 first step: represent alignment columns as ancestral probability distributions

B. Stage 1 second step: create initial grouping (encoding) based on evolutionary similarity and frequency distribution

ESPERR: Evolutionary Sequence and Pattern Extraction using Reduced Representations

(colored circles represent groups of columns from clustering)
Stage 2: Improve encoding

C. Stage 2: search for best encoding based on classification rate:

1) Initialize from clustering
2) Generate candidate encodings with a random set of collapses and expansions
3) Encode training data with each candidate and evaluate with cross validation
4) Accept candidate with best performance
5) Iterate until stable
Train models for classification

D. Use final encoding on alignments for training and classification. Encoding symbols can be visualized with “logos”.

6 6 2 may occur frequently in positive training set and rarely in the negative training set, and thus contribute to discrimination.
If the positive training set is known regulatory regions, this would contribute to a positive RP.

Note that many different columns are reduced to single “encoding” (a number in the figure). E.g. Four different columns are each called “3”.
Using Galaxy to find predicted CRMs.
Example that suggests turnover

GATA-1 BSs

chr6:38702207–38702376 (170 bp)

mouse
non-rod.

RP

phastCons

Log norm. expression

Days HMBA
Additional methods find CACC box as distinctive for validation

### Background:
Mouse chr 19 (42.8% C+G) - NCBI Build 30

### CLOVER (Zlab)
- EKLF PWM (Dr. Perkins)

### Hexamer Counting

#### ELPH (UMaryland)

<table>
<thead>
<tr>
<th>6-mer</th>
<th>TTATYT</th>
<th>GGCAGR</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-mer</td>
<td>CCWCAGM</td>
<td>RGRCAGR</td>
</tr>
<tr>
<td>8-mer</td>
<td>CASCCWGC</td>
<td>CAGGGAWR</td>
</tr>
<tr>
<td>9-mer</td>
<td>CCWGGCWGM</td>
<td>CWGRGAWRA</td>
</tr>
</tbody>
</table>

### Output for validated preCRMs
- Motif: EKLF
- Probability: 0.0008

### Output for nonvalidated preCRMs
- Motif: none
- Probability: none

### Expected Counts

<table>
<thead>
<tr>
<th>Motif</th>
<th>Validated</th>
<th>Non-Validated</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCACCC</td>
<td>60</td>
<td>32</td>
</tr>
<tr>
<td>CACCCW</td>
<td>56</td>
<td>27</td>
</tr>
<tr>
<td>NCACCC</td>
<td>16.31</td>
<td>5.81</td>
</tr>
<tr>
<td>CACCCW</td>
<td>11.74</td>
<td>4.36</td>
</tr>
</tbody>
</table>

Additional methods find CACC box as distinctive for validation.
Alignability is a more sensitive method for detecting constraint than overlap with MCSs.

pTRR = putative transcriptional regulatory region