POIMs: Positional Oligomer Importance Matrices
(Understanding Support Vector Machine Based Signal Detectors)

Sören Sonnenburg
Fraunhofer FIRST.IDA, Berlin

joint work with
Alexander Zien, Petra Philips and Gunnar Rätsch
Friedrich Miescher Laboratory
The Motivating Application - Splice Site recognition

*Discriminate true signal positions against all other positions*

- \( \approx 150 \) nucleotides window around dimer

```
CT...GTCGTA...GAAGCTAGGAGCGC...ACGCGT...GA
```

- **True sites:** fixed window around a true splice site
- **Decoy sites:** all other consensus sites

- Sequences are compared via String-Kernels
  - For each position a Weighted Degree Kernel compares all \( k \)-mers up to maximal length \( K \)

**SVM \( \approx 3 \) times more accurate than IMCs**

(54.4% vs. 16.2% auPRC)
The Motivating Application - Splice Site recognition

*Discriminate true signal positions against all other positions*

≈ 150 nucleotides window around dimer

CT...GTCGTA...GAAGCTTAGGAGCGC...ACGCGT...GA

- **True sites:** fixed window around a true splice site
- **Decoy sites:** all other consensus sites

Sequences are compared via String-Kernels

- For each position a Weighted Degree Kernel compares all k-mers up to maximal length K

SVM ≈ 3 times more accurate than IMCs
(54.4% vs. 16.2% auPRC)
The Motivating Application - Splice Site recognition

*Discriminate true signal positions against all other positions*

≈ 150 nucleotides window around dimer

CT...GTCGTA...GAAGCTTAGGAGCGC...ACGCGT...GA

- **True sites**: fixed window around a true splice site
- **Decoy sites**: all other consensus sites

Sequences are compared via String-Kernels
- For each position a Weighted Degree Kernel compares all k-mers up to maximal length K

**SVM** ≈ 3 times more accurate than IMCs
(54.4% vs. 16.2% auPRC)
**Beauty in Generality**

- Transcription Start (Sonnenburg et al., Eponine Down et al.)
- Acceptor Splice Site (Philips et al.)
- Donor Splice Site (Philips et al.)
- Alternative Splicing (Rätsch et al., -)
- Transsplicing (Schweikert et al., -)
- Translation Initiation (Sonnenburg et al., Saeys et al.)

Drawback: SVM solution is hard to interpret!!
Beauty in Generality

- Transcription Start (Sonnenburg et al., Eponine Down et al.)
- Acceptor Splice Site (Philips et al.)
- Donor Splice Site (Philips et al.)
- Alternative Splicing (Rätsch et al., -)
- Transsplicing (Schweikert et al., -)
- Translation Initiation (Sonnenburg et al., Saeys et al.)

Drawback: SVM solution is hard to interpret!!
Goal

For PWMs we have sequence logos:

We would like to have similar means to understand Support Vector Machines.
Why Are SVM’s Hard to Interpret?

**SVM decision function is** \( \alpha \) **weighting of training points**

\[
s(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b
\]

\( \alpha_1 \cdot \) AAAACAAAATAGTAACATATCTTTTAGAGAAGACGTTCACCATTGAG

\( \alpha_2 \cdot \) AAGATTTAAAAAAAACAAATTTTAGCATTACAGATATAATAATCTAATT

\( \alpha_3 \cdot \) CACTCCCCAAATCAACCGATATTATTAGGTCAAATACACATCCGTCGTGCC

\( \vdots \)

\( \alpha_N \cdot \) TTAATTTCACCTCCACATACTTCCAGATCATCAATCTCCAAAAACACAC

**But we are interested in weights over features.**
Support Vector Machines

**SVM Scoring Function**

\[
w = \sum_{i=1}^{N} \alpha_i y_i \Phi(x_i)
\]

\[
s(x) := \sum_{k=1}^{K} \sum_{i=1}^{L-k+1} w(x[i]^k, i) + b
\]

<table>
<thead>
<tr>
<th>k-mer</th>
<th>pos. 1</th>
<th>pos. 2</th>
<th>pos. 3</th>
<th>pos. 4</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>+0.1</td>
<td>-0.3</td>
<td>-0.2</td>
<td>+0.2</td>
<td>...</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>-0.1</td>
<td>+2.4</td>
<td>-0.2</td>
<td>...</td>
</tr>
<tr>
<td>G</td>
<td>+0.1</td>
<td>-0.7</td>
<td>0.0</td>
<td>-0.5</td>
<td>...</td>
</tr>
<tr>
<td>T</td>
<td>-0.2</td>
<td>-0.2</td>
<td>0.1</td>
<td>+0.5</td>
<td>...</td>
</tr>
<tr>
<td>AA</td>
<td>+0.1</td>
<td>-0.3</td>
<td>+0.1</td>
<td>0.0</td>
<td>...</td>
</tr>
<tr>
<td>AC</td>
<td>+0.2</td>
<td>0.0</td>
<td>-0.2</td>
<td>+0.2</td>
<td>...</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>TT</td>
<td>0.0</td>
<td>-0.1</td>
<td>+1.7</td>
<td>-0.2</td>
<td>...</td>
</tr>
<tr>
<td>AAA</td>
<td>+0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>+0.1</td>
<td>...</td>
</tr>
<tr>
<td>AAC</td>
<td>0.0</td>
<td>-0.1</td>
<td>+1.2</td>
<td>-0.2</td>
<td>...</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>TTT</td>
<td>+0.2</td>
<td>-0.7</td>
<td>0.0</td>
<td>0.0</td>
<td>...</td>
</tr>
</tbody>
</table>
The Scoring System - Examples

\[ s(x) := \sum_{k=1}^{K} \sum_{i=1}^{L-k+1} w(x[i]^k, i) + b \]

Examples:
- WD-kernel (Rätsch, Sonnenburg, 2005)
- WD-kernel with shifts (Rätsch, Sonnenburg, 2005)
- Spectrum kernel (Leslie, Eskin, Noble, 2002)
- Oligo Kernel (Meinicke et al., 2004)

Not limited to SVM’s:
- Markov Chains (higher order/inhomogeneous/mixed order)
The SVM Weight Vector $w$

- Explicit representation of $w$ allows for (some) interpretation!
- String kernel SVMs capable of efficiently dealing with large $k$-mers $k > 10$

**But:** Weights for substrings not independent
Interdependence of $k$–mer Weights

What is the score for TAC?
- Take $w_{TAC}$?
- But substrings and overlapping strings contribute too!

Problem
The SVM-$w$ does **NOT** reflect the score for a motif
Positional Oligomer Importance Matrices (POIMs)

Idea:

- Given $k$-mer $z$ at position $j$ in the sequence, compute expected score $E[s(x) | x[j] = z]$ (for small $k$)

  
  \[
  \begin{array}{c}
  \text{AAAAAAAATAC}
  \\
  \text{AAAAAAAATACAA}
  \\
  \text{AAAAAAAATACAAAA}
  \\
  \\vdots
  \\
  \text{TACTTTTTTTTT}
  \end{array}
  \]

- Normalize with expected score over all sequences

\[
Q(z, j) := E[s(x) | x[j] = z] - E[s(x)]
\]

$\Rightarrow$ Needs efficient algorithm for computation
Efficient Computation

Effort of naive approach \textbf{exponential} $O(|\Sigma|^L + L|\Sigma|^k)$ (e.g. Splice Sites $10^{120}$)

$$Q(z,j) := \mathbb{E}[s(x) \mid x[j] = z] - \mathbb{E}[s(x)]$$

- Number of k-mers grows linearly with size of input
- Only features which are dependent on $(z,j)$ matter
- Computation can be split in contributions from 4 cases

Main contribution of the paper

\textbf{Efficient Recursive Algorithm:} Effort \textbf{linear} in length of input: $O(LN + L|\Sigma|^k)$
Ranking Features and Condensing Information

- Obtain highest scoring $z$ from $Q(z, i)$ (Enhancer or Silencer)

- Visualize POIM as heat map;
  - x-axis: position
  - y-axis: k-mer
  - color: importance

- For large $k$: Differential POIMs;
  - x-axis: position
  - y-axis: k-mer length
  - color: importance

<table>
<thead>
<tr>
<th>$z$</th>
<th>$i$</th>
<th>$Q(z, i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATTACA</td>
<td>10</td>
<td>+30</td>
</tr>
<tr>
<td>AGTAGTG</td>
<td>30</td>
<td>+20</td>
</tr>
<tr>
<td>AAAAAAAA</td>
<td>10</td>
<td>-10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Comparison with SVM-w

GATTACA and AGTAGTG at Fixed Positions 10 and 30
Comparison with SVM-\textit{w}

\textbf{GATTACA and AGTAGTG at fixed positions 10 and 30}

\begin{verbatim}
TGAGCGCGTGATTACAGTCCGTCTGGGCCCAGTAGTGCGTAGTCGCCGGGA
GGCATGTCGATTACAACAGAGCCCTCTCAGTAGTGGGGAGCCACGAAA
CCCGTCGAAAGATTACACACGGGGCGTGGAGTAGTGCCGATTACGGGCTC
GGTCGGCAGGATTACACGACGCGTACGTTACGAGTAGTGAAACACTGACTC
\end{verbatim}
Comparison with SVM-w

GATTACA and AGTAGTG at fixed positions 10 and 30

```
TGAGCGCGTGATTACAGTCCTGTTGGGCGGACGTAGTGCGTAGTCGGCAGGA
GGCATGGTGCATTACAAACGAGCCCTCTCAGTAGTGCGGGGAGCCACAGAA
CCCGTCGAAGATTACACACGGGGCGTGGGAGTAGTGCGCGATTACGGGCTC
GGTCGCGCACGGATTACACGACGCCTTTACGAGTAGTAACACTGACTCCTC
```
Comparison with SVM-w

GATTACA and AGTAGTG at fixed positions 10 and 30

TGAGCGCGTGATTACA GTCCGTCTGGGCCAGTAGTG CGTAGTCGCCGGGA GGCATGGTCGATTACA AACGAGCCCTCTCAGTAGTG GGGGAGCCACGAAAC CCCGTAGAAAGATTACACACGGGGCGTGAGGAGTAGTG CGCGATTACCGGCTCG TGGTCGCAGGAATTACACGACGCGTGTACGAGTAGTGAACACTGACTCTC
GATTACA at variable positions

TGAGCGCGTGATTACAGTCCGTCT
GGCTCGATCAAAACGAGCCCGAT
CCCGTCAACAGGATTACACACGG
GGTCGGCAGCTTACACGACAGCGT
Toy Example motif at Variable Positions

GATTACA at variable positions

TGAGCGCGTGGATTACAGTCCGTCT
GGCTCGATCACAAACGAGCCCGAT
CCCGTCGAACAGGATTACACACGG
GGTCGGCAGCTTACACGACAGCGT
Toy Example motif at Variable Positions

GATTACA at variable positions

Differential POIM Overview – GATTACA shift
C. elegans Acceptor Splice Site Recognition

- **Upstream**
  - AGGTAAAGT  
    - Position: -44/++  
    - Description: Donor
  - GGGGGGGG  
    - Position: -16/- -  
    - Description: Silencer?
  - TAATAA  
    - Position: -16/++  
    - Description: Branch

- **Central**
  - TTTTTTTC  
    - Position: -06/+  
  - TTTCAAG^A_G  
    - Position: -03/++  
    - Description: Acceptor

- **Downstream**
  - TTTTTTTT  
    - Position: +07/- -  
  - TTTTTT  
    - Position: +26/- -
Drosophila Transcription Starts

Differential POIM Overview – Drosophila TSS

<table>
<thead>
<tr>
<th>Motif Length (k)</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>-70</td>
</tr>
<tr>
<td>7</td>
<td>-60</td>
</tr>
<tr>
<td>6</td>
<td>-50</td>
</tr>
<tr>
<td>5</td>
<td>-40</td>
</tr>
<tr>
<td>4</td>
<td>-30</td>
</tr>
<tr>
<td>3</td>
<td>-20</td>
</tr>
<tr>
<td>2</td>
<td>-10</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>40</td>
</tr>
</tbody>
</table>

TATA-box

Inr TCA $\begin{array}{c} \frac{G}{T} \\ \frac{T}{C} \end{array}$

CpG

- TATAAAAA -29/++
- GTATAAAA -30/++
- ATATAAAA -28/++
- CAGTCAGT -01/++
- TCAGTTGT -01/++
- CGTCAGTT -03/++
- CGTCGCGG +18/++
- GCGCGC +23/++
- CGCGCGC +22/++
Positional Oligomer Importance Matrices

- Support Vector Machines often are state-of-the-art classifiers
- POIMs systematically compute the importances of positional motifs for the expected decision score
  - Useful to rank motifs and for visualization
  - Can even identify motif length
  - Applicable for a large class of popular scores (SVM+Spectrum/WD/Oligo kernel; Markov Chain)
- Promising results on toy and real world data

Efficient implementation [http://www.shogun-toolbox.org](http://www.shogun-toolbox.org)
Webinterface [http://galaxy.fml.tuebingen.mpg.de](http://galaxy.fml.tuebingen.mpg.de)
Positional Oligomer Importance Matrices

- Support Vector Machines often are state-of-the-art classifiers
- POIMs systematically compute the importances of positional motifs for the expected decision score
  - Useful to rank motifs and for visualization
  - Can even identify motif length
  - Applicable for a large class of popular scores
    (SVM+Spectrum/WD/Oligo kernel; Markov Chain)
- Promising results on toy and real-world data

Tables: http://www.fml.mpg.de/raetsch/projects/POIM
Efficient implementation: http://www.shogun-toolbox.org
Webinterface: http://galaxy.fml.tuebingen.mpg.de
Positional Oligomer Importance Matrices

- Support Vector Machines often are state-of-the-art classifiers.
- POIMs systematically compute the importances of positional motifs for the expected decision score.
  - Useful to rank motifs and for visualization.
  - Can even identify motif length.
  - Applicable for a large class of popular scores (SVM+Spectrum/WD/Oligo kernel; Markov Chain).
- Promising results on toy and real world data.

Tables: [http://www.fml.mpg.de/raetsch/projects/POIM](http://www.fml.mpg.de/raetsch/projects/POIM)
Efficient implementation: [http://www.shogun-toolbox.org](http://www.shogun-toolbox.org)
Webinterface: [http://galaxy.fml.tuebingen.mpg.de](http://galaxy.fml.tuebingen.mpg.de)

Poster P38
ISMB 2008 Travel Fellowship Award

Participant travel costs to present the project described was partially supported by the U.S. National Science Foundation. The content is solely the responsibility of the author(s) and does not necessarily represent the official views of the National Science Foundation.

Machine Learning Open Source Software

- A repository of several machine learning algorithms.
- All freely available under open source licenses.
- Bioinformatics projects available.

Visit http://mloss.org!