SVMs and Probabilistic Approaches for Classifying Promoters

Anirvan M. Sengupta
Physics Dept./ BioMaPS Inst.
Rutgers University
Plan of the Talk

• The problem
• “Philosophical” issues
• Probability models and support vector machines
• Issues in combining heterogeneous input data
Transcriptional Regulation: Promoters as Computing Devices

**E. coli**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Gene Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLUCOSE + LACTOSE</td>
<td>OPERON OFF</td>
</tr>
<tr>
<td>GLUCOSE - LACTOSE</td>
<td>OPERON OFF</td>
</tr>
<tr>
<td>-GLUCOSE + LACTOSE</td>
<td>OPERON ON</td>
</tr>
</tbody>
</table>

**Sea Urchin**

A Module A functions:

- Vegetal plate expression in early development:
- Synergism with modules B and G enhancing endoderm expression in later development:
- Repression in ectoderm (modules E and F) and skeletogenic mesenchyme (module DC):

Modules E, F and DC with L:G treatment:
Classifying Promoters
Classifiers for Biological Problems

• Accuracy versus interpretability
• Off-the-shelf versus domain-specific tools
• Perfect versus imperfect labeling
Positives-only Data: Big picture

Wrap the “tightest” surface around the known examples!
An Example: Identification of Targets of a Transcription Factor

Supervised learning problem:
Find all functional targets of a factor in the genome from the knowledge a few examples.
Some Known Targets of lacI and of crp

lacI binding sites:

\[
\begin{align*}
AATTTGTTGAGCGGATAAAC & \text{AATT} \\
AAATTTGTTGAGCGGAGT & \text{AAACCA} \\
GGCAGTGGACGCGCAA & \text{ACGCAATT}
\end{align*}
\]

some crp binding sites:

\[
\begin{align*}
& \text{TAAAAATTTACGTCCCTTTGTAC} \\
& \text{GAAGCGGACTTGTTCAATGCGA} \\
& \text{GGTG TGAATTTGATCAGCATTTC} \\
& \text{GATG CGAGGCGGATCGAAAAA} \\
& \text{AAA TTTCAATATTCATCACACCTT} \\
& \text{TAAAAATTTACGTCCCTTTGTAC} \\
& \text{AAAAATTTACGTCCCTTTGTAC}
\end{align*}
\]

Consensus:

\[
\text{AAATTTGTTGATCTAGATCACATT}
\]
Describing Fuzzy Motifs

Weight Matrix
[Berg, vonHippel, Studen, Stormo, …]

Given a set of known factor binding motifs, like,

\[
\begin{align*}
\text{TAATGTGACGTCCTTTGCATAC} \\
\text{GAAGGCGACCTGGGTATGCTG} \\
\text{CGATGCAGGCGGATCGAAAA} \\
\text{ATTTGAACCAGATCGCATT} \\
\text{AAATGTAAAGCTGTGCACGTTT}
\end{align*}
\]

construct a frequency matrix \( n_{ib} \)

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\text{Position} & 1 & 2 & 3 & \cdots \cdots & 22 \\
\hline
\text{A} & 3 & 4 & 5 & \cdots \cdots & 3 \\
\hline
\text{C} & 2 & 1 & 1 & \cdots \cdots & 2 \\
\hline
\text{G} & 2 & 2 & 2 & \cdots \cdots & 2 \\
\hline
\text{T} & 2 & 2 & 2 & \cdots \cdots & 2 \\
\hline
\end{array}
\]
Fuzzy Motifs

Weight Matrix Continued...

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

For any sequence \( S \), the score \( W \) is given by: \( W = \sum_{ib} \omega_{ib} S_{ib} \)

For example:
\[
W(TTAGCA.....) = \omega_{1T} + \omega_{2T} + \omega_{3A} + \omega_{4G} + \omega_{5C} + \omega_{6A} + ...... \\
\]

Sequences with higher \( W \) are better binders.
Precise relationship with binding energy in certain limits.
Problem of Threshold Selection
Independent Base Model

\[ E(TTAGCAA) = \varepsilon_1 T + \varepsilon_2 T + \varepsilon_3 A + \varepsilon_4 G + \varepsilon_5 C + \varepsilon_6 A + \varepsilon_7 A = \sum_{ib} \varepsilon_{ib} S_{ib} = \varepsilon \cdot S \]
Binding Probability

\[ f(E(S)) = \frac{n}{n + Ke^{\beta E(S)}} = \frac{1}{e^{\beta (E(S) - \mu)} + 1} \]

Remember \( \mu = k_B T \ln(n/K) \)

It is the Fermi (Logistic) function!
Threshold Set by Concentration of Transcription Factor
The Probability Model for Data:
Low Stringency SELEX
Maximum Likelihood Method for Estimating $\epsilon$ and $\mu$

\[ e^{\mathcal{L}(\epsilon, \mu | O)} = \prod_{S \in O} [\gamma f(E(S))] \prod_{S' \notin O} [1 - \gamma f(E(S'))] \]

Non-degenerate Limit
Low $\mu$
\[ f(E(S)) \rightarrow e^{\beta \mu} e^{-\beta E(S)} \]
Info. Theory Weight Matrix

Zero Temperature Limit
Low $T$
\[ f(E(S)) \rightarrow \Theta(\mu - E(S)) \]
Support Vector Machine (QPMEME)
Increasing Width of $D(E)$ increases number of ‘False Positives’/ Random Background

Minimize the Variance!
Quadratic Programming Method for Energy Matrix 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.

Applied to $\sim50$ *E. coli* TFs in the DPInteract Database
The hyperplane farthest from the origin consistent with data

\[ \vec{\varepsilon} = \sum_a \alpha_a \vec{S}_a = \alpha_1 \vec{S}_1 + \alpha_2 \vec{S}_2 \]

Parallel Work in Machine Learning Community

Probability models and SVM
Platt (1999)

One class SVM
Schoelkopf et al. (2001)
Manevitz and Yousef (2001)
Tax and Duin (2002)
A biophysical approach to transcription factor binding site discovery

Summary

Identification of transcription binding sites within the regulatory segments of genomic DNA is an important step towards understanding of regulatory circuits that control expression of genes. It is also a task where methods of bio-informatics can be very effective. A powerful general approach to bio-informatic identification of binding sites is based on a "weight matrix" which assigns a position dependent value to each of the possible bases of a sequence segment and combines them into a "score" used for classification. Currently, the widely used method for defining the weight matrix is based on the information theoretic considerations and assigns each sequence an "information score" (for review see Stormo G.D. (2000), "DNA binding sites: representation and discovery", Bioinformatics 16, 16-23). Here we describe a novel method, which is based on the bio-physical considerations and defines the weight matrix by estimating the sequence dependent (free) energy of binding, which is then used for site classification. The new method also provides for each transcription factor an estimate of the chemical potential which acts as a "binding threshold". Although derived from physical considerations, our method is algorithmically related to the "support vector machine" approach to pattern recognition (Cristianini, N. and Shawe-Taylor, J., (2001), Intro to support vector machines, Cambridge Univ. Press). The new method for binding site discovery provides a significant improvement over the information score based weight matrix approach, particularly in the ubiquitous case of low specificity factors where it allows to reduce the expected number of false positives without sacrifice in the number of false negatives. The new method is used to identify likely genomic binding sites for the E.coli transcription factors collected in DPInteract database.

Reference: Marko Djordjevic, Anirvan M. Sengupta and Boris I. Shraiman, "A biophysical approach to transcription factor binding site discovery" (Genome Research 2003, submitted)

Summary of binding sites found in E. coli genome search:

The table below summarizes search results for E. coli transcription factors compiled in the DPInteract database, and compares with the information score search results (Robison et al, (1998) J. Mol. Biol. 284, 241-254.) Transcription factor names link to search parameters (energy matrices and binding thresholds) and complete lists of candidate sites.

<table>
<thead>
<tr>
<th>Name</th>
<th>Length</th>
<th>Number of examples</th>
<th>Information score &quot;hits&quot;</th>
<th>QPMEME &quot;hits&quot;</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>AraC</td>
<td>48</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7*10^-5</td>
</tr>
<tr>
<td>AraA</td>
<td>15</td>
<td>14</td>
<td>391</td>
<td>52</td>
<td>6.4</td>
</tr>
<tr>
<td>AraR</td>
<td>18</td>
<td>17</td>
<td>320</td>
<td>79</td>
<td>8.9</td>
</tr>
<tr>
<td>CspA</td>
<td>25</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1*10^-5</td>
</tr>
<tr>
<td>CspA</td>
<td>22</td>
<td>49</td>
<td>3093</td>
<td>796</td>
<td>27.2</td>
</tr>
<tr>
<td>CspA</td>
<td>20</td>
<td>4</td>
<td>15</td>
<td>4</td>
<td>2*10^-3</td>
</tr>
<tr>
<td>CynR</td>
<td>21</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3*10^-4</td>
</tr>
</tbody>
</table>

- Solves the problem of threshold selection
- Better sensitivity/specificity tradeoff than conventional methods
- False negative rate 25% (for CAP)
- Positive predictive value 60-70% (for CAP)
Comparison with Conventional Weight Matrix Results
Significant over-abundance of *E. coli* sites under the threshold
Binding at physiological concentration:
Separation of bound sequences from the rest

Result of weight matrix misestimating
the orientation of the separating plane

Effect of TF concentration:
Lower->Stringent, Higher->Relaxed

QPMEME estimates the orientation
and the location from marginal examples
EMSA for Predicted Sites

Positive predictive value = \( \frac{TP}{TP + FP} \) \(~ 60-70\%\)
Improvements?

• Corrections to independent base model: adding nearest neighbor terms
  (with O’Flanagan, Paillard, Lavery)
• “Unbiased” datasets: high-throughput SELEX experiments on CAP
  (with Nagaraj, O’Flanagan, Shraiman)
• Incorporation other type of information
  (gene expression, inter-species comparison,..)
DNA Deformation

Protein-DNA complex free energy
= Direct Protein DNA terms
+ DNA deformation terms
= $\sum_{ib} \varepsilon_{ib} S_{ib} + \sum_{ib} J_{i+1;i;b} S_{ia} S_{i+1;b}$
+......
TBP Binding TATA Box
Problems of Generalizing QPMEME to Include Deformation

• Four times more parameters.
• Number of sequences needed to train in hundreds
• Can possibly be done with SELEX SAGE
• However, for the time being, why not use atomistic calculation to get the best binders use that to test ideas? (O’Flanagan, Paillard, Lavery, Sengupta, Bioinformatics, to appear)
Performance of Algorithms on Computationally Generated Data
Performance vs # of Examples

• One could estimate optimal number of sites necessary (e.g. 60-70 for TBP)

• Could use structural insights to make it more sparse (informative priors).
Many Sequences: SELEX SAGE

A

Random sequence library

\[ 5' - 
\text{TCATTCCTGATGAGCTAT} \cdot \text{N(25)} \cdot 
\text{TACAGCTTAAACCGACTCGTATTTAATT-3'} \]

Second-strand synthesis by PCR

\[ 5' - 
\text{TCATTCCTGATGAGCTAT} \cdot \text{N(25)} \cdot 
\text{TACAGCTTAAACCGACTCGTATTTAATT-3'} \]

\[ 3' - 
\text{AGCTAGAGAGCGATACAGATCTAGAT} \cdot \text{N(25)} \cdot 
\text{ATCTAGAGGGAGGGCTAGGCAATTAA-5'} \]

Selection of binding sequences (gel shift)

Amplification

Digestion (BglII)

\[ 5' - \text{GATCTA} \cdot \text{N(25)} \cdot \text{TA} \]

\[ \text{AT} \cdot \text{N(25)} \cdot \text{TACAGT-3'} \]

Concatemerization and cloning

\[ 5' - \text{GATCTA} \cdot \text{N(25)} \cdot \text{TACAGT} \cdot \text{N(25)} \cdot \text{TACAGT} \cdot \text{N(25)} \cdot \text{TACAGT} \cdot \text{N(25)} \cdot \text{TACAGT-3'} \]

\[ \text{AT} \cdot \text{N(25)} \cdot \text{TACAGT} \cdot \text{N(25)} \cdot \text{TACAGT} \cdot \text{N(25)} \cdot \text{TACAGT} \cdot \text{N(25)} \cdot \text{TACAGT-3'} \]

Site 1

Site 2

Site 3

HTS sequencing

Roulet et al.
Nat. Biotech.
2002
Improvement of Correlation with Affinity

Comparison between E. Coli and Salmonella:
Fraction of conserved predicted sites improves: 55-60% --> 75-80%
(Nagaraj, O’Flanagan, Shraiman, Sengupta, ms. in preparation)
Phylogenetic Footprinting

From http://www.genetics.wustl.edu/saccharomycesgenomes/yeast_phylogeny.html
Functional weak sites

**HO(10) Strong site, highly conserved**

Scer  --------GTGTTGCGCGTTAAAACCTACATC-AAAAAAGG-CGGATCA
Spar  gtcaTAACGGTTTGGCCGCTAAAACCTACATC-AAAAAAGGCGGATCA
Smik  CAAt------TTTTACCGCGTTAAAACATACATCgAAAAAGGCGGATCA
Skud  ------TACGTTTACCGCGTTAAAACATACATC-AAAAAAGGCGGATCA
Sbay  AAgtTACATTTACCGCGTTAAAACATACATC-AAAAAAGGCGGATCA

**Output of Dialign, Morgenstern et al.**

**HO(7) Medium strength site, highly conserved**

Scer  CCAAAAGGGGATCAAAATATGGATGCTTTTTTCACCTACGATGATC
Spar  CCAAAAGGGGATCAAAATATGGATGCTTTTTTCACCTACGATGATC
Smik  CCAAAAGGGGATCAAAATATGGATGCTTTTTTCACCTACGATGATC
Skud  CTGAAGGGGATATCAAAATATGGATGCTTTTTTCACCTACGATGATC
Sbay  CTGAAGGGG--ATGATGATACATATTACACGATGATC

**HO(2) Medium strength site, not well conserved**

Scer  AATTCA-TGCAT-GTCCACATTAACATCTTGG-CAGAGCAAACTTACATC
Spar  AATTCA-TGTTAATGTTTACTTACATTAACATCTTGCAGAGAAACGGCTCGT
Smik  AACcttaTGCGAacGTTTACATTAACATCTCACTACAGAAAATAAAT
Skud  AAagaA-TTTATTTGTTTACATCAACTCTCTAGAGAAACTTCATC
Sbay  AACTGA-TGAAATGTTTACTTACATCAACTCTCTGAGAACATCATC

22100102221111222222222222232111112222222221111122
Scores of Evolutionarily Conserved Sites
Constrained Optimization

\[ E^{(A)}_a = \varepsilon \cdot S^{(A)}_a \leq -1, \forall a, A \]

\[ c_{AB}(\varepsilon) = \text{cov}(E^{(A)}, E^{(B)}) \]  

\[ \max \sum_{AB} c(\varepsilon)^{-1}_{AB} \]

Kinder approach:
Optimization of a quartic function.
Soluble, by iterated QP.

\[ E^{(A)}_a = \varepsilon \cdot S^{(A)}_a \leq -1, \forall a, A \]

\[ c_{AB}(\varepsilon) = \text{cov}(E^{(A)}, E^{(B)}) \]

\[ \min \sum_{AB} \gamma_A c_{AB}(\varepsilon)\gamma_A + \sum_A \gamma_A \]
Integration of Expression Data and Sequence Analysis: Too many Thresholds to choose?
Combining Multiple Scores

Suppose we have multiple scores \((x_{1g}, \ldots, x_{ng})\) for a gene \(g\) that are uncorrelated for the “generic” gene, but correlated for the regulated ones.

\[
p_g = \prod_i \Pr_i(x_i > x_{ig})
\]

Caveat: Need some feature selection method not to throw in irrelevant (or very weakly predictive) scores.
Yeast life cycle

Haploid invasive growth

Glucose starvation

Mating

Sporulation

Nitrogen starvation

Pseudohyphal development
Combinatorial Control of Yeast Mating Type Identity

- Combinatorial control by three regulated factors (and a constitutive one) regulate cell type identity.
- Detection of direct targets by combining sequence analysis and microarray data (Nagaraj, O’Flanagan, Bruning, Mathias, Vershon, Sengupta, BMC Genomics 2004)
Mutational data

Measurement of fold repression caused by single base mutations of the “consensus” sequence in a heterologous promoter (Jin, Zhong and Vershon, MCB, 1999) allows us to score other sequences.
# Microarray data for polyploids

<table>
<thead>
<tr>
<th>orf</th>
<th>a</th>
<th>aa</th>
<th>aaa</th>
<th>aaaa</th>
<th>x</th>
<th>xx</th>
<th>xxx</th>
<th>xxxx</th>
<th>ax</th>
<th>aax</th>
<th>aaxx</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAL069W</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>YAL067C</td>
<td>16</td>
<td>33</td>
<td>10</td>
<td>31</td>
<td>13</td>
<td>1</td>
<td>15</td>
<td>8</td>
<td>15</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>YAL066W</td>
<td>32</td>
<td>34</td>
<td>30</td>
<td>5</td>
<td>26</td>
<td>23</td>
<td>35</td>
<td>9</td>
<td>30</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

(YO: haploid specific gene)

(YKL178C | 49| 43 | 67  | 80   | 436 | 327| 310 | 520  | 30  | 6   | 49   |
(STE3: α-specific gene)

(YKL209C | 342| 332| 289 | 261  | 44  | 57 | 49  | 80   | 59  | 40  | 55   |
(STE6: α-specific gene)

(Galitski, Saldhana, Styles, Lander and Fink, Science, 1999)
Assigning Scores for Expression

A high score is given to genes which are repressed in the presence of α1 but not in α1−α2.

Expression p-value = The fraction of genes with a better score in α1−α2 type cells

Binding p-value = Probability of finding a good binding site in the promoter

Combined p-value = p-value for high expression score and p-value for good binding site in the promoter.
## Ordering of Candidate Targets

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene</th>
<th>Sub-classa</th>
<th>Expression</th>
<th>Binding</th>
<th>Combined</th>
<th>a1-α2</th>
<th>ChIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>YDL227C</td>
<td>HO</td>
<td>1</td>
<td>0.0006</td>
<td>0.0017</td>
<td>1.1e-6</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>YLR265C</td>
<td>NEJ1</td>
<td>1</td>
<td>0.0003</td>
<td>0.0053</td>
<td>1.7e-6</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>YBL016W</td>
<td>FUS3</td>
<td>2</td>
<td>0.0001</td>
<td>0.0991</td>
<td>1.6e-5</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>YOR212W</td>
<td>STE4</td>
<td>2</td>
<td>0.0020</td>
<td>0.0082</td>
<td>1.7e-5</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>YJR086W</td>
<td>STE18</td>
<td>2</td>
<td>0.0008</td>
<td>0.0218</td>
<td>1.7e-5</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>YHR005C</td>
<td>GPA1</td>
<td>2</td>
<td>0.0005</td>
<td>0.0437</td>
<td>2.1e-5</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>YDR103W</td>
<td>STE5</td>
<td>2</td>
<td>0.0017</td>
<td>0.0298</td>
<td>5.2e-5</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>YBR073W</td>
<td>RDH54</td>
<td>1</td>
<td>0.0053</td>
<td>0.0116</td>
<td>6.2e-5</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>YGR044C</td>
<td>RME1</td>
<td>3</td>
<td>0.0009</td>
<td>0.0720</td>
<td>6.8e-5</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>YGL248W</td>
<td>PDE1</td>
<td>4</td>
<td>0.0182</td>
<td>0.0040</td>
<td>7.3e-5</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>YPL038W</td>
<td>MET31</td>
<td>4</td>
<td>0.0292</td>
<td>0.0027</td>
<td>8.0e-5</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>YDR088C</td>
<td>SLU7</td>
<td>4</td>
<td>0.0303</td>
<td>0.0038</td>
<td>1.2e-4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>YGL052W</td>
<td></td>
<td></td>
<td>0.0117</td>
<td>0.0109</td>
<td>1.3e-4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>YYL157C</td>
<td>FAR1</td>
<td>2</td>
<td>0.0013</td>
<td>0.1141</td>
<td>1.4e-4</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>YPR122W</td>
<td>AXL1</td>
<td>2</td>
<td>0.0091</td>
<td>0.0163</td>
<td>1.5e-4</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>YIL099W</td>
<td>SGA1</td>
<td>3</td>
<td>0.0063</td>
<td>0.0267</td>
<td>1.7e-4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>YLR233C</td>
<td>EST1</td>
<td>1</td>
<td>0.0226</td>
<td>0.0090</td>
<td>2.0e-4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>YKL182W</td>
<td>FAS1</td>
<td></td>
<td>0.0578</td>
<td>0.0035</td>
<td>2.1e-4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>YMR053C</td>
<td>STB2</td>
<td>3</td>
<td>0.0028</td>
<td>0.0884</td>
<td>2.5e-4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>YNL319W</td>
<td></td>
<td></td>
<td>0.0123</td>
<td>0.0222</td>
<td>2.7e-4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>YFR012W</td>
<td></td>
<td></td>
<td>0.0088</td>
<td>0.0125</td>
<td>2.8e-4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>YNL188W</td>
<td>KAR1</td>
<td>2</td>
<td>0.0019</td>
<td>0.1890</td>
<td>3.6e-4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>YGL193C</td>
<td></td>
<td></td>
<td>0.0014</td>
<td>0.2654</td>
<td>3.8e-4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>YMR157C</td>
<td></td>
<td></td>
<td>0.1557</td>
<td>0.0026</td>
<td>4.0e-4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>YIL117C</td>
<td>PRM5</td>
<td></td>
<td>0.0011</td>
<td>0.3689</td>
<td>4.1e-4</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Significant combinations

Choice of cutoff
Ordering of Candidate Targets

Table 1. List Potential α1-α2 Binding Sites in Haploid-specific Genes

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene</th>
<th>Sub-class</th>
<th>Expression P-val</th>
<th>Binding P-val</th>
<th>Combined P-val</th>
<th>α1-α2 ChIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>YDL227C</td>
<td>HO</td>
<td>1</td>
<td>0.0006</td>
<td>0.0017</td>
<td>1.1e-6</td>
<td>+</td>
</tr>
<tr>
<td>YLR265C</td>
<td>NEJ1</td>
<td>1</td>
<td>0.0003</td>
<td>0.0053</td>
<td>1.7e-6</td>
<td>+</td>
</tr>
<tr>
<td>YBL016W</td>
<td>FUS3</td>
<td>2</td>
<td>0.0001</td>
<td>0.0991</td>
<td>1.6e-5</td>
<td>+</td>
</tr>
<tr>
<td>YOR212W</td>
<td>STE4</td>
<td>2</td>
<td>0.0020</td>
<td>0.0082</td>
<td>1.7e-5</td>
<td>+</td>
</tr>
<tr>
<td>YJR086W</td>
<td>STE18</td>
<td>2</td>
<td>0.0008</td>
<td>0.0218</td>
<td>1.7e-5</td>
<td>+</td>
</tr>
<tr>
<td>YHR005C</td>
<td>GPA1</td>
<td>2</td>
<td>0.0005</td>
<td>0.0437</td>
<td>2.1e-5</td>
<td>+</td>
</tr>
<tr>
<td>YDR103W</td>
<td>STE5</td>
<td>2</td>
<td>0.0017</td>
<td>0.0298</td>
<td>5.2e-5</td>
<td>+</td>
</tr>
<tr>
<td>YBR073W</td>
<td>RDH54</td>
<td>1</td>
<td>0.0053</td>
<td>0.0116</td>
<td>6.2e-5</td>
<td>+</td>
</tr>
<tr>
<td>YGR044C</td>
<td>RME1</td>
<td>3</td>
<td>0.0009</td>
<td>0.0720</td>
<td>6.8e-5</td>
<td>+</td>
</tr>
<tr>
<td>YGL248W</td>
<td>PDE1</td>
<td>4</td>
<td>0.0182</td>
<td>0.0040</td>
<td>7.3e-5</td>
<td>+</td>
</tr>
<tr>
<td>YPL038W</td>
<td>MET31</td>
<td>4</td>
<td>0.0292</td>
<td>0.0027</td>
<td>8.0e-5</td>
<td>+</td>
</tr>
<tr>
<td>YDR088C</td>
<td>SLU7</td>
<td></td>
<td>0.0303</td>
<td>0.0038</td>
<td>1.2e-4</td>
<td>-</td>
</tr>
<tr>
<td>YGL052W</td>
<td></td>
<td></td>
<td>0.0117</td>
<td>0.0109</td>
<td>1.3e-4</td>
<td>-</td>
</tr>
<tr>
<td>YJL157C</td>
<td>FAR1</td>
<td>2</td>
<td>0.0013</td>
<td>0.1141</td>
<td>1.4e-4</td>
<td>+</td>
</tr>
<tr>
<td>YFR122W</td>
<td>AXL1</td>
<td>2</td>
<td>0.0091</td>
<td>0.0163</td>
<td>1.5e-4</td>
<td>+</td>
</tr>
<tr>
<td>YIL099W</td>
<td>SGA1</td>
<td>3</td>
<td>0.0063</td>
<td>0.0267</td>
<td>1.7e-4</td>
<td>-</td>
</tr>
<tr>
<td>YLR233C</td>
<td>EST1</td>
<td>1</td>
<td>0.0226</td>
<td>0.0090</td>
<td>2.0e-4</td>
<td>-</td>
</tr>
<tr>
<td>YKL182W</td>
<td>FAS1</td>
<td></td>
<td>0.0578</td>
<td>0.0035</td>
<td>2.1e-4</td>
<td>-</td>
</tr>
<tr>
<td>YMR053C</td>
<td>STB2</td>
<td>3</td>
<td>0.0028</td>
<td>0.0884</td>
<td>2.5e-4</td>
<td>-</td>
</tr>
<tr>
<td>YNL319W</td>
<td></td>
<td></td>
<td>0.0123</td>
<td>0.0222</td>
<td>2.7e-4</td>
<td>-</td>
</tr>
<tr>
<td>YFR012W</td>
<td></td>
<td></td>
<td>0.0088</td>
<td>0.0125</td>
<td>2.8e-4</td>
<td>-</td>
</tr>
<tr>
<td>YNL188W</td>
<td>KAR1</td>
<td>2</td>
<td>0.0019</td>
<td>0.1890</td>
<td>3.6e-4</td>
<td>-</td>
</tr>
<tr>
<td>YGL193C</td>
<td></td>
<td></td>
<td>0.0014</td>
<td>0.2654</td>
<td>3.8e-4</td>
<td>-</td>
</tr>
<tr>
<td>YMR157C</td>
<td></td>
<td></td>
<td>0.1557</td>
<td>0.0026</td>
<td>4.0e-4</td>
<td>-</td>
</tr>
<tr>
<td>YIL117C</td>
<td>PRM5</td>
<td></td>
<td>0.0011</td>
<td>0.3689</td>
<td>4.1e-4</td>
<td>-</td>
</tr>
</tbody>
</table>
ChIP experiments

Predicted Direct Targets

Predicted Indirect Targets
Indirect Effects via Signaling Pathways?

Ste12

Direct Targets in mating pathway: G protein components, Ste5, Fus3, but not the downstream TF Ste12. Possibly affects baseline activity.
Orphan Sites

Could it be regulating *IME4* via an antisense transcript?
Ongoing follow-up work in Vershon lab.
Conclusion

• Ability to design classifying surfaces appropriate for the problem
• Principled way of determining cutoffs
• Experimental tests encouraging
• Need studies of generalization properties
Collaborators

Viji Nagaraj
Ruadhan O’Flanagan

Boris Shraiman
Marko Djordjevic (Columbia)

Adrian Bruning
Sean Hanlon (UNC)

Jonathan Mathias (U. Wisc.)
Andrew Vershon

Richard Lavery (IBPC, Paris)
Guillaume Paillard (IBPC, Paris)