A Biophysical Approach to Transcription Factor Binding Site Discovery

Anirvan Sengupta
Dept. of Physics and Astronomy
and
BioMaPS Institute
Rutgers University

Plan of the Talk

• Global Regulators and DNA-binding Specificity
• Predicting Binding Sites: Bioinformatics
• Experimental Efforts
Examples of Global Regulators in E. Coli

<table>
<thead>
<tr>
<th>Name of Protein</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crp/CAP</td>
<td>Hunger (cAMP) sensor</td>
</tr>
<tr>
<td>LexA</td>
<td>SOS response</td>
</tr>
<tr>
<td>Fnr</td>
<td>Oxygen sensing</td>
</tr>
<tr>
<td>Lrp</td>
<td>Leucine?? response</td>
</tr>
<tr>
<td>Various sigma factors</td>
<td>Response to different stresses</td>
</tr>
</tbody>
</table>

Textbook Example: Lac Operon
Some Known Targets of lacI and of crp

lacI binding sites:

\begin{verbatim}
AATTGTGAGCGGATAAACAAATT
AAATGTGAGCGGATAAACACCC
GCGAGTGAACCCAGCAAGGAAATT
\end{verbatim}

some crp binding sites:

\begin{verbatim}
TAATGTGACGTCTTTGCACTAC
GAAGCGACCTGGGTACGCA
GGTGTTAAATTGACAGTTTC
GATGCAGGCCAGCTCAAAA
AAAATTCCTATTATCTACCTT

TTTGCAGTCAAAATACACTT
AAACGATCAACCCCTCAATT
TAATGTGATAGCTACACTCAT
AATCGTAGCGATAACAAATT

Consensus:
AAATGTGATCTAGATCACATT
\end{verbatim}

Describing Fuzzy Motifs

IUPAC way:
- A, C, G, T
- R (A or G) from pUrine
- Y (C or T) from pYrimidine
- W (A or T) from Weak
- S (G or C) from Strong
- M (A or C) from aMino
- K (G or T) from Keto
- B (C or G or T) from not-A
- D (A or G or T) from not-C
- H (A or C or T) from not-G
- V (A or C or G) from not-T
- N from aNy

For CRP, we have to look for
WWWNTGCAGNNNNNNNTCTCANDWW
or something less stringent
WWWNYKDYVNNNNNNBHRMRNW
Biophysical Approach to Transcription Factor Binding Site Discovery

Weight Matrix

[Less Berg, von Hippel, Studen, Stormo, …]

Given a set of known factor binding motifs, like,

TAATGTGACGTCCCTTTGCATA
GAAGCGACCTGGTCATGCTG
CGATGCAGGGCAGATCGAAAA
……

ATTTGAACCAGATCGATT
AAATGTAAGCTGTGCCACGT

construct a frequency matrix $n_{ib}$

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>……</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>……</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>……</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>……</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>……</td>
<td>2</td>
</tr>
</tbody>
</table>

Weight Matrix Continued...

Calculate weights by taking logarithm: $w_{ib} = \frac{\log(n_{ib} / n_s)}{

For any sequence $S$, the score $W$ is given by: $W = \sum_{ib} w_{ib} S_{ib}$

For example:

$W(TTAGCA….)= w_{IT} + w_{2T} + w_{3A} + w_{4G} + w_{5C} + w_{6A} + ……$

Sequences with higher $W$ are better binders.
Precise relationship with binding energy in certain limits.
Compromise

Robison, McGuire, Church

Problem of Threshold Selection
A Model for Transcription Factor Binding DNA

Independent Nucleotide Model for Binding Energy
(Berg, vonHippel, Stormo, …)

The energy is the sum of independent contributions from bases.

\[ E(TTAGCAA) = \varepsilon_{1T} + \varepsilon_{2T} + \varepsilon_{3A} + \varepsilon_{4G} + \varepsilon_{5C} + \varepsilon_{6A} + \varepsilon_{7A} = \sum_{ib} \varepsilon_{ib} S_{ib} = \varepsilon S \]
Biophysical Approach to Transcription Factor Binding Site Discovery

**Binding Probability**

\[ f(E(s)) = \frac{1}{\exp(\beta(E(S) - \mu)) + 1} = \frac{K \exp(-\beta E(S)n)}{1 + K \exp(-\beta E(S)n)} \]

- **Finite Temperature**
- **Zero Temperature**

**The Probability Model for Data:**

**Low Stringency SELEX**

Dr. Anirvan Sengupta, Rutgers Univ (KITP Bio Networks 2/25/03)
From Sequences to Energies

Maximum Likelihood Method

\[ e^L = \prod_{S \in O} (p_f(E(S))) \prod_{S' \in \bar{O}} (1 - p_f(E(S'))) \]

Leads to minimization of

\[ -L \approx -n_s \ln(p) - \sum_{S \in O} \ln(f(E(S))) + p \int dED(E) f(E) \]

Simplifies in the ‘low temperature’ limit

Temperature<< Variation of Energies

Leads to Minimization of

\[ -\sum_{S \in O} \ln(\Theta(\mu - E(S))) + p \int_{-\infty}^{\mu} dED(E) \]

The first term forces binding energy of sample sequences to be less than the chemical potential.

The second term forces the number of random strings that bind to be minimum.
Increasing Width of D(E) increases number of ‘False Positives’/ Random Background

Minimize the Variance!

Quadratic Programming Method for Energy Parameter Estimation

Minimize variance $\varepsilon^2$
Subject to constraints
$E(S_a) = \varepsilon, S_a < \mu = -1$
for each example $a$.

Solvable by Quadratic Programming.
Similar to Support Vector Machine (SVM) pattern finder.
Support Vectors Machines

S1 and S2 supports the separating hyperplane.

Low Concentration Limit: Weight Matrix Method

\[ \mu \to -\infty, \quad f(E(S)) \to e^{\beta \mu} e^{-\beta E(S)} \]

Maximum likelihood estimate of parameters leads to weight matrix formula.
Comparison to Weight Matrix based Search

Training set for 55 E. Coli. transcription factors from DPInteract Database (G. Church, Harvard Medical School).

Get additional sites from RegulonDB (Collado-Vides, UNAM)

Compare the success of our method with that of Weight Matrix Method:

Problem: No natural threshold choice for weight matrix.

Resolution: Compare for more than one thresholds.

False Negatives vs List Size

![Graph showing False Negatives vs List Size](image)
Biophysical Approach to Transcription Factor Binding Site Discovery

Statistical Significance

Table 1: Statistical summary of E. coli search results (see website for details)

<table>
<thead>
<tr>
<th>Name</th>
<th>Length</th>
<th>Number of examples</th>
<th>Weight matrix</th>
<th>QMF</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArcA</td>
<td>15</td>
<td>14</td>
<td>392</td>
<td>52</td>
<td>6.3</td>
</tr>
<tr>
<td>ArcR</td>
<td>18</td>
<td>17</td>
<td>320</td>
<td>79</td>
<td>8.9</td>
</tr>
<tr>
<td>CRP</td>
<td>25</td>
<td>48</td>
<td>3850</td>
<td>766</td>
<td>27.3</td>
</tr>
<tr>
<td>CytR</td>
<td>18</td>
<td>5</td>
<td>740</td>
<td>42</td>
<td>2.3</td>
</tr>
<tr>
<td>DnaA</td>
<td>15</td>
<td>8</td>
<td>08748</td>
<td>465</td>
<td>0.5</td>
</tr>
<tr>
<td>FabR</td>
<td>17</td>
<td>7</td>
<td>28</td>
<td>18</td>
<td>9.0</td>
</tr>
<tr>
<td>FisR</td>
<td>10</td>
<td>4</td>
<td>1880</td>
<td>241</td>
<td>3.7</td>
</tr>
<tr>
<td>Fis</td>
<td>35</td>
<td>19</td>
<td>7867</td>
<td>259</td>
<td>4.1</td>
</tr>
<tr>
<td>Fnu</td>
<td>22</td>
<td>14</td>
<td>174</td>
<td>36</td>
<td>19.9</td>
</tr>
<tr>
<td>FovR</td>
<td>16</td>
<td>12</td>
<td>35</td>
<td>28</td>
<td>14.8</td>
</tr>
<tr>
<td>CapH</td>
<td>16</td>
<td>7</td>
<td>20</td>
<td>3</td>
<td>2.9</td>
</tr>
<tr>
<td>CapA</td>
<td>30</td>
<td>13</td>
<td>9130</td>
<td>192</td>
<td>1.6</td>
</tr>
<tr>
<td>H-NS</td>
<td>11</td>
<td>15</td>
<td>3430</td>
<td>2340</td>
<td>2.7</td>
</tr>
<tr>
<td>IGF</td>
<td>48</td>
<td>28</td>
<td>82404</td>
<td>393</td>
<td>13.6</td>
</tr>
<tr>
<td>LexA</td>
<td>20</td>
<td>19</td>
<td>30</td>
<td>39</td>
<td>23.9</td>
</tr>
<tr>
<td>LuxF</td>
<td>25</td>
<td>12</td>
<td>05676</td>
<td>2487</td>
<td>32.9</td>
</tr>
<tr>
<td>MalA</td>
<td>16</td>
<td>10</td>
<td>36</td>
<td>71</td>
<td>8.7</td>
</tr>
<tr>
<td>MetJ</td>
<td>16</td>
<td>15</td>
<td>401</td>
<td>45</td>
<td>1.6</td>
</tr>
<tr>
<td>MetJ</td>
<td>15</td>
<td>8</td>
<td>344</td>
<td>58</td>
<td>0.2</td>
</tr>
<tr>
<td>MagG</td>
<td>23</td>
<td>6</td>
<td>72</td>
<td>8</td>
<td>7.0</td>
</tr>
<tr>
<td>NapL</td>
<td>16</td>
<td>11</td>
<td>2800</td>
<td>19</td>
<td>7.5</td>
</tr>
<tr>
<td>CytR</td>
<td>10</td>
<td>9</td>
<td>4600</td>
<td>94</td>
<td>3.6</td>
</tr>
<tr>
<td>FloR</td>
<td>22</td>
<td>15</td>
<td>258</td>
<td>28</td>
<td>14.6</td>
</tr>
<tr>
<td>FurR</td>
<td>26</td>
<td>22</td>
<td>47</td>
<td>28</td>
<td>37.0</td>
</tr>
<tr>
<td>s**</td>
<td>15</td>
<td>27</td>
<td>31937</td>
<td>639</td>
<td>2.3</td>
</tr>
<tr>
<td>s**</td>
<td>16</td>
<td>48</td>
<td>30087</td>
<td>922</td>
<td>2.6</td>
</tr>
<tr>
<td>s**</td>
<td>17</td>
<td>51</td>
<td>34988</td>
<td>3920</td>
<td>0.6</td>
</tr>
<tr>
<td>s**</td>
<td>30</td>
<td>34</td>
<td>30130</td>
<td>381</td>
<td>0.6</td>
</tr>
<tr>
<td>s**</td>
<td>10</td>
<td>26</td>
<td>35996</td>
<td>351</td>
<td>0.4</td>
</tr>
<tr>
<td>s**</td>
<td>18</td>
<td>18</td>
<td>640</td>
<td>7</td>
<td>6.0</td>
</tr>
<tr>
<td>s**</td>
<td>25</td>
<td>15</td>
<td>30660</td>
<td>240</td>
<td>2.0</td>
</tr>
<tr>
<td>SsrA</td>
<td>35</td>
<td>14</td>
<td>3560</td>
<td>49</td>
<td>2.2</td>
</tr>
<tr>
<td>CytR</td>
<td>22</td>
<td>17</td>
<td>3840</td>
<td>78</td>
<td>6.2</td>
</tr>
</tbody>
</table>

Dr. Anirvan Sengupta, Rutgers Univ (KITP Bio Networks 2/25/03) 12
Statistics of $\sigma^{70}$ Hits: Orientation Dependence

False Positives: Experiments

- Checking examples by gel-retardation experiments
- Effort to do genome-wide location analysis for some pleiotropic factors in *E. Coli.*
- Low stringency SELEX experiments.
Some of Predicted Binding Sites

<table>
<thead>
<tr>
<th>Site</th>
<th>Score</th>
<th>Location</th>
<th>Score</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAGTGTGACCCGTTTCACGTAG</td>
<td>1.23</td>
<td>1402612 in between 1401279 1402604 -1; 1402765 1403673 +1;</td>
<td>TTATGGAAGAGATACATTTT</td>
<td>1.04</td>
</tr>
<tr>
<td>TCCTGCGCTTGGCTGACAATC</td>
<td>1.05</td>
<td>1846106 in between 1844909 1846032 -1; 1846149 1846700 -1;</td>
<td>TAACCGGATCCGCTCAAAAT</td>
<td>1.16</td>
</tr>
<tr>
<td>TATCGAGATAACGATCACAAAA</td>
<td>1.17</td>
<td>2175284 in between 2174370 2175230 -1; 2175532 2176656 -1;</td>
<td>CAAATTTGCTTACATCTCTTAA</td>
<td>1.04</td>
</tr>
</tbody>
</table>

Gel-shift Experiments

Dr. Anirvan Sengupta, Rutgers Univ (KITP Bio Networks 2/25/03)
Summary

- Low DNA-binding specificity for global regulators and need to quantify variability.
- New bio-informatic tool for binding site prediction with a built-in threshold.
- Preliminary experimental results encouraging.

Collaborators

Computation: Boris Shraiman (Rutgers)
Marco Djordjevic (Columbia)

Experiments: Viji Nagaraj (Rutgers)
Boris Shraiman (Rutgers)
Richard Ebright (Rutgers)