Bioinformatics Approaches for Studying Transcription Regulation and Protein-DNA Interactions

Xiaole Shirley Liu
8/11/2003

Outline

• Biology of transcription regulation
• Scan for known TF motif sites
• De novo method
  – Regular expression enumeration
  – Position weight matrix update
  – Using microarray to guide motif search
• Practical issues in motif finding
  – Lower organisms
  – Higher eukaryotes

Biology of Transcription Regulation

Motif can only be computational discovered when there are enough cases for machine learning
Goal of Understanding Regulation

- Which TFs are involved in the regulation?
- What are the binding motifs of these TFs?
- Does a TF enhance / repress gene expression?
- Which genes are regulated by this TF?
- Are there binding partner / competitor for a TF?
- Why there is disease when a TF went wrong?

<table>
<thead>
<tr>
<th>Gene Expression Profile Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Gene Expression" /></td>
</tr>
</tbody>
</table>

**Upstream Regions**

- GATGCTGCACCATGGTATCG
- CACATGGCTACGGCTGGCG
- AGCTGGAGCTGCGAGCTG
- CGCTAGCTGGCTG

**Co-expressed Genes**

- Pho 5
- Pho 8
- Pho 81
- Pho 84

**Transcription Start Site (TSS)**

**Gene Expression Profile Clusters**

- ![Gene Expression](image4) | ![Profile Cluster](image5) | ![Upstream Motif Finding](image6) |

**Upstream Regions**

- GATGCTGCACCATGGTATCG
- CACATGGCTACGGCTGGCG
- AGCTGGAGCTGCGAGCTG

**Co-expressed Genes**

- Pho 4 binding
ChIP-chip Experiments

• Chromatin immunoprecipitation + microarray
  (Chromatin IP, ChIP array, ChIP chip)

• Detects in vivo protein DNA interaction

- Crosslink protein-DNA interaction
- Shear DNA
- Immuno-precipitation
- Hybridize with microarray
  (intergenic sequence probes)

Computational Motif Finding

• Input data:
  – Upstream sequences of gene expression profile cluster
    or ChIP-chip selected sequences
  – 20-800 sequences, each 300-5000 bps long

• Output: enriched sequence patterns (motifs)

• Challenges:
  – False positive sequences, variable sites / sequence
  – Motif sites have substitution from consensus
  – Many non-functional repeats
  – Many motifs may be involved

Scan for Known TF Motif Sites

• TRANSFAC database: experimental TF sites

• Motif representation:
  – Regular expression: Consensus CACAAAA
  – Degenerate CRCAAAW
  – IUPAC A/G A/T
Scan for Known TF Motif Sites

- **TRANSFAC database:** experimental TF sites
- **Motif representation:**
  - Regular expression: Consensus CACAAAA
  - Degenerate CRCAAAW
  - Position weight matrix (PWM):

```
Pos  A  G  C  T  Cnt
1    0.1 0.8 0 0.1 12345678
2    0.7 0.1 0.1 0.6 43215678
3    0.2 0.2 0.7 0.1 43215678
4    0.6 0.4 0.2 0.1 43215678
5    0.1 0.3 0.1 0.7 43215678
6    0.8 0.2 0.1 0.1 43215678
7    0.2 0.1 0.8 0.1 43215678

Motif Matrix
```

De novo Sequence Motif Finding

- **Goal:** look for common sequence patterns enriched in the input data (compared to the genome background)
- **Regular expression enumeration**
  - Pattern driven approach
  - Enumerate patterns, check significance in dataset
- **Position weight matrix update**
  - Data driven approach
  - Initialize random matrices, use dataset to refine
- **Using microarray measures to guide motif search**
  - Motif occurrence best correlated with expression

Regular Expression Enumeration

- **For every oligonucleotide w check for over-representation:**
  - Expected w occurrence in data
    - Consider genome sequence + current data size
  - Observed w occurrence in data
  - Over-represented w is potential TF binding motif
- **Exhaustive, guaranteed to find global optimum, and can find multiple motifs**
- **Not as flexible with base substitutions, long list of similar good motifs, and limited with motif width**
RE Enumeration Derivatives

- oligo-analysis, spaced dyads \( w_1 H_{w_2} \) (J van Helden)
- IUPAC alphabet, Markov background (M Tompa)
- Suffix trie (A Brazma) to represent all intergenic sequences, prune nodes with two few sites
- 2-bit encoding (P Baldi), fast index access
- WINNOWER (P Pevzner), find cliques in a graph
- MobyDick (H Bussemaker), build long motifs from shorter ones

De novo Sequence Motif Finding

- Goal: look for common sequence patterns enriched in the input data (compared to the genome background)
- Regular expression enumeration
  - Pattern driven approach
  - Enumerate patterns, check significance in dataset
- Position weight matrix update
  - Data driven approach
  - Initialize random matrices, use dataset to refine
- Using microarray measures to guide motif search
  - Motif occurrence best correlated with expression

Consensus

- Starting from the 1st sequence, add one sequence at a time to look for the best motifs obtained with the additional sequence

\[
\text{Seq1} \quad \text{Seq2} \quad \ldots
\]
Consensus

- Starting from the 1st sequence, add one sequence at a time to look for the best motifs obtained with the additional sequence

- G Stormo, algorithm runs very fast

- Sequence order plays a big role in performance
  - First two sequences better contain the motif
  - Sites stop accumulating at the first bad sequence
  - Newer version allowing [0-n] is much slower

Expectation Maximization and Gibbs Sampling Model

- Objects:
  - Seq: sequence data to search for motif
  - $\theta_0$: non-motif (genome background) probability
  - $\Theta$: motif probability matrix parameter
  - $\pi$: unknown variable, site locations
- Problem: $P(\theta, \pi \mid seq, \theta_0)$
- Approach: alternately estimate
  - $\pi$ by $P(\pi \mid \theta, seq, \theta_0)$
  - $\theta$ by $P(\theta \mid \pi, seq, \theta_0)$
- EM and Gibbs differ in the estimation methods
Expectation Maximization

- E step: $\pi | \theta, \text{seq}, \theta_0$
  
  $P_1 = \text{likelihood ratio} = \frac{P(TTGAC|\theta)}{P(TTGAC|\theta_0)}$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pos 1</td>
<td>1.0 0.7 0.3 0.1 0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pos 2</td>
<td>0.9 0.01 0.05 0.1 0.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pos 3</td>
<td>0.8 0.02 0.22 0.1 0.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pos 4</td>
<td>0.3 0.3 0.1 0.4 0.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pos 5</td>
<td>0.2 0.4 0.1 0.2 0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$P_1 = P(TTGAC|\theta) = 0.3 \times 0.3 \times 0.2 \times 0.3 \times 0.2$

- M step: $\theta | \pi, \text{seq}, \theta_0$

  $\arg\max_\theta p(\pi | \theta, \text{seq}, \theta_0)$

$\theta$ reflects weighted average of $\pi$

EM Derivatives

- First EM motif finder (C Lawrence)
  - Deterministic algorithm, guarantee local optimum

- Random projection (J. Buhler)
  - Sample h / w positions, hash words agreeing at h into bucket
  - Run EM on buckets with enough size

- MEME (TL Bailey)
  - Prior probability allows 0-n site / sequence
  - Parallel running multiple EM with different seed
  - User friendly results
Gibbs Sampling

- Stochastic process, although still may need multiple initializations
  - Sample $\pi$ from $P(\pi | \theta, \text{seq}, \theta_0)$
  - Sample $\theta$ from $P(\theta | \pi, \text{seq}, \theta_0)$

- Collapsed form:
  - $\theta$ estimated with counts, not sampling from Dirichlet
  - Sample site from one seq based on sites from other seqs

- Converged motif matrix $\theta$ and converged motif sites $\pi$ represent stationary distribution of a Markov Chain

Gibbs Sampler

- Randomly initialize a probability matrix

$$P_{AI} = \frac{n_{AI} + s_i}{a_{AI} + s_A + a_{AT} + s_T + a_{AC} + s_C + a_{AG} + s_G + a_{AT} + s_T}$$

$\theta$ estimated with counts

Initial $\theta_1$

- Take out one sequence with its sites from current motif

$\theta_1$ Without $\pi_{11}$ Segment
Gibbs Sampler

- Score each possible segment of this sequence

Segment (1-6): 1.5

θ₁ Without π₁₁ Segment

Gibbs Sampler

- Score each possible segment of this sequence

Segment (2-7): 3

θ₁ Without π₁₁ Segment

Gibbs Sampler

- Sample site from one seq based on sites from other seqs

Modified θ₁
θ estimated with counts
Gibbs Sampler

• Repeat the process until motif converges

\[ \pi_1 \]

\[ \pi_2 \]

\[ \pi_3 \]

\[ \pi_4 \]

\[ \pi_5 \]

\[ \theta \]

Without

\[ \pi_1 \]

\[ \pi_2 \]

\[ \pi_3 \]

\[ \pi_4 \]

\[ \pi_5 \]

Segment

--

Gibbs Sampler

• Column shift

\[ \pi_1 \]

\[ \pi_2 \]

\[ \pi_3 \]

\[ \pi_4 \]

\[ \pi_5 \]

• Metropolis algorithm:
  – Propose \( \pi^* \) as \( \pi \) shifted 1 column to left or right
  – Calculate motif score \( u(\pi) \) and \( u(\pi^*) \)
  – Accept \( \pi^* \) with prob = \( \min(1, \frac{u(\pi^*)}{u(\pi)}) \)

Gibbs Sampling Derivatives

• Gibbs Motif Sampler (JS Liu)
  – Add prior probability to allow 0-n site / seq
  – Sample motif positions to consider

• AlignACE (GM Church)
  – Mask out one motif to find more different motifs

• BioProspector (XS Liu)
  – Use background model with Markov dependencies
  – Sampling with threshold (0-n sites / seq), new scoring function
  – Can find two-block motifs with variable gap
**Position Weight Matrix Update**

- **Advantage**
  - Can look for motifs of any widths
  - Flexible with base substitutions

- **Disadvantage:**
  - No guaranteed global optimum

---

**Using microarray measures to guide motif search**

- Motif occurrence best correlated with variations in gene expression

- REDUCE
- GMEP
- MDscan
- Motif Regressor

---

**REDUCER**

- Incorporating TFBS copy number with microarray values (H Bussemaker)
  - Single microarray experiment (no clustering)
  - Enumerate all possible $w$-mers
  - Check $w$-mer copy number in each upstream seq
  - Check downstream expression of every gene
  - See whether there is any correlation
Example

<table>
<thead>
<tr>
<th>Copy #</th>
<th>Expr</th>
<th>Copy #</th>
<th>Expr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq1</td>
<td>5</td>
<td>Seq1</td>
<td>5</td>
</tr>
<tr>
<td>Seq2</td>
<td>0</td>
<td>Seq2</td>
<td>0</td>
</tr>
<tr>
<td>Seq3</td>
<td>4</td>
<td>Seq3</td>
<td>4</td>
</tr>
<tr>
<td>Seq4</td>
<td>2</td>
<td>Seq4</td>
<td>2</td>
</tr>
<tr>
<td>Seq5</td>
<td>0</td>
<td>Seq5</td>
<td>0</td>
</tr>
<tr>
<td>Seq6</td>
<td>1</td>
<td>Seq6</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

REDUCER

- Incorporating TFBS copy number with microarray values (H Bussemaker)
  - Enumerate all possible $w$-mers
  - Check $w$-mer copy number in each upstream seq
  - Check downstream expression of every gene
  - See whether there is any correlation
    - More upstream sites, more expression $\Rightarrow$ inducer
    - More upstream sites, less expression $\Rightarrow$ repressor
- Exhaustive, multiple motifs, global optimum
- Many similar motifs, limited in motif width

Genome Mean Expression Profiles

- M Eisen, single microarray experiment
- For each $w$-mer, find the $G$ genes whose upstream contain the $w$-mer and calculate GMEP
  \[
  GMEP(w) = \sum_{g=1}^{G} N_{wg} \cdot E_g / \sum_{g=1}^{G} N_{wg}
  \]
- Randomly pick $G$ genes, calculate their expression distribution $N(\mu, \sigma)$
- Report $w$-mer as potential motif if GMEP is extremely distributed on $N(\mu, \sigma)$
MDscan

• ChIP dip results insights:
  – High ChIP ranking => true targets
  – Highest ChIP ranking => contain more sites

• Basic strategy:
  – Search TF motif from highest ranking targets first
    (high signal / background ratio)
  – Refine candidate motifs with all targets

• Deterministic algorithm, very fast (XS Liu)

Similarity defined by $m$-match

For a given $w$-mer and any other random $w$-mer:

<table>
<thead>
<tr>
<th>$w$-mer</th>
<th>$m$-matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGTAACGT</td>
<td>8-mers</td>
</tr>
<tr>
<td>TGTTAAGCT</td>
<td>matched 8</td>
</tr>
<tr>
<td>AGTTAAGCT</td>
<td>matched 7</td>
</tr>
<tr>
<td>TGCAACAT</td>
<td>matched 6</td>
</tr>
<tr>
<td>TGACAGG</td>
<td>matched 5</td>
</tr>
<tr>
<td>AATACAG</td>
<td>matched 4</td>
</tr>
</tbody>
</table>

Pick a reasonable $m$ to call two $w$-mers similar

<table>
<thead>
<tr>
<th>$w$</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

MDscan Algorithm:
Finding candidate motifs

Extreme
High
Rank

Seed1

m-matches

All ChIP selected targets
MDscan Algorithm: Finding candidate motifs

Seed2

m-matches

Extreme
High
Rank

All ChIP selected targets

MDscan Algorithm: Scoring candidate motifs

- Motif scoring function:

\[
\frac{\log(x_m)}{w} \times \left[ \sum_{i=1}^{w} \sum_{j=1}^{x_i} \log p_j \log p_j \left( \frac{1}{x_m} \sum_{x=1}^{x_m} \log(p_x(x)) \right) \right]
\]

Motif Signal
Positions
Specific (unlikely in genome background)

Prefer: conserved motifs with many sites, but are not often seen in the genome background
Keep best 30 candidate motifs

MDscan Algorithm: Update motifs with remaining seqs

Seed1

m-matches

Extreme
High
Rank

All ChIP selected targets
MDscan Algorithm:
Update motifs with remaining seqs

Seed2
m-matches

Extreme
High
Rank

All ChIP selected targets

MDscan Algorithm:
Refine the motifs

Seed1
m-matches

Extreme
High
Rank

All ChIP selected targets

Motif Regressor

- EM Conlon

Look for candidate motifs
Refine motifs
MDscan

Genes

Regress b/t upstream mtf match score and downstream expression
Motif Regressor Rational

- For each TF:
  - Upstream Seq Mtf Match Gene Exp
  - Gene1: 3.2 1.8
  - Gene2: 2.8 0.3
  - Gene3...

- Upstream sequence X motif matching score measures:
  - Number of sites
  - Strength of matching

Motif Regressor Strategy

- Rank genes by log2 (expression fold change)
- Try MDscan (width 5-17) on induced and repressed genes separately
  - Find 50 candidate motifs from top 100 genes
  - Refine candidate motifs with top 500 genes
  - Report <= 30 distinct motifs
- Score each upstream sequence with each motif
- Linear regression to eliminate insignificant motifs

Linear Regression Example

<table>
<thead>
<tr>
<th>Person</th>
<th>IQ</th>
<th>Age</th>
<th>Education</th>
<th>Height</th>
<th>Eye color</th>
<th>Spend/week</th>
<th># of CD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>120</td>
<td>30</td>
<td>High</td>
<td>175</td>
<td>blue</td>
<td>$4000</td>
<td>30</td>
</tr>
<tr>
<td>B</td>
<td>250</td>
<td>41</td>
<td>PhD</td>
<td>155</td>
<td>brown</td>
<td>$1500</td>
<td>32</td>
</tr>
<tr>
<td>C</td>
<td>150</td>
<td>8</td>
<td>Grade10</td>
<td>115</td>
<td>black</td>
<td>$1000</td>
<td>30</td>
</tr>
<tr>
<td>D</td>
<td>180</td>
<td>16</td>
<td>Grade12</td>
<td>140</td>
<td>grey</td>
<td>$200</td>
<td>15</td>
</tr>
<tr>
<td>E</td>
<td>98</td>
<td>4</td>
<td>Preschool</td>
<td>88</td>
<td>green</td>
<td>$500</td>
<td>24</td>
</tr>
<tr>
<td>F</td>
<td>130</td>
<td>17</td>
<td>High</td>
<td>170</td>
<td>black</td>
<td>$500</td>
<td>300</td>
</tr>
<tr>
<td>G</td>
<td>110</td>
<td>21</td>
<td>Collage</td>
<td>182</td>
<td>blue</td>
<td>$800</td>
<td>220</td>
</tr>
</tbody>
</table>

- Gene Expression
  - Mtf1 Mtf2 Mtf3 Mtf4 Mtf5 Mtf6

- Single Regression
Motif Regressor Strategy

- Stepwise regression to find multiple motifs that work together
  - Each step find the motif that explains the remaining expression the best
  - Remove its effects from expression
- Multiple regression model: expression explained as the sum of motifs' effects

\[
Y_g = \alpha + \sum_{m=1}^{M} \beta_m S_{mg} + \epsilon_g
\]

Expression of gene \(g\)
Baseline expression
Regression coefficient/ motif match score

Stepwise Regression Example

<table>
<thead>
<tr>
<th>Person</th>
<th>IQ</th>
<th>Age</th>
<th>Education</th>
<th>Height</th>
<th>Eye color</th>
<th>Spend/Week</th>
<th># of CD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>120</td>
<td>30</td>
<td>High</td>
<td>175</td>
<td>blue</td>
<td>$6000</td>
<td>34</td>
</tr>
<tr>
<td>B</td>
<td>250</td>
<td>41</td>
<td>PhD</td>
<td>155</td>
<td>brown</td>
<td>$1500</td>
<td>18</td>
</tr>
<tr>
<td>C</td>
<td>150</td>
<td>8</td>
<td>Grade10</td>
<td>140</td>
<td>gray</td>
<td>$200</td>
<td>15</td>
</tr>
<tr>
<td>D</td>
<td>180</td>
<td>16</td>
<td>Grade12</td>
<td>150</td>
<td>green</td>
<td>$200</td>
<td>21</td>
</tr>
<tr>
<td>E</td>
<td>90</td>
<td>4</td>
<td>President</td>
<td>170</td>
<td>black</td>
<td>$500</td>
<td>26</td>
</tr>
<tr>
<td>F</td>
<td>130</td>
<td>17</td>
<td>High</td>
<td>178</td>
<td>black</td>
<td>$800</td>
<td>180</td>
</tr>
<tr>
<td>G</td>
<td>110</td>
<td>21</td>
<td>College</td>
<td>182</td>
<td>blue</td>
<td>$800</td>
<td>220</td>
</tr>
</tbody>
</table>

Gene | Expr | Mt1 | Mt2 | Mt3 | Mt4 | Mt5 | Mt6 |
--- | --- | --- | --- | --- | --- | --- | --- |
X   | X    | X   | --  | --  | --  | --  |

Regression

- Single
- Stepwise

Outline

- Biology of transcription regulation
- Scan for known TF motif sites
- De novo method
  - Regular expression enumeration
  - Position weight matrix update
  - Using microarray to guide motif search
- Practical issues in motif finding
  - Lower organisms
  - Higher eukaryotes
Motif Finding in Bacteria

• Promoter sequences are short (200–300 bp)
• Motif are usually very long (10–20 bases)
• There are many two-block motifs
  – Sigma factors motifs: two blocks with a variable gap
  – Many HTH proteins binds to palindrome motifs
• Long motifs are usually very degenerate
  – Often requires each sequence to contain ≥ one site
  – Adding orthologous sequences from other species can aid discovery of weak motifs

Motif Finding in Lower Eukaryotes

• Upstream sequences longer (800 bp), with some simple repeats
• Motif width varies (5–17 bases)
• Expression clusters provide decent input sequences quality for TF motif finding
• Motif combination appears, although single motifs are usually significant enough for identification

Motif Finding in Higher Eukaryotes

• Upstream sequences very long (3KB–20KB), TF motif could appear downstream
• Usually [–800, 200] TSS has the highest TF density, finding TSS is critical (RefSeq)
• Motifs are usually short (6–8 bases), and work in combination, so individual motif may not be significant
• Needs to run RepeatMasker to remove simple repeats
• Gene expression cluster not good enough input
Comparative Genomics

- TF sites across species are more conserved than random background due to functional constraints, comparative genomics can narrow down the search space
- Many genes across 2 species (WW Wasserman)
  - Align orthologous sequences (Vista, LAGAN, Pipmaker, Bayes block aligner)
  - get rid of sequence too mutated between species before running motif finding algorithms
- One gene across multiple species, phylogenetic foot printing
  - Zoo project (E Rubin)
  - Phylogenetic foot printer (M Tompa)

Motif Site Clusters

- Higher eukaryote TF motif sites appear in clusters
- Use scanning or de novo methods to find individual TF motifs
  - Check whether there are site combinations always occurring in close proximity (M Levine)
  - If known one motif, check sites appearing near the known motif sites, e.g. E2F motif
  - Consider comparative genomics as well

Summary

- Understanding transcription regulation is important
- Scan for known TF motif sites, TRANSFAC
- De novo method
  - Regular expression enumeration
  - Position weight matrix update (consensus, EM, Gibbs)
  - Using microarray to guide motif search (REDUCER, GMEP, MDscan, Motif Regressor)
- Practical issues in motif finding
  - Lower organisms
  - Higher eukaryotes (Comparative genomics, motif site clusters)
- Despite wide computational studies on transcription regulation, we are far from reaching the goal
- Questions: xsliu@jimmy.harvard.edu