ARTS: Accurate Recognition of Transcription Starts in human

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

FIRST

Fraunhofer Institut
Rechnerarchitektur
und Softwaretechnik

MAX-PLANCK-GESELLSCHAFT

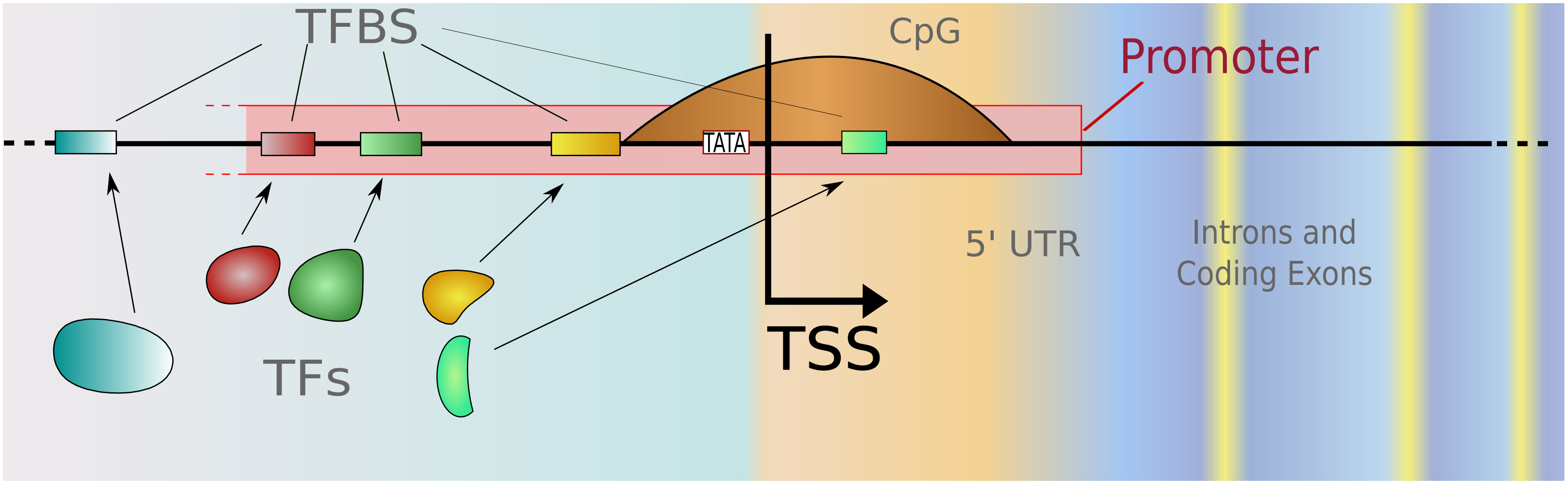
[†] Fraunhofer FIRST.IDA, Kekuléstr. 7, 12489 Berlin, Germany
[♮] Friedrich Miescher Laboratory of the Max Planck Society,
[*] Max Planck Institute for Biological Cybernetics,
Spemannstr. 37-39, 72076 Tübingen, Germany

Soeren.Sonnenburg@first.fraunhofer.de,
{Alexander.Zien,Gunnar.Raetsch}@tuebingen.mpg.de

Overview:

- Transcription Start Site (TSS)
- Features to describe the TSS
- Our approach
- Evaluation with current methods
- Example - Protocadherin-α
- Summary

Transcription Start Site - Properties



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

$\Rightarrow$ **Promoter Prediction is non-trivial**

Features to describe the TSS

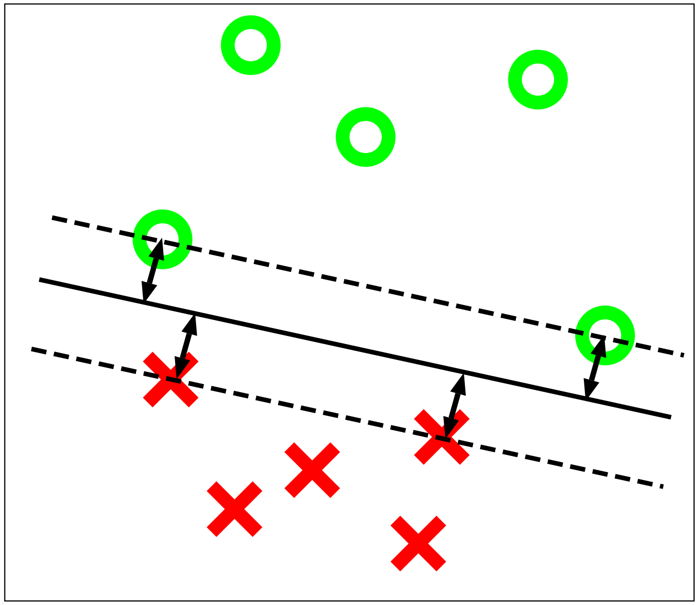
- TFBS in Promoter region
- condition: DNA should not be too twisted
- CpG islands (often over TSS/first exon; in most, but not all promoters)
- TSS with TATA box (≈ -30 bp upstream)
- Exon content in UTR 5" region
- Distance to first donor splice site

Idea: Combine weak features to build strong promoter predictor

THE ARTS APPROACH

- use SVM classifier

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



- key ingredient is kernel $k(\boldsymbol{x}, \boldsymbol{x}')$ — similarity of two sequences
- use 5 sub-kernels suited to model the aforementioned features

$$\mathrm{k}(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{k}_{TSS}(\boldsymbol{x}, \boldsymbol{x}') + \mathrm{k}_{CpG}(\boldsymbol{x}, \boldsymbol{x}') + \mathrm{k}_{coding}(\boldsymbol{x}, \boldsymbol{x}') + \mathrm{k}_{energy}(\boldsymbol{x}, \boldsymbol{x}') + \mathrm{k}_{twist}(\boldsymbol{x}, \boldsymbol{x}')$$

The 5 sub-kernels

1. TSS signal (including parts of core promoter with TATA box)
   – use **Weighted Degree Shift kernel**

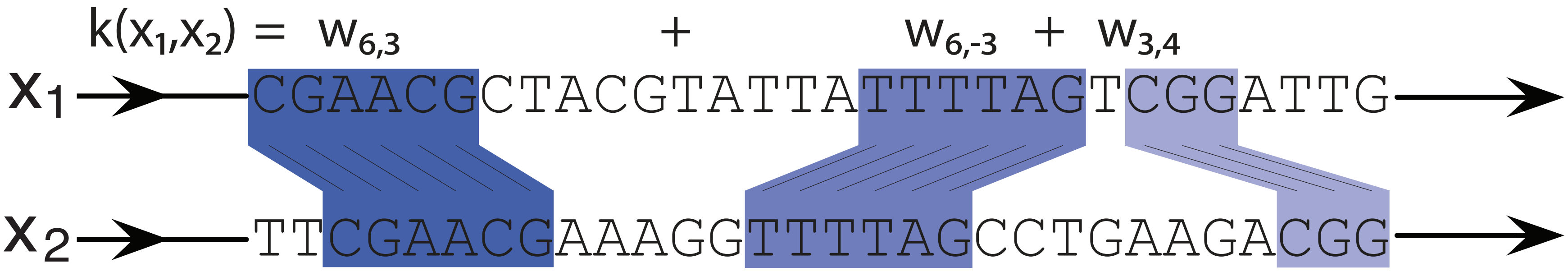
2. CpG Islands, distant enhancers and TFBS upstream of TSS
   – use **Spectrum kernel** (large window upstream of TSS)

3. Model coding sequence TFBS downstream of TSS
   – use another **Spectrum kernel** (small 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**

Weighted Degree Shift Kernel



$k(x_1,x_2) = w_{6,3} + w_{6,-3} + w_{3,4}$

x_1 CGAACGCTACGTATTATTTTAGTCGGATTG

x_2 TTCGAACGAAAGGTTTTAGCCTGAAGACGG

- Count matching substrings of length $1 \ldots d$
- Weight according to length of the match $\beta_1 \ldots \beta_d$
- Position dependent but tolerates “shifts” of up to S

$$\mathrm{k}(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^{d} \beta_k \sum_{l=1}^{L-k+1} \sum_{\substack{s=0 \\ s+l \le L}}^{S} \delta_s \left(\mathbf{I}(\mathbf{x}[k:l+s] = \mathbf{x}'[k:l]) + \mathbf{I}(\mathbf{x}[k:l] = \mathbf{x}'[k:l+s]) \right)$$

$\mathbf{x}[k:l] :=$ subsequence of $\mathbf{x}$ of length k starting at position l

Training – Data Generation

True TSS:

- From dbTSSv4 (based on hg16) extract putative TSS windows of size $[-1000, +1000]$

Decoy TSS:

- Annotate dbTSSv4 with transcription-stop (via *BLAT* alignment of mRNAs)
- From the interior of the gene ($+100bp$ to gene end) sample negatives for training (10 per positive), again windows $[-1000, +1000]$

Processing:

- 8508 positive, 85042 negative examples
- Split into disjoint training and validation set ($50\% : 50\%$)

Training – Model Selection

16 kernel parameters + SVM regularization to be tuned!

- Full grid search infeasible
- Local axis-parallel searches instead

SVM training/evaluation on $> 10,000$ **examples computationally too demanding**

Speedup trick:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N_s} \alpha_i \mathrm{k}(\boldsymbol{x}_i, \boldsymbol{x}) + b = \underbrace{\sum_{i=1}^{N_s} \alpha_i \Phi(\boldsymbol{x}_i)}_{\boldsymbol{w}} \cdot \Phi(\boldsymbol{x}) + b = \boldsymbol{w} \cdot \Phi(\boldsymbol{x}) + b$$

$f(x)$ before: $O(N_s dLS)$ now: $= O(dL) \Rightarrow$ **speedup factor** up to $N_s \cdot S$

$\Rightarrow$ **Large Scale Training and Evaluation possible**

Comparison

Current state-of-the-art methods:

- **FirstEF** [Davuluri, Grosse, Zhang; 2001, Nat Genet]
  QDF: for promoter, donor, first exon, WM
  Range: $[-1500, +500]$

- **McPromoter** [Ohler, Liao, Niemann, Rubin; 2002, Genome Biol]
  GHMM with IMC for 6 regions (e.g. upstream, TATA) NN
  Range: $[-250, +50]$

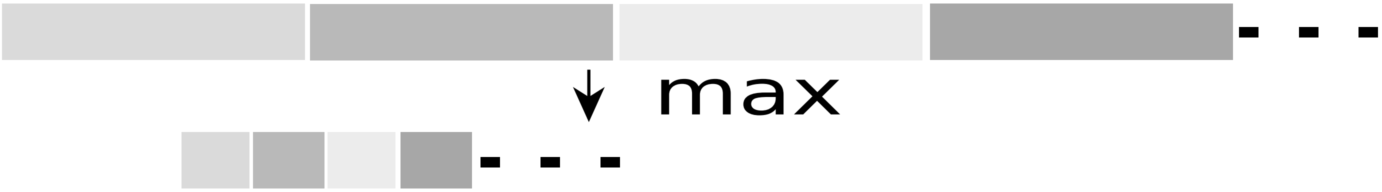
- **Eponine** [Down, Hubbard; 2002 Genome Res]
  RVM: WM with positional distribution for 4 regions (e.g. TATA, CpG)
  Range: $[-200, +200]$

⇒ **Do a genome wide evaluation!**
⇒ **How to do a fair comparison?**

Evaluation

Idea: Only consider “new” TSS from dbTSSv5-dbTSSv4, with max 30% overlap

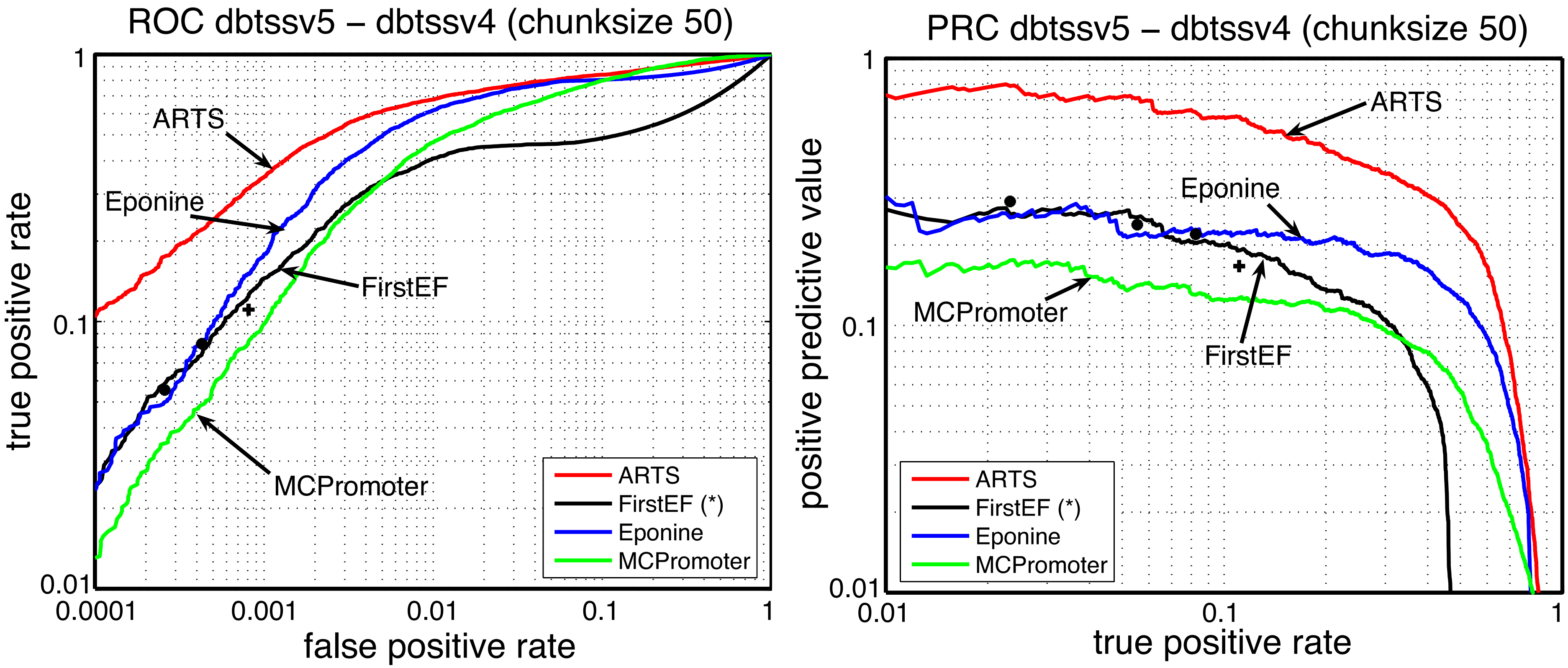
1. Compute genome wide outputs for each TSF
2. Decrease resolution: divide genome into non-overlapping fixed size chunks (e.g. 50 or 500)



3. Annotate dbTSSv5 TSS with gene end
4. Label chunk positive if intersects with $[TSS - 20bp, TSS + 20bp]$
5. Label chunk negative $[TSS + 21bp, GeneEnd]$

Results

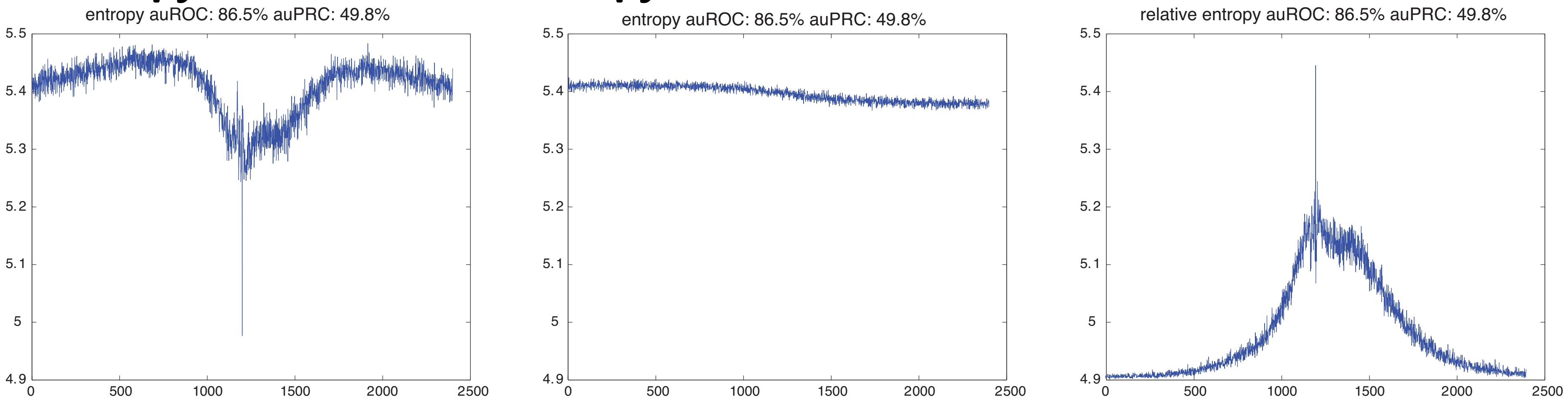
Receiver Operator Characteristic Curve *and* Precision Recall Curve



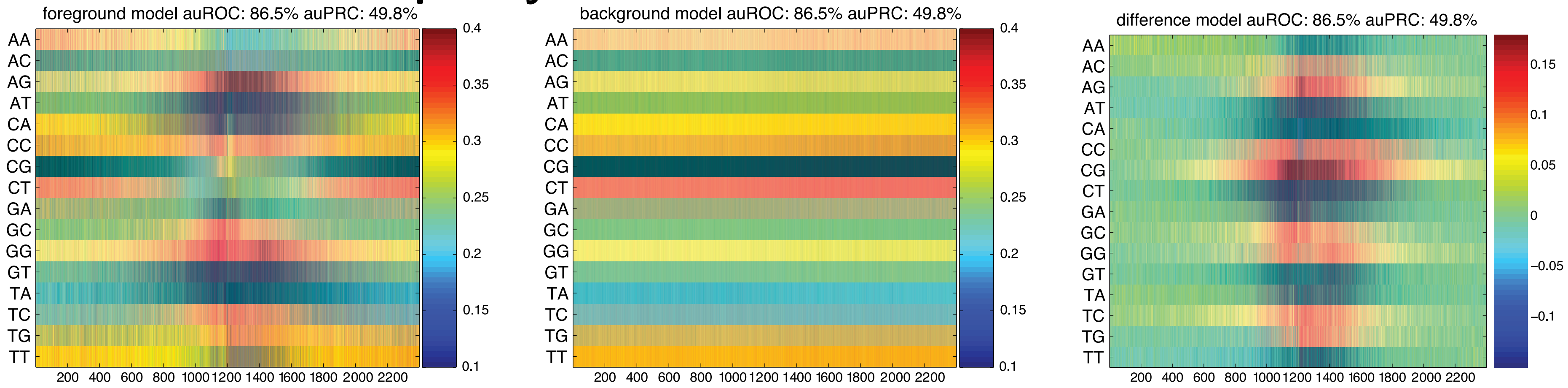
⇒ 35% true positives at a false positive rate of $1/1000$ (best other method find about a half (18%))

What does ARTS do better ?

Entropy and Relative Entropy

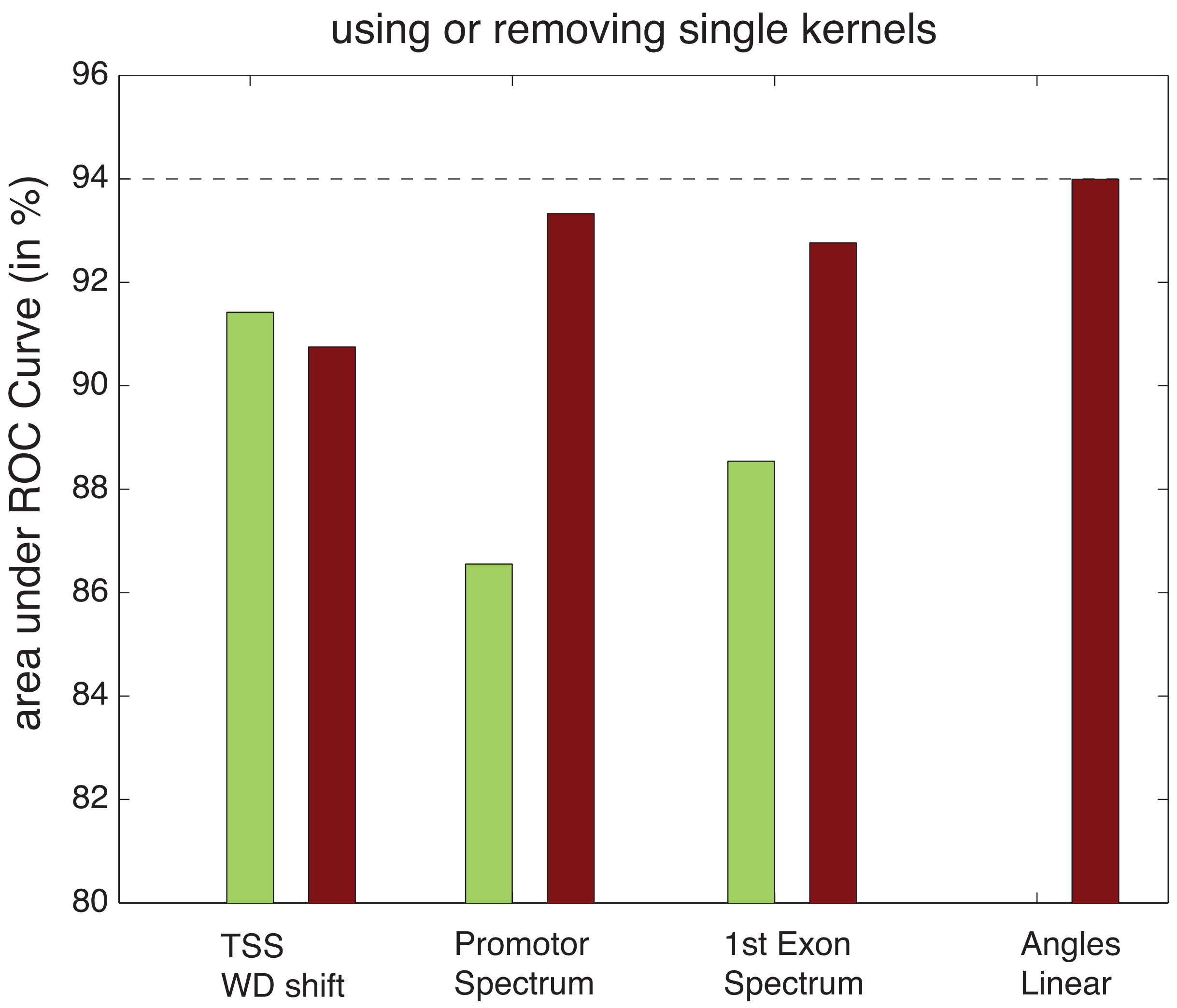


Di-nucleotide Frequency



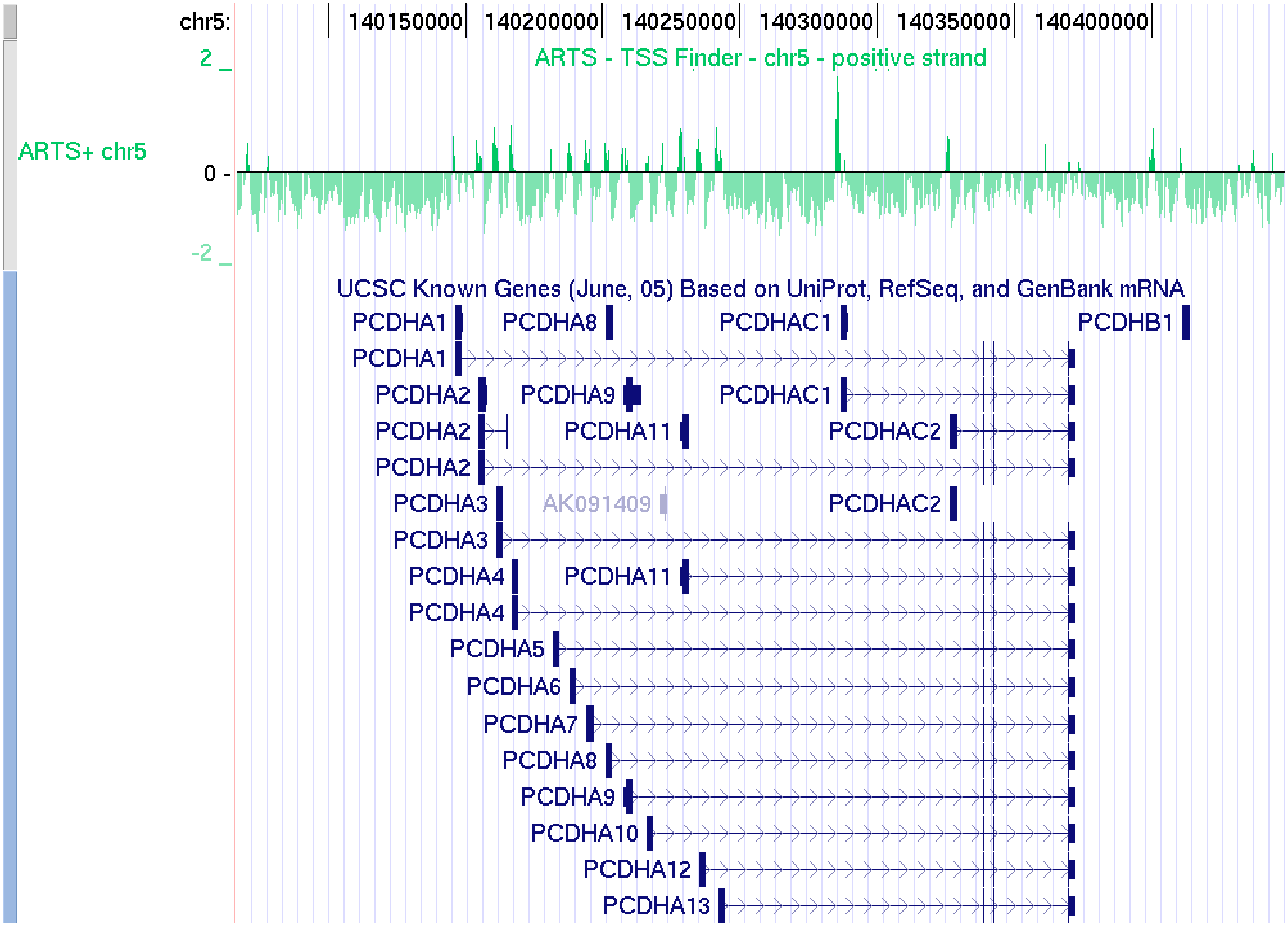
⇒ strong discriminative signal around TSS

Which kernel captures most information ?



⇒ Most important Weighted Degree Shift kernel modelling the TSS signal

Alternative TSS - Protocadherin-α



Conclusion

- Developed a new TSF 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: intensively modelling the TSS region, large scale svm training/evaluation with string kernels
- Future work: Drosophila, C.elegans, Zebrafish,. . .

Poster:

H56

Datasets, Genomebrowser custom track, a lot more details:

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

Source code of SHOGUN toolbox used to train ARTS *freely* available:

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