Learning Methods for DNA Binding in Computational Biology

Mark Kon
Dustin Holloway
Yue Fan
Chaitanya Sai
Charles DeLisi

Boston University

IJCNN
Orlando
August 16, 2007
Outline

• Background on Transcription Factors and Regulation
• Motivation
• Use of SVM
• Inferences
• Regulatory pathway prediction
• Human genome work
Wealth of Sequence and Biochemical Data

The amount of sequence data available is rapidly increasing. Over 1,500 genome projects are ongoing.

There is a need for techniques that can rapidly determine which sequences in a genome are functional.
Biology: Transcription and Regulatory Control
Transcription: key to gene expression

DNA is transcribed into RNA and eventually proteins

Our concern: the first step - initiation of transcription
The transcription process

RNA Polymerase runs along DNA to produce RNA copy

Initiation of this process occurs when a TF binds to DNA at the start of transcription
The beginning: TF binds to DNA
**Basics of Transcription**

1. **Promoter** is a region of the DNA which tries to attract RNA polymerase so that transcription can be initiated. When cell transcribed it is expressed as a protein.

2. Differences between cells are determined by which proteins they produce, which are determined by which genes are expressed.

3. Promoter region of DNA contains regulatory sequences which attract proteins called transcription factors (TF). The presence of these proteins is required for transcription with RNA polymerase to begin.

4. Regulatory sequences consist of inexactly repeating patterns (motifs).

5. Motifs stand out as highly similar patterns across species - their function is to attract very specific transcription factors.
Regulatory sequences on the DNA attract the TF.

Recurring attracting sequences are *motifs or consensus sequences*.
Transcription Factor Binding

Binding between DNA and transcription factors (TF’s) is hard to predict chemically.

**Goal:** For a given TF in yeast or human, determine which genes’ promoters it binds to, and where.
Summary

• High throughput technologies, including ChIP-chip data, are rapidly increasing experimental information about transcription factor binding to DNA
• Identification of TF binding sites in the genome remains difficult and incomplete
• Machine learning approaches have potential to supplant difficult experimental methods
• SVM methods studied here have sensitivity of 70% and positive predictive value of 90% on the average.
Summary

• Applications to inferences on biochemical pathway information are given
PSSM = Position Specific Scoring Matrix

• G. D. Stormo, DNA Binding Sites: Representation and Discovery., Bioinformatics 16 16-23,2000
Assume a fixed species $S$ (e.g. baker's yeast, $S. \textit{cerevisiae}$) has genome $G$ (full set of genes).

Gene $g$ begins transcription (for protein production) when a transcription factor (TF) $t$ (a protein) chemically binds to it.

**Question:** given a fixed TF $t$, which genes $g \in G$ does it bind to?

Chemically hard to solve -
Machine learning approach

Consider training data set

\[ \mathcal{D}_0 = \{(g_i, y_i)\}_{i=1}^n, \]

where \( g_i \in \mathcal{G} \) and \( y_i \in \mathbb{B} = \{-1, 1\} \).

Assume

\[ y_i = \begin{cases} 1 & \text{if } g_i \text{ attaches the TF} \\ -1 & \text{otherwise} \end{cases}. \]

How to learn \( f_0 : \mathcal{G} \rightarrow \mathbb{B} \) from examples?

First define \( g \in \mathcal{G} \) formally by its promoter sequence
Machine learning approach

\[ p = p(g) = ACGGTCTGGT...CGT \]

= DNA sequence of promoter region of gene

(promoter region is where TF will attach; in yeast it has \(~1000\) bases).

Effectively

\[ f_0 : \mathcal{A}^{1000} \rightarrow \mathbb{B}, \]

with \( \mathcal{A} = \{A, G, C, T\} \)
Use of feature maps

**Remark:** If we map $A$ into numbers (e.g., $g = 0211323113...213$), $f_0$ difficult to guess from examples $D$.

**Feature maps**

A solution: map $G$ into a space where it's easier to classify.
Sample feature maps

Example: Feature map

\[ \Phi_1(g) = \mathbf{x}(\mathbf{p}(g)) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{104} \end{bmatrix} \]

with \( x_i \) = \# hits in \( \mathbf{p}(g) \) by PSSM for TF \( t_i \) (e.g., from a list of 104 TF's for yeast).

Example: Microarray expression data for gene \( g \)

\[ \Phi_2(g) = \mathbf{x} = \text{vector of expression levels of } g \text{ in 25 microarray experiments} \]
Sample feature maps

Example:

$$\Phi_3(g) = x(p(g)) = \text{vector of string counts in } p(g)$$

Consider ordered list

<table>
<thead>
<tr>
<th>string1</th>
<th>AAAAAAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>string2</td>
<td>AAAAAAC</td>
</tr>
<tr>
<td>string3</td>
<td>AAAAAAG</td>
</tr>
<tr>
<td>string4</td>
<td>AAAAAAT</td>
</tr>
<tr>
<td>string5</td>
<td>AAAAACA</td>
</tr>
</tbody>
</table>

of all strings of 6 base pairs.
Sample feature maps

Note $\mathbf{x} = \mathbf{x}(g)$ has components:

$$x_i = \# \text{ appearances of string } i \text{ in (upstream region of) } g.$$ 

$\mathbf{x}$ has $4^6 = 4,096$ components; $F = \mathbb{R}^{4,096}$.

We thus have a set of feature maps $\Phi_i : G \rightarrow F_i$ — feature spaces:

$$\Phi_i(g) = \mathbf{x}_i \in F_i;$$

For yeast, we use $k = 26$ such feature maps (some of them highly discriminatory).
Concatenation of feature spaces

Now form full feature space as direct sum:

\[ F = F_1 \oplus F_2 \oplus \ldots \oplus F_k, \]

i.e.,

\[ \mathbf{x} = (x_1, x_2, \ldots, x_k) \]
Concatenation of feature spaces

Note: The kernel (thus geometry) of the full feature space \( F \) is the sum of the individual kernels of \( F_i \):

\[
K(x, y) = \sum_i K_i(x, y),
\]

with \( y = (y_1, \ldots, y_k) \).

In particular, combining information contained in a collection of kernels \( K_i(\cdot, \cdot) \) is obtained from just taking their sum.
Basic SVM setup: the discriminating function $f$

With data $\mathcal{D}$, can we find function $f_1 : F \to B$ which generalizes above examples, so $f_1(\mathbf{x}) = y$ (i.e., correct prediction) for all feature vectors $\mathbf{x}$?

Easier: find $f : F \to \mathbb{R}$ where

$$f(\mathbf{x}) > 0 \text{ if } f_1(\mathbf{x}) = 1; \quad f(\mathbf{x}) < 0 \text{ if } f_1(\mathbf{x}) = -1.$$
Basic SVM setup: the kernel

Now define geometry of space $F$ by defining dot product:

Assume we are given any *kernel function* $K(x, y)$ which is *positive definite* and symmetric in $x, y$.

We then define geometry of $F$ by defining the nonlinear dot product

$$x \cdot y \equiv K(x, y).$$

Then apply SVM algorithm using geometry induced by $K$ to find optimized choice of $f$ (here $\mathbf{x}_i$ are examples)

$$f(x) = \sum_i \alpha_i K(\mathbf{x}_i, x) + b = \sum_i K(\mathbf{w}, x) + b.$$
Basic SVM setup: the kernel

Linear kernel case: $K(x, y) = x \cdot y$ (linear dot product). Then

$$f(x) = \sum_i \alpha_i x_i \cdot x + b = \left( \sum_i \alpha_i x_i \right) \cdot x + b \equiv w \cdot x + b.$$  

Final classification rule: $f(x) > 0 \Rightarrow y = 1$ (TF binds gene); $f(x) < 0 \Rightarrow y = -1$ (TF does not bind).

Learning from training data:

$$N_f = (f(x_1), \ldots, f(x_n)) = (y_1, \ldots, y_n).$$

Consider separating hyperplane $H : f(x) = 0$: 
Basic SVM setup: diagram

Geometric interpretation

Recall:

\[ f(x) = w \cdot x + b; \]

Fig 2: SVM geometry (2 dimensions)
Feature spaces

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOT:</td>
<td>Motif hits in S. cerevisiae</td>
</tr>
<tr>
<td>CON:</td>
<td>Motif hits conservation 18 organisms</td>
</tr>
<tr>
<td>PHY:</td>
<td>Phylogenetic profile</td>
</tr>
<tr>
<td>EXP:</td>
<td>Expression correlation</td>
</tr>
<tr>
<td>GO:</td>
<td>GO term profile</td>
</tr>
<tr>
<td>KMER:</td>
<td>K-strings – 4, 5, 6-mers</td>
</tr>
<tr>
<td>S1:</td>
<td>Split 6-string 1 gap kkk_kkk</td>
</tr>
<tr>
<td></td>
<td>Description</td>
</tr>
<tr>
<td>---</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>S2</td>
<td>Split 6-string 2 gaps</td>
</tr>
<tr>
<td>S3</td>
<td>Split 6-string 3 gaps</td>
</tr>
<tr>
<td>S4</td>
<td>Split 6-string 4 gaps</td>
</tr>
<tr>
<td>S5</td>
<td>Split 6-string 5 gaps</td>
</tr>
<tr>
<td>S6</td>
<td>Split 6-string 6 gaps</td>
</tr>
<tr>
<td>S7</td>
<td>Split 6-string 7 gaps</td>
</tr>
<tr>
<td>S8</td>
<td>Split 6-string 8 gaps</td>
</tr>
</tbody>
</table>
## Feature spaces

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>M01</td>
<td>6-string with 1 mismatch (count 0.1)</td>
</tr>
<tr>
<td>M05</td>
<td>6-string with 1 mismatch (count 0.5)</td>
</tr>
<tr>
<td>ENT</td>
<td>Condition specific TF-target correlation</td>
</tr>
<tr>
<td>BIT</td>
<td>Nucleotide sparse binary encoding</td>
</tr>
<tr>
<td>CRV</td>
<td>Promoter Curvature prediction</td>
</tr>
<tr>
<td>HC</td>
<td>Homolog Conservation</td>
</tr>
<tr>
<td>HYD</td>
<td>Hydroxyl Cleavage</td>
</tr>
</tbody>
</table>
## Feature spaces

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPo</td>
<td>Kmer median positions from start</td>
</tr>
<tr>
<td>KPr</td>
<td>Kmer Probabilities (-log pval)</td>
</tr>
<tr>
<td>MT</td>
<td>Promoter Melting Temperature-20bp window</td>
</tr>
<tr>
<td>DG</td>
<td>Promoter Melting Delta G profile-20bp win</td>
</tr>
<tr>
<td>BND</td>
<td>Promoter bend prediction</td>
</tr>
</tbody>
</table>
Feature spaces

Many of these methods are not so reliable on their own, but can combine using statistical inference to yield a more powerful prediction scheme.
Promoter Sequences

Motif Detection using Position Specific Scoring Matrices for 163 TFs

Selection of Features: Rationale

Overrepresentation (Degeneracy) Analysis
Count motifs for each TF-target pair

Conservation Analysis
Using 18 Genomes

Expression Correlation Analysis

Experiments

1 2 3 4 5 ... 1011
Degeneracy: Repetitive TF Binding Sites

P(\text{True} | 2 \text{ hits}) = 2 \cdot P(\text{True} | 1 \text{ hit})

Having more than one detected binding site for a TF in the upstream region of a gene increases the likelihood that the TF truly binds the gene.

Some transcription factors have a preference for repetitive motifs.

This is Supplementary Table 5 From C. Harbison, E. Fraenkel, R. Young and e. al., Transcriptional Regulatory Code of a Eukaryotic Genome, *Nature* 431 99-104, 2004.
Conservation of a TF binding site in several orthologous upstream regions increases the likelihood that a potential site is a True site.
Expression Correlation Analysis

Two methods can be used to explore expression relationships:

1. Transcription factors that are highly correlated with potential targets are more likely to regulate those targets.

2. Pairs of genes with highly correlated expression are more likely to be regulated by the same TF.
SVM Algorithm

- 26 feature spaces lead to 26 kernels
- SVM forms hyperplane

\[ f(x) = w \cdot x + b = 0 \]

- Kernel

\[ K_{ij} = x_i \cdot x_j \]

(generalized inner product)
## Kernel Choices

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear</td>
<td>none</td>
<td>$K(x,y) = x \cdot y$</td>
</tr>
<tr>
<td>polynomial</td>
<td>poly degree $d$</td>
<td>$K(x,y) = (x \cdot y + 1)^d$</td>
</tr>
<tr>
<td>Gaussian radial basis function (RBF)</td>
<td>$\sigma$</td>
<td>$K(x,y) = \exp\left(\frac{-</td>
</tr>
<tr>
<td>Gaussian</td>
<td>$\sigma$</td>
<td>$K(x,y) = \frac{1}{2\pi\sigma^2} e^{- \frac{x^2+y^2}{2\sigma^2}}$</td>
</tr>
</tbody>
</table>
Probabilistic Interpretation (Platt)

• Rank the data by

\[ P(y_i = 1 | \mathbf{w} \cdot \mathbf{x}_i + b) \]

= posterior probability of positive classification given distance of \( \mathbf{x} \) from hyperplane.

Result: empirically based confidence levels given to SVM predictions.
Overall Algorithm

Synthesizing a single classifier from various data sources
Train a classifier on selected features.

Under-sample the negative set.

features → genes

Trainin g Set

Feature Reduction and Classifier Construction

- SVM-RFE to select top 1500 features.
- Train Platt’s SVM on selected features.

Final Training Classifier

Testing Set

Evaluate on test set

Single Accuracy Estimate

Repeat validation with new resampling of negatives. Average the Accuracy estimates over 100 repeats.

100X Average Accuracy
Weighting schemes for kernel sums

- Weighted sums of kernels are taken:

\[ K(x, y) = \sum_{i=1}^{26} \alpha_i K_i(x, y) \]

---

Scale with \( \alpha_i = \)

- Scaled \( F1 \) score
- Square of scaled \( F1 \) score
- Squared tangent of \( F1 \) score

(note latter have effect of emphasizing higher and better \( F1 \) values)
Kernels: accuracy scores
Summary of accuracy

• Best single kernel has sensitivity of .71 and PPV of .82
• Squared-tan weighting gives sensitivity .73 and PPV of .89
Summary of accuracy

TFs For Which Each Method is Significant (p ≤ 0.05)
F1 Scores: Random vs. Genomic Data
Sensitivity vs. Example Size
SVM vs. PSSM Scan

![Graph showing comparison between PSSM and SVM for 104 transcription factors. The graph compares Sensitivity, Specificity, PPV, and F1 for both 104 PSSMs and 104 SVMs.]
Implications for Pathways: GCN4 and Amino Acid Biosynthesis

[Flowchart description]
- L-aspartate 4-P-transferase (HOM3)
- L-aspartyl 4-P
- L-aspartate-semialdehyde
- Homoserine dehydrogenase (HOM6)
- Homoserine kinase (THR1)
- O-phospho-L-homoserine
- Threonine synthase (THR4)
- L-threonine
- O-acetylhomoserine (thiol)-lyase (MET17)
- Homocysteine
- Homoserine O-trans-acetylase (MET2)
- O-Acetyl-L-homoserine
- N5-methyltetrahydropterylglycine reductase (MET6)
- Tetrahydropteroyltri-L-glutamate
- L-methionine

Targets of GCN4 in an amino-acid biosynthetic pathway:
- Previously known to be regulated by GCN4
- New Predictions
- Reaction in metabolic pathway
- Transcriptional regulation
Implications for Pathways: RAP1 and Glycolytic/TCA Cycle
Degeneracy Significance

Degeneracy 0 means not detected by Motifscanner
Conservation Results

\[ P(T \mid k) \]

\[ P(k \mid T) \]

\[ P(T \mid k) \]
**Genomes:**

- *S. cerevisiae* (SC) with genes P1, P2, P3, P4, P5, P6, P7
- *E. coli* (EC) with genes P1, P2, P3, P4, P5, P6, P7
- *B. subtilis* (BS) with genes P1, P2, P3, P5, P6, P7

**Phylogenetic Profile:**

<table>
<thead>
<tr>
<th>EC</th>
<th>SC</th>
<th>BS</th>
<th>HI</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>P2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>P3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P6</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P7</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Profile Clusters:**

```
  P1  P5
   |   |
  P4  P2 P7
       |
      P3 P6
```

**Conclusion:**
- P2 and P7 are functionally linked.
- P3 and P6 are functionally linked.
Degeneracy

<table>
<thead>
<tr>
<th></th>
<th>Gene1</th>
<th>Gene2</th>
<th>Gene3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motif1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Motif2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Motif3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Conservation

<table>
<thead>
<tr>
<th></th>
<th>Gene1</th>
<th>Gene2</th>
<th>Gene3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motif1</td>
<td>4</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Motif2</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Motif3</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Phylogenetic Profile

<table>
<thead>
<tr>
<th></th>
<th>Genome1</th>
<th>Genome2</th>
<th>Genome3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Gene2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Gene3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Expression

<table>
<thead>
<tr>
<th></th>
<th>Exp1</th>
<th>Exp2</th>
<th>Exp3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene1</td>
<td>0.32</td>
<td>0.001</td>
<td>0.5</td>
</tr>
<tr>
<td>Gene2</td>
<td>-0.2</td>
<td>0.04</td>
<td>-0.001</td>
</tr>
<tr>
<td>Gene3</td>
<td>-0.6</td>
<td>0.4</td>
<td>-0.3</td>
</tr>
</tbody>
</table>

Dot products

Kernel Matrices

Combined Kernel Matrix

SVM
K-fold Random Resampling

- Random Subsampling performs K data splits of the dataset
  - Each split randomly selects a (fixed) no. examples without replacement
  - For each data split we retrain the classifier from scratch with the training examples and estimate $E_i$ with the test examples

- The true error estimate is obtained as the average of the separate estimates $E_i$
  - This method is significantly better than simple split sample techniques
    \[
    E = \frac{1}{K} \sum_{i=1}^{K} E_i
    \]
Some human target predictions

WT1 - a TF involved in Wilms' Tumor - makes up 8% of childhood cancers.

SVM predictions for WT1 targets suggest new Wilms tumor models.

Genes in significant loci include several oncogenes and tumor suppressors which are candidates for involvement in cancer progression.
Some human target predictions

Example: chromosomal region 11p15.5

- known to be involved in Wilms' Tumor.

Newly predicted targets for WT1 are statistically enriched (.0005) for genes falling in this region.

Three of these are possible tumor suppressors, i.e., RNH1, IGF2AS, and CD151.

Other regions known to play a role in Wilms' Tumor also contain new target predictions (16q, 1p36.3, 16p13.3, 17q25, and 4p16.3).

Anti-apoptotic (anti-programmed cell death) effects of WT1 are possibly related several new target genes, including BAX and PDE4B - may help mediate the effect.
Some human target predictions

Motif discovery used for new candidate WT1 binding motif:

Fig. 3 - Wt1 target motifs:
(A) From literature
(B) Rankings of candidate motif strings as determined by application of SVM to a string feature space str, and from another oligo-analysis.
(C) Top ranked motifs using the Weeder algorithm on SVM-based rankings.
Acknowledgments

Dustin Holloway is responsible for much of the above analysis.

Charles DeLisi initiated this project.
Machine Learning Predictions:

http://visant.bu.edu/