ARTS: Accurate Recognition of Transcription Starts in *human*

(A SHOGUN Machine Learning Toolbox Application)

Sören Sonnenburg[†], Alexander Zien[*,‡], Gunnar Rätsch[‡]

[†] Fraunhofer FIRST.IDA, Berlin
[*] Max Planck Institute for Biological Cybernetics, Tübingen
[‡] Friedrich Miescher Laboratory of the Max Planck Society, Tübingen

FIRST

Fraunhofer Institut
Rechnerarchitektur
und Softwaretechnik

MAX-PLANCK-GESELLSCHAFT

1. SHOGUN
    - Features - Overview
    - Kernels
    - Going Large Scale
    - Multiple Domains
    - Interpretability
2. ARTS: A Method for TSS Finding
    - Finding Transcription Start Sites
    - Properties
    - Kernels
    - Results
    - Interpretability
3. Summary
    - Discussion

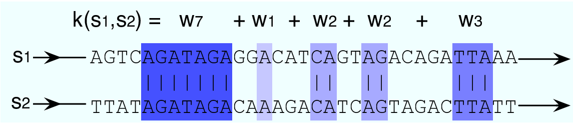
Machine Learning Toolbox SHOGUN

Main Features:

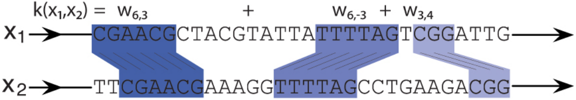
- Toolbox's focus is on kernel methods esp. Support Vector Machines (SVMs) for computational biology
- Includes a variety of common kernels (Linear, Polynomial, Gaussian) and recent String Kernels
- Kernels can be combined; weighting can be learned using Multiple Kernel Learning.
- Tuned for large scale data sets (parallelized SVM training on 10,000,000 DNA sequences in 27hrs, parallelized SVM testing on 7 billion examples)
- For string kernels: ⇒ **interpretability**

String Kernels

- Spectrum Kernel
  - Count k-mers in each sequence, Spectrum Kernel is sum of product of counts
- Weighted Degree Kernel


k(s1,s2) = w7 + w1 + w2 + w2 + w3
s1
AGTCAGATAGAGGACATCAGTAGACAGATTAAA
s2
TTATAGATAGACAAAGACATCAGTAGACTTATT

- Weighted Degree Kernel with Shifts


k(x1,x2) = w6,3 + w6,-3 + w3,4
x1
CGAACGCTACGTATTATTTTAGTCGGATTG
x2
TTCGAACGAAAGGTTTTAGCCTGAAGACGG

Linadd Optimization

Update rule: $f_i \leftarrow f_i^{old} + \sum_{j \in W} (\alpha_j - \alpha_j^{old}) y_j \, \mathrm{k}(x_i, x_j)$

Exploiting $\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ and $\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \Phi(\mathbf{x}_i)$:

$$f_i \leftarrow f_i^{old} + \sum_{j \in W} (\alpha_j - \alpha_j^{old}) y_j \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) = f_i^{old} + \mathbf{w} \cdot \Phi(\mathbf{x}_i)$$

Key Idea: Store $\mathbf{w}$ and compute $\mathbf{w} \cdot \Phi(\mathbf{x})$ *efficiently*

- `Clear:` $\mathbf{w} = \mathbf{0}$
- `Add:` $w_u \leftarrow w_u + v$ (only needed $|W|$ times per iteration)
- `Lookup:` obtain w_u (must be highly efficient)

⇒ speedup of factor 60 (7) for Spectrum (Weighted Degree Kernel) ⇒ parallelized additional factor 2 (5)

Multiple Kernel Learning

- Multiple input domains (binding energies, DNA sequence, ...)
- Kernel $k(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}')$ used in standard SVM Classifier

$$f(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{\ell} y_i \alpha_i \mathrm{k}(\mathbf{x}, \mathbf{x}_i) + b\right)$$

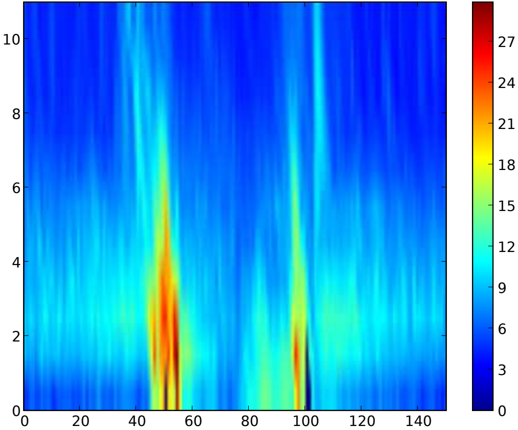
- Now: linear combination of kernels (again a kernel)

$$\mathrm{k}(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{M} \beta_j \, \mathrm{k}_j(\mathbf{x}, \mathbf{x}'), \;\; \beta_j \geq 0$$

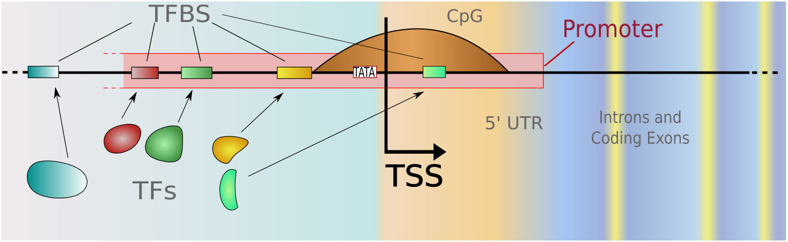
- Possible to learn weights β_j

Questions:

- Where is which $k-$mer of importance ?
- Where is which $k-$mer - length of importance ?



Properties of Transcription Start Sites (TSS)



- POL II binds to a rather vague region of ≈ $[-20, +20]$ bp
- Upstream of TSS: promoter containing transcription factor binding sites
- Downstream of TSS: 5’ UTR, and further downstream coding regions and introns (different oligomer statistics)
- 3D structure of the promoter must allow the transcription factors to bind

Features to describe the TSS

- TFBS in Promotor region
- Condition: DNA should not be too twisted
- CpG islands (often over TSS/first exon; in most, but not all promoters)
- TSS with TATA box (≈ -30 bp upstream)
- Exon content in UTR 5" region
- Distance to first donor splice site

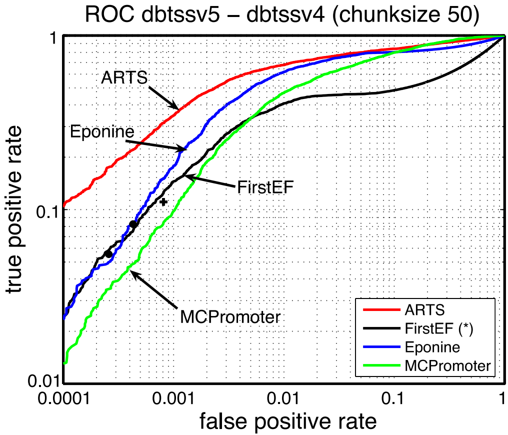
Idea: Combine weak features to build strong promoter predictor

Combine (Five) Sub-Kernels

Simply add up kernel for different features:

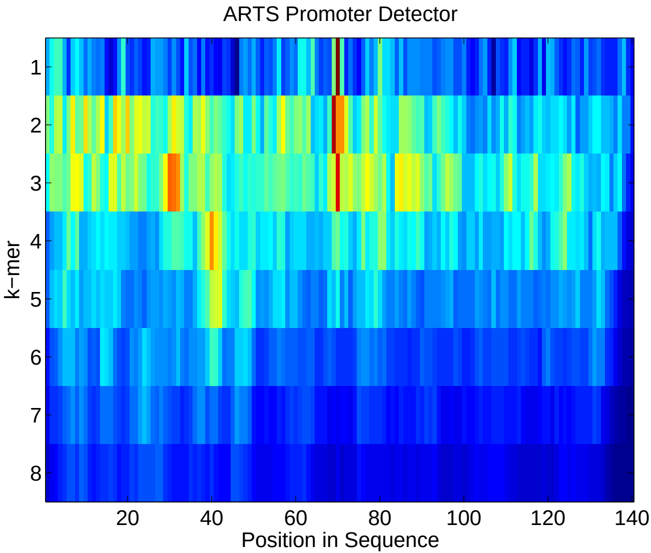
1. TSS signal (including parts of core promoter with TATA box)
   – use **Weighted Degree Shift kernel**
2. CpG Islands, distant enhancers and TFBS upstream of TSS
   – use **Spectrum kernel** (large window upstream of TSS)
3. model UTR and coding sequence downstream of TSS
   – another **Spectrum kernel** (window downstream of TSS)
4. stacking energy of DNA
   – use *btwist* energy of dinucleotides with **linear kernel**
5. twistedness of DNA
   – use btwist angle of dinucleotides with **linear kernel**

Receiver Operator Characteristic Curve



⇒ 35% **true positives at a false positive rate of** 1/1000 **(best other method find about a half (**18%**))**

Overview over Discriminative Features



Conclusions

- Developed a new TSS finder, “ARTS”
- In genome wide evaluation achieves state-of-the-art results: ARTS about 35% true positives at a false positive rate of 1/1000 (best other method about a half, 18%)
- Reason: large scale SVM training/evaluation with string kernels, intensively modelling the TSS region
- Future work:
  - Drosophila, C. elegans, Arabidopsis, ...
  - Motif Discovery
  - Alternative Transcription Start Sites

Availability

Datasets, Genomebrowser custom track, a lot more details:

`http://www.fml.tuebingen.mpg.de/raetsch/projects/arts`

Free source code of SHOGUN toolbox used to train ARTS:

`http://www.fml.tuebingen.mpg.de/raetsch/projects/shogun`

Thank you!