# 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*

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

Friedrich Miescher Laboratory of the Max Planck Society

# The Motivating Application - Splice Site recognition

*Discriminate true signal positions against all other positions*

≈ 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

```
AAACAAATAAGTAACTAATCTTTTAGGAAGAACGTTTCAACCATTTTGAG
AAGATTAAAAAAAAACAAATTTTTAGCATTACAGATATAATAATCTAATT
CACTCCCCAAATCAACGATATTTTAGTTCACTAACACATCCGTCTGTGCC
TTAATTTCACTTCCACATACTTCCAGATCATCAATCTCCAAAACCAACAC
```

- 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*

$\approx$ 150 nucleotides window around dimer

```
CT...GTC[GTA...GAAGCTAGGAGCGC...ACG]CGT...GA
```

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

```
AAACAAATAAGTAACTAATCTTTTAGGAAGAACGTTTCAACCATTTTGAG
AAGATTAAAAAAAAACAAATTTTTAGCATTACAGATATAATAATCTAATT
CACTCCCCAAATCAACGATATTTTAGTTCACTAACACATCCGTCTGTGCC
TTAATTTCACTTCCACATACTTCCAGATCATCAATCTCCAAAACCAACAC
```

- 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*

$\approx$ 150 nucleotides window around dimer

CT...GTC GTA...GAAGCTAGGAGCGC...ACG CGT...GA

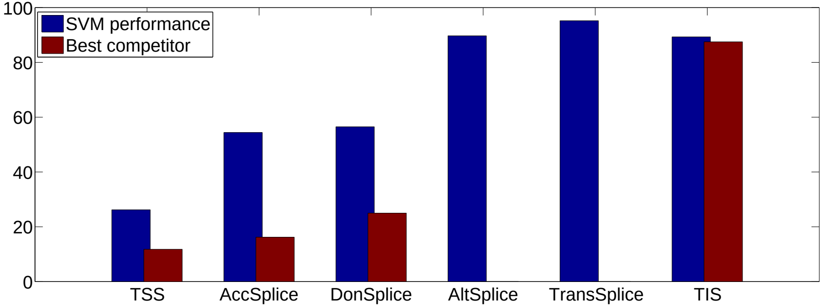
- True sites: fixed window around a true splice site
- Decoy sites: all other consensus sites

```
AAACAAATAAGTAACTAATCTTTTAGGAAGAACGTTTCAACCATTTTGAG
AAGATTAAAAAAAAACAAATTTTTAGCATTACAGATATAATAATCTAATT
CACTCCCCAAATCAACGATATTTTAGTTCACTAACACATCCGTCTGTGCC
TTAATTTCACTTCCACATACTTCCAGATCATCAATCTCCAAAACCAACAC
```

- 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)

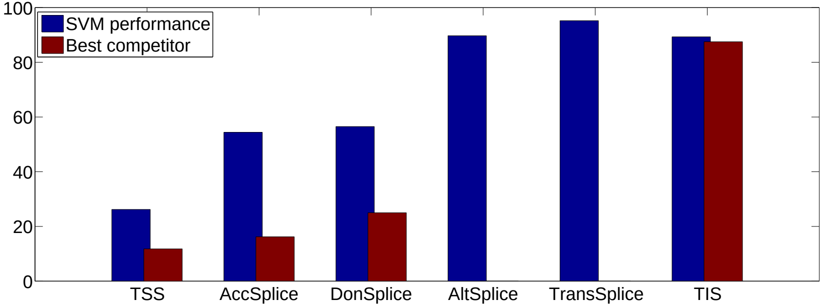
# Beauty in Generality




- 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



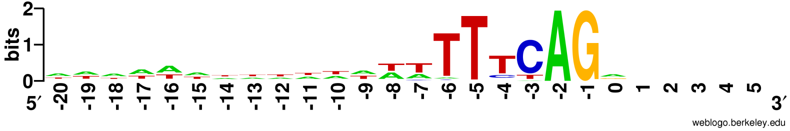

- Transcription Start (Sonnenburg et al., Eponine Down et al.)
- Acceptor Splice Site (Philips et al.)
- Donor Splice Site (Philips et al.)
- Alternative Splicing (Rätsch et al., -)
- Transsplicing (Schweikert et al., -)
- Translation Initiation (Sonnenburg et al., Saeys et al.)

**Drawback: SVM solution is hard to interpret!!**

# Understanding Support Vector Machines

## 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**

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

$\alpha_1 \cdot$ AAACAAATAAGTAACTAATCTTTTAGGAAGAACGTTTCAACCATTTTGAG

$\alpha_2 \cdot$ AAGATTAAAAAAAAACAAATTTTTAGCATTACAGATATAATAATCTAATT

$\alpha_3 \cdot$ CACTCCCCAAATCAACGATATTTTAGTTCACTAACACATCCGTCTGTGCC

$\vdots$ $\vdots$

$\alpha_N \cdot$ TTAATTTCACTTCCACATACTTCCAGATCATCAATCTCCAAAACCAACAC

**But we are interested in weights over features.**

# SVM Scoring Function

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \Phi(\mathbf{x}_i) \qquad s(\mathbf{x}) := \sum_{k=1}^{K} \sum_{i=1}^{L-k+1} w\left(\mathbf{x}[i]^k, i\right) + b$$

| k-mer | pos. 1 | pos. 2 | pos. 3 | pos. 4 | $\cdots$ |
|---|---|---|---|---|---|
| **A** | +0.1 | -0.3 | -0.2 | +0.2 | $\cdots$ |
| **C** | 0.0 | -0.1 | +2.4 | -0.2 | $\cdots$ |
| **G** | +0.1 | -0.7 | 0.0 | -0.5 | $\cdots$ |
| **T** | -0.2 | -0.2 | 0.1 | +0.5 | $\cdots$ |
| **AA** | +0.1 | -0.3 | +0.1 | 0.0 | $\cdots$ |
| **AC** | +0.2 | 0.0 | -0.2 | +0.2 | $\cdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ |
| **TT** | 0.0 | -0.1 | +1.7 | -0.2 | $\cdots$ |
| **AAA** | +0.1 | 0.0 | 0.0 | +0.1 | $\cdots$ |
| **AAC** | 0.0 | -0.1 | +1.2 | -0.2 | $\cdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ |
| **TTT** | +0.2 | -0.7 | 0.0 | 0.0 | $\cdots$ |

# The Scoring System - Examples

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

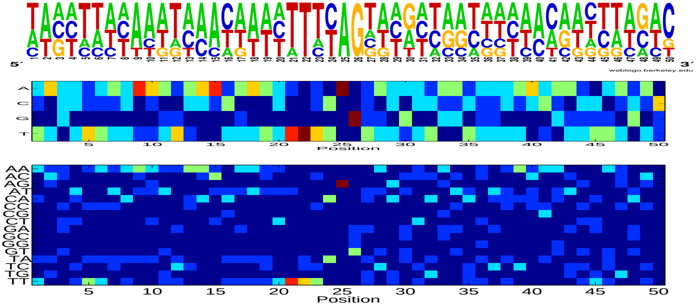
**Examples:**

- 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**

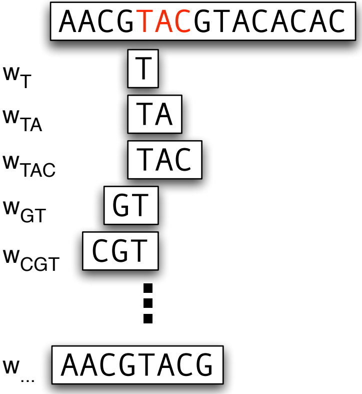



⋮

- 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 $\mathbf{z}$ at position $j$ in the sequence, compute expected score $\mathbb{E}\left[\, s(\mathbf{x}) \mid \mathbf{x}\,[j] = \mathbf{z}\,\right]$ (**for small k**)

```
AAAAAAAAAATACAAAAAAAAAA
AAAAAAAAAATACAAAAAAAAAC
AAAAAAAAAATACAAAAAAAAAG
           ⋮
TTTTTTTTTTTACTTTTTTTTTT
```

- Normalize with *expected score* over **all sequences**

**POIMs**

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

$\Rightarrow$ **Needs efficient algorithm for computation**

# Efficient Computation

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

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

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

**Main contribution of the paper**

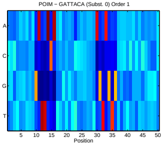
**Efficient Recursive Algorithm:**
**Effort linear in length of input:** $\mathcal{O}(LN + L|\Sigma|^k)$

# Ranking Features and Condensing Information

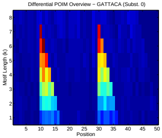
- Obtain highest scoring $\mathbf{z}$ from $Q(\mathbf{z}, i)$ (Enhancer or Silencer)

| $\mathbf{z}$ | i | $Q(\mathbf{z}, i)$ |
|---|---|---|
| GATTACA | 10 | +30 |
| AGTAGTG | 30 | +20 |
| AAAAAAA | 10 | -10 |
| ... | ... | ... |

- 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




Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

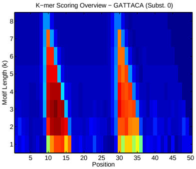
# `GATTACA` and `AGTAGTG` at Fixed Positions 10 and 30

```
TGAGCGCGTGATTACAGTCCGTCTGGGCCAGTAGTGCGTAGTCGCCGGGA
GGCATGGTCGATTACAAACGAGCCCTCTCAGTAGTGGGGAGCCACGAAA
CCCGTCGAAGATTACACACGGGGCGTGGGAGTAGTGGCGATTACGGGCTC
GGTCGGCAGGATTACACGACGCGTTTACGAGTAGTGAACACTGACTCCTC
```

# GATTACA and AGTAGTG at fixed positions 10 and 30

```
TGAGCGCGTGATTACAGTCCGTCTGGGCCAGTAGTGCGTAGTCGCCGGGA
GGCATGGTCGATTACAAACGAGCCCTCTCAGTAGTGGGGGAGCCACGAAA
CCCGTCGAAGATTACACACGGGGCGTGGGAGTAGTGGCGATTACGGGCTC
GGTCGGCAGGATTACACGACGCGTTTACGAGTAGTGAACACTGACTCCTC
```

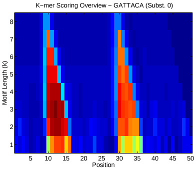
**w**



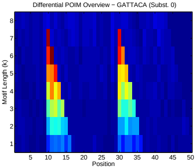
# GATTACA and AGTAGTG at fixed positions 10 and 30

TGAGCGCGTGATTACAGTCCGTCTGGGCCAGTAGTGCGTAGTCGCCGGGA
GGCATGGTCGATTACAAACGAGCCCTCTCAGTAGTGGGGGAGCCACGAAA
CCCGTCGAAGATTACACACGGGGCGTGGGAGTAGTGGCGATTACGGGCTC
GGTCGGCAGGATTACACGACGCGTTTACGAGTAGTGAACACTGACTCCTC

**w**

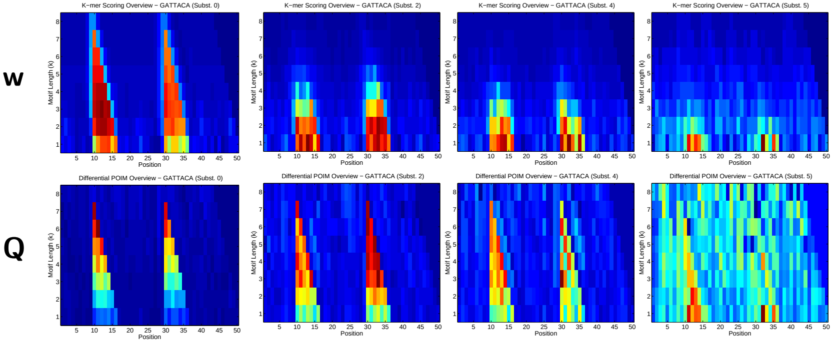



**Q**

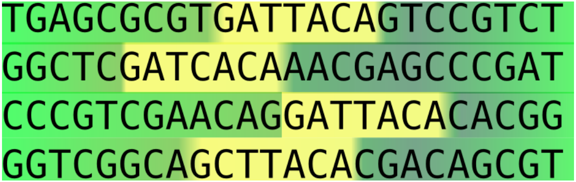


# GATTACA and AGTAGTG at fixed positions 10 and 30

```
TGAGCGCGTGATTACAGTCCGTCTGGGCCAGTAGTGCGTAGTCGCCGGGA
GGCATGGTCGATTACAAACGAGCCCTCTCAGTAGTGGGGAGCCACGAAA
CCCGTCGAAGATTACACACGGGGCGTGGGAGTAGTGGCGATTACGGGCTC
GGTCGGCAGGATTACACGACGCGTTTACGAGTAGTGAACACTGACTCCTC
```

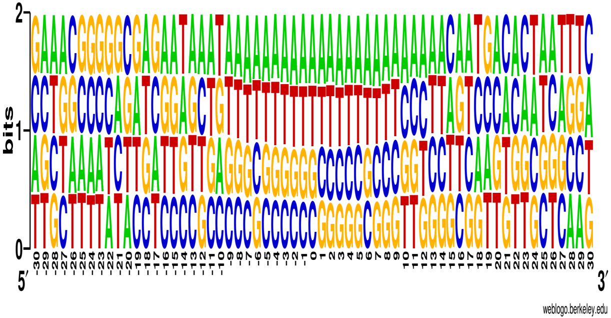


# GATTACA at variable positions

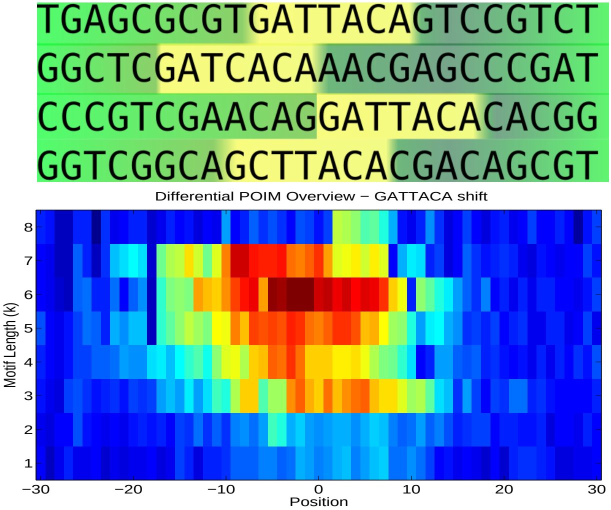


# GATTACA at variable positions

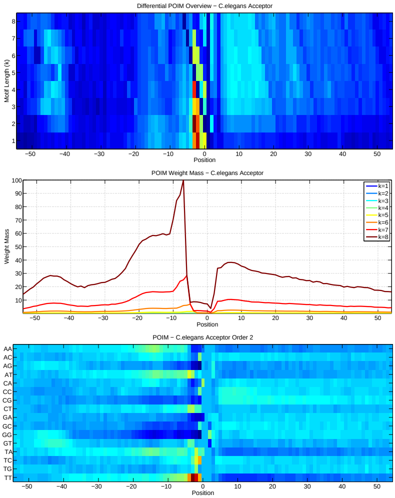
TGAGCGCGTGATTACAGTCCGTCT
GGCTCGATCACAAACGAGCCCGAT
CCCGTCGAACAGGATTACACACGG
GGTCGGCAGCTTACACGACAGCGT



# GATTACA at variable positions



# *C.elegans* Acceptor Splice Site Recognition




- Upstream

| | | |
|---|---|---|
| AG**GT**AAGT | -44/++ | Donor |
| GGGGGG | -16/- - | Silencer? |
| **TAAT**AA | -16/++ | Branch |

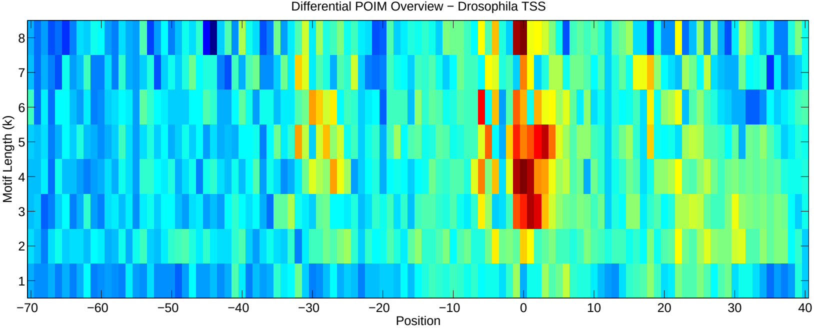
- Central

| | | |
|---|---|---|
| TTTTTTC | -06/+ | |
| TTTC**AG** $\frac{A}{G}$ | -03/++ | Acceptor |

- Downstream

| | |
|---|---|
| TTTTTTTT | +07/- - |
| TTTTT | +26/- - |

# Drosophila Transcription Starts




| | | | | | |
|---|---|---|---|---|---|
| TATAAAA | -29/++ | **CAGT**CAGT | -01/++ | CGTCGCG | +18/++ |
| GTATAAA | -30/++ | T**CAGT**TGT | -01/++ | GCGCGCG | +23/++ |
| ATATAAA | -28/++ | CGT**CAGTT** | -03/++ | CGCGCGC | +22/++ |

**TATA-box** **Inr** $\mathbf{TCA}\frac{G}{T}\mathbf{T}\frac{T}{C}$ **CpG**

# Conclusions

## 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`

Poster P38

# Conclusions

## 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`

Poster P38

# Conclusions

## 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`

Poster P38

# Acknowledgements


## 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 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`!