Current Bioinformatics Projects

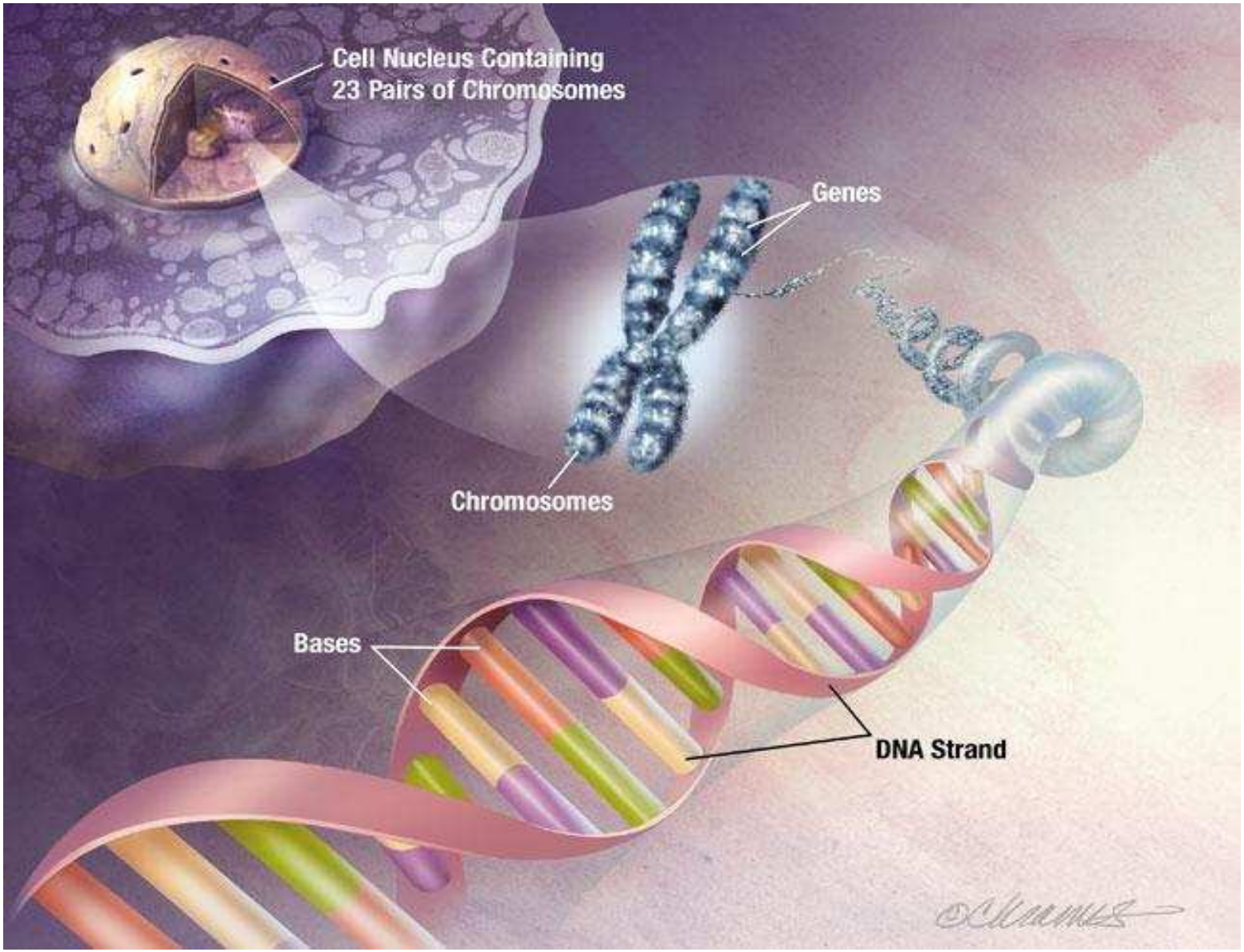
Sören Sonnenburg[†], Mikio Braun,[†] Petra Philipps[♮], Cheng Soon Ong[♮],
Alex Zien,[♮] Gunnar Rätsch[♮], Klaus-Robert Müller[†]

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

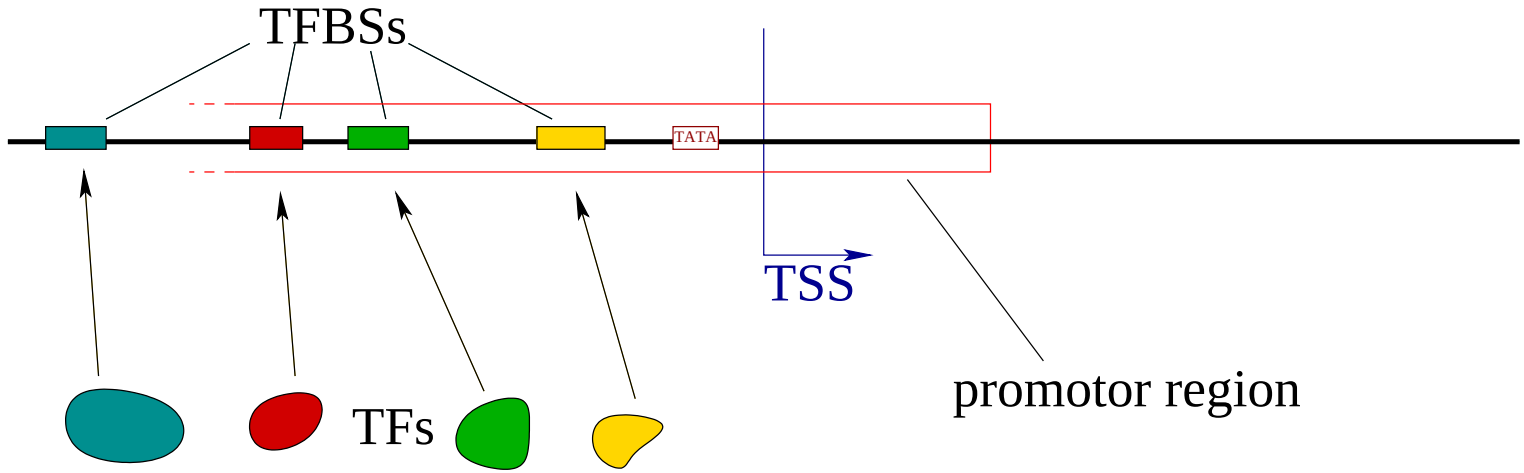
Project Overview:

- Promoter Prediction
- Splice Form Prediction
- Alternative Splicing
- TFBS Module Discovery
- siRNA Prediction
- PolyA Prediction
- Genome wide SNP discovery on *A. Thaliana*
- Secondary Structure Prediction

Promoter Prediction



Definition



Properties:

- TSS has no exact location, more like a range of $[-20, +20]$ base pairs
- TSS - no consensus sequence
- position, order and number of TFBS in Promotor region variable

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

Features

- TFBS in Promotor region
- condition: DNA should not be too twisted
- CpG islands (often over TSS/first exon, seem to be 2 general types CpG and non-CpG island promotors)
- TSS with TATA box (≈ -30 bp upstream)
- exon content in UTR 5" region

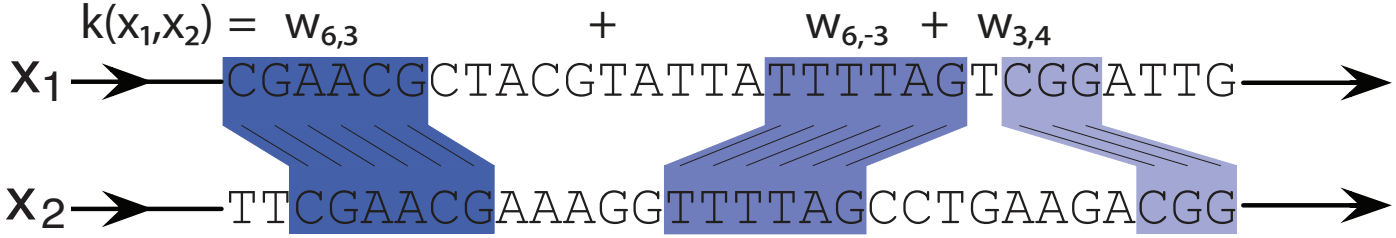
Idea: Combine weak features to build strong promotor predictor

Method

Use Support Vector Machine Classifier classifier

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

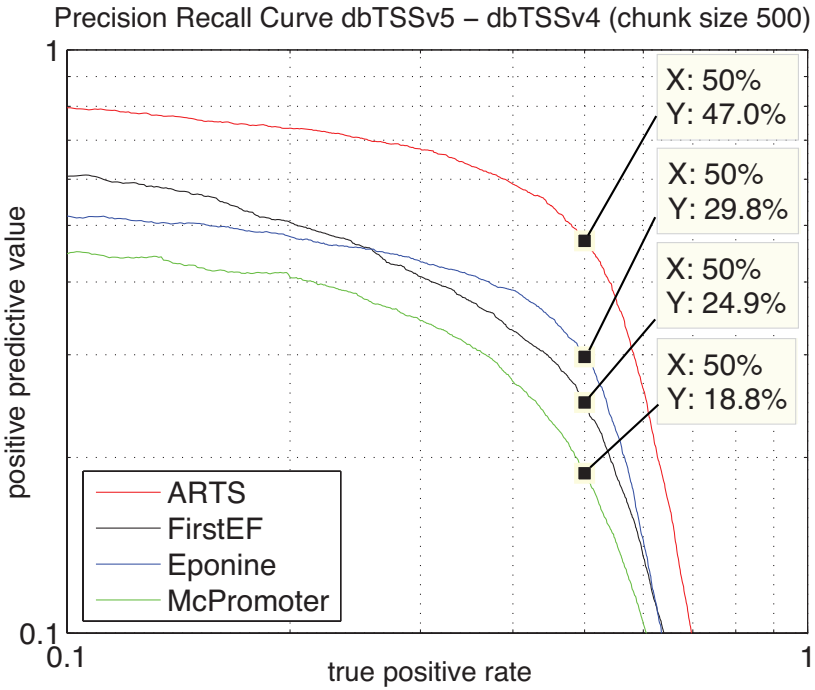
- key ingredient is kernel $k(\boldsymbol{x}, \boldsymbol{x}')$ – gives means to compare 2 sequences
- to detect TSS (including parts of core promotor with TATA box) – use Weighted Degree Shift kernel

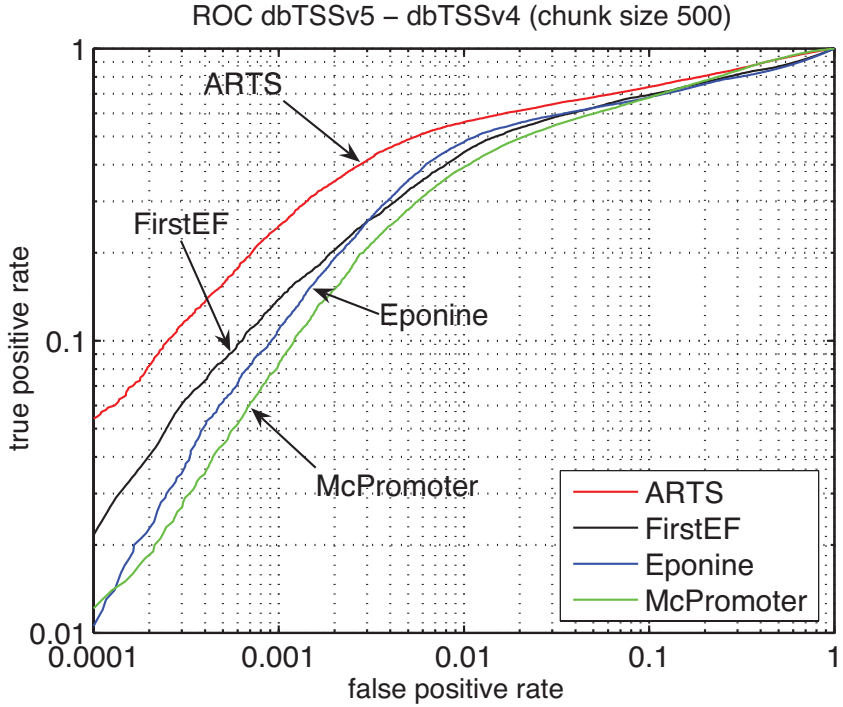


- CpG Islands, distant Enhancers and TFBS upstream of TSS, coding sequence downstream (Spectrum kernel)
- stacking energy of DNA (linear kernel)

Results

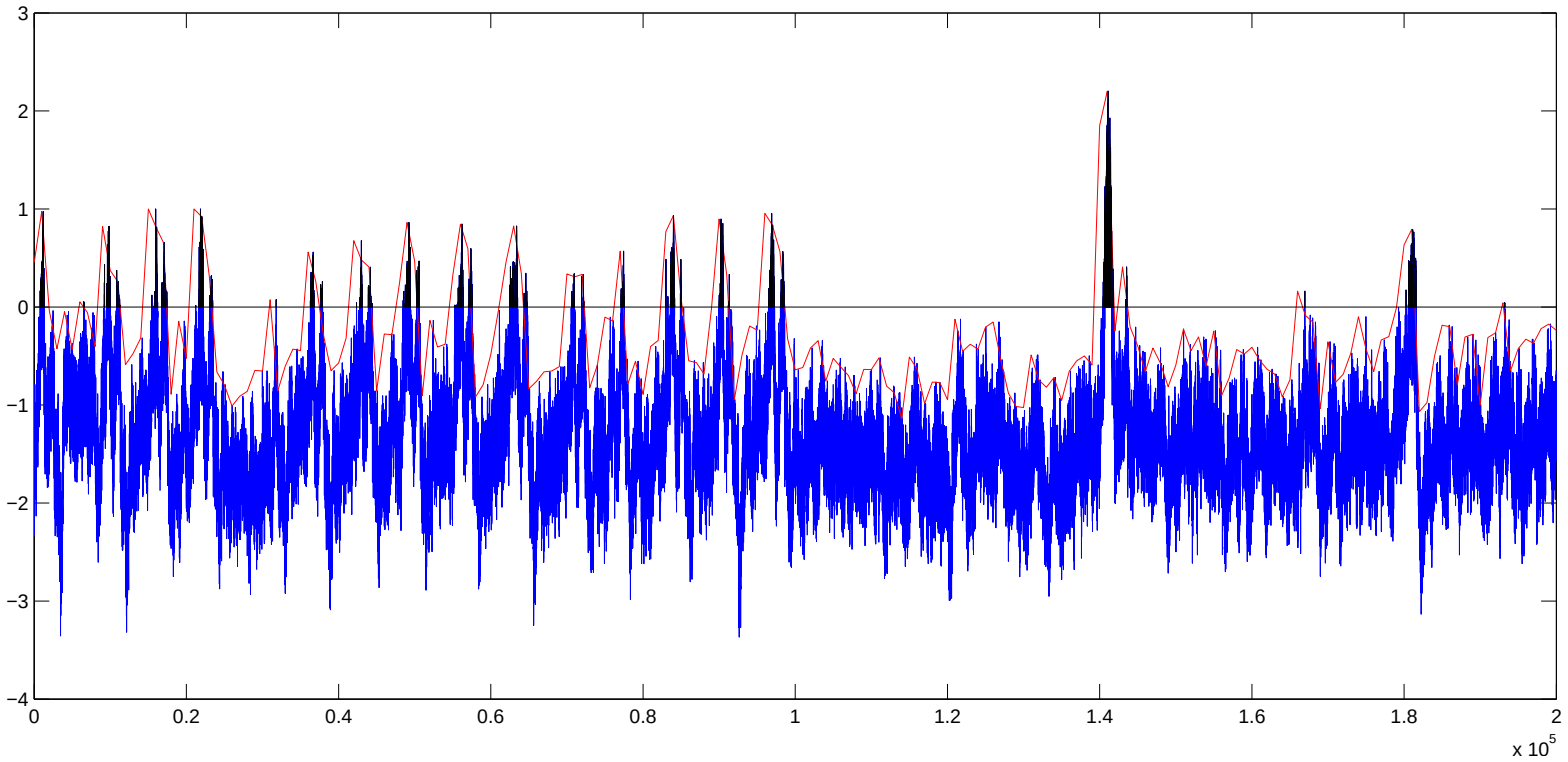
Genomewide (human) evaluation





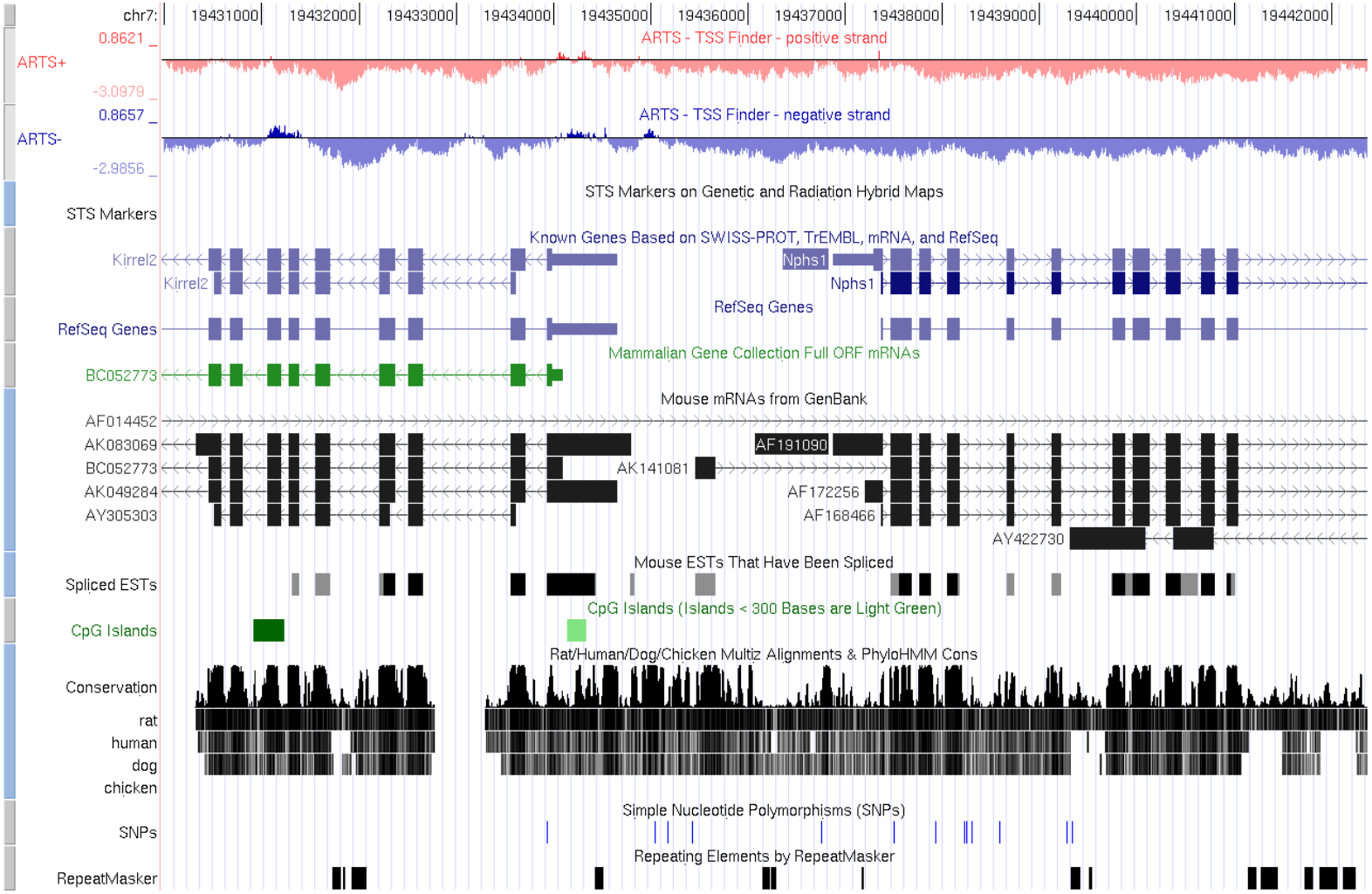
Example

Protocadherin Alpha

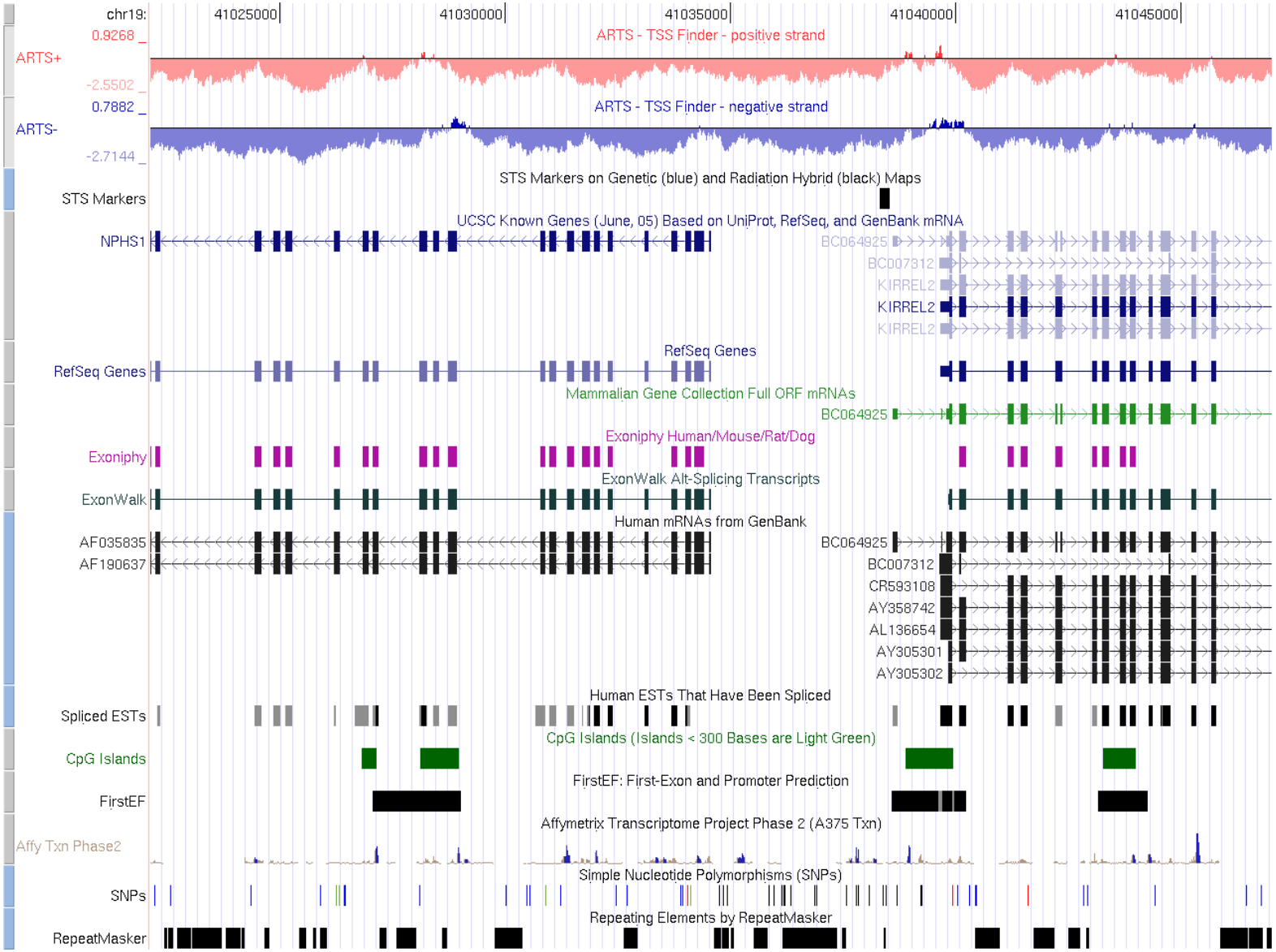


⇒ **might even detect alternative promoters**
⇒ **accepted at ISMB 2006**

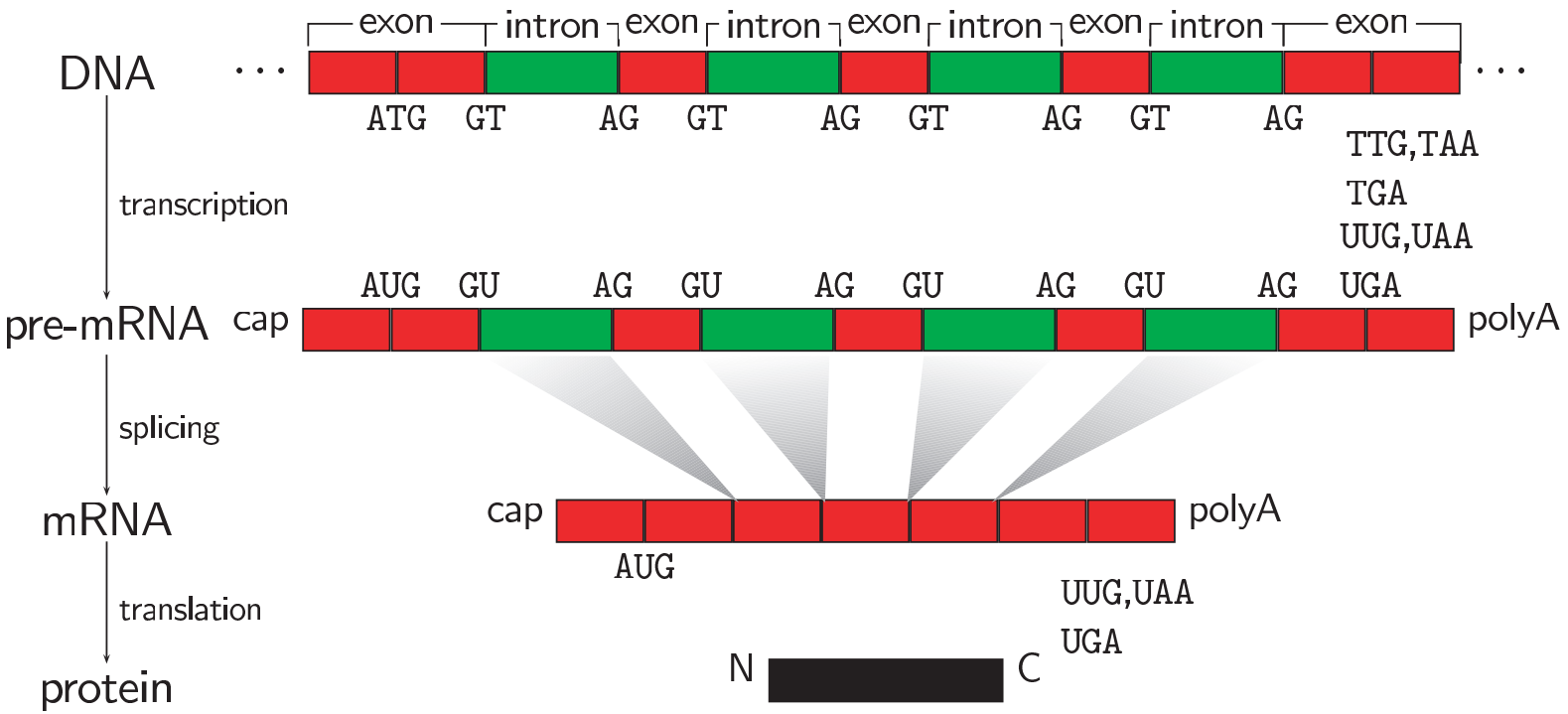
Nephrin mm5


chr7:
19431000
19432000
19433000
19434000
19435000
19436000
19437000
19438000
19439000
19440000
19441000
19442000
0.8621
ARTS+
-3.0979
ARTS - TSS Finder - positive strand
0.8657
ARTS-
-2.9856
ARTS - TSS Finder - negative strand
STS Markers on Genetic and Radiation Hybrid Maps
STS Markers
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
Kirrel2
Kirrel2
Nphs1
Nphs1
RefSeq Genes
RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
BC052773
Mouse mRNAs from GenBank
AF014452
AK083069
BC052773
AK049284
AY305303
AK141081
AF191090
AF172256
AF168466
AY422730
Mouse ESTs That Have Been Spliced
Spliced ESTs
CpG Islands (Islands < 300 Bases are Light Green)
CpG Islands
Rat/Human/Dog/Chicken Multiz Alignments & PhyloHMM Cons
Conservation
rat
human
dog
chicken
Simple Nucleotide Polymorphisms (SNPs)
SNPs
Repeating Elements by RepeatMasker
RepeatMasker

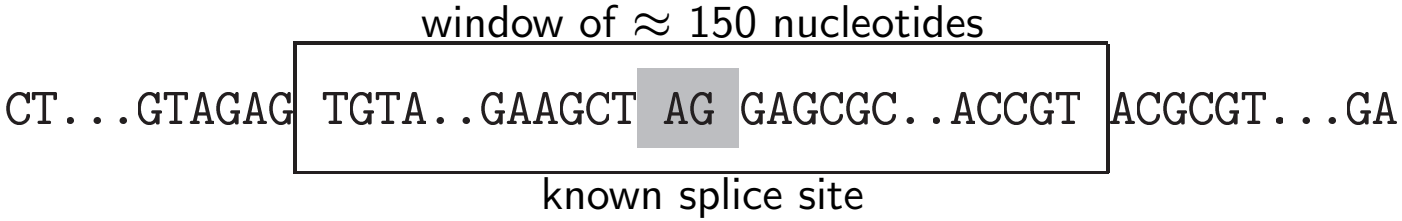
Nephrin hg17



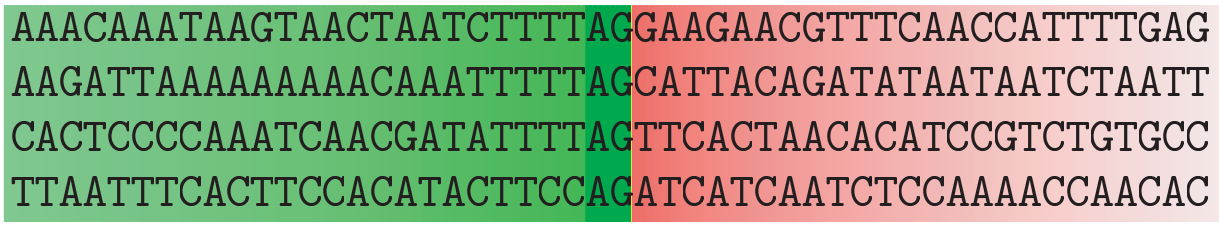
Splice Form Prediction



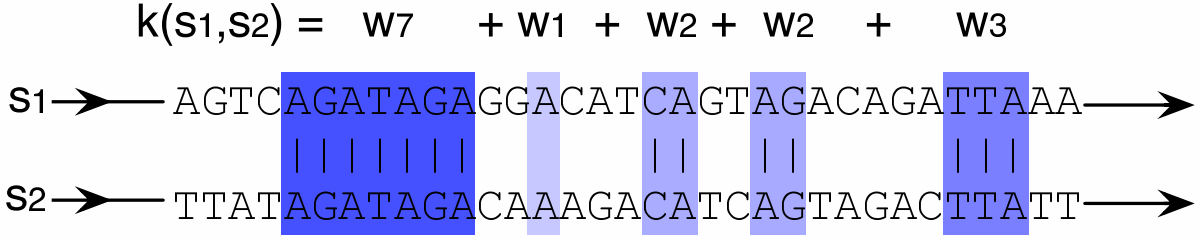
2-class Splice Site Detection



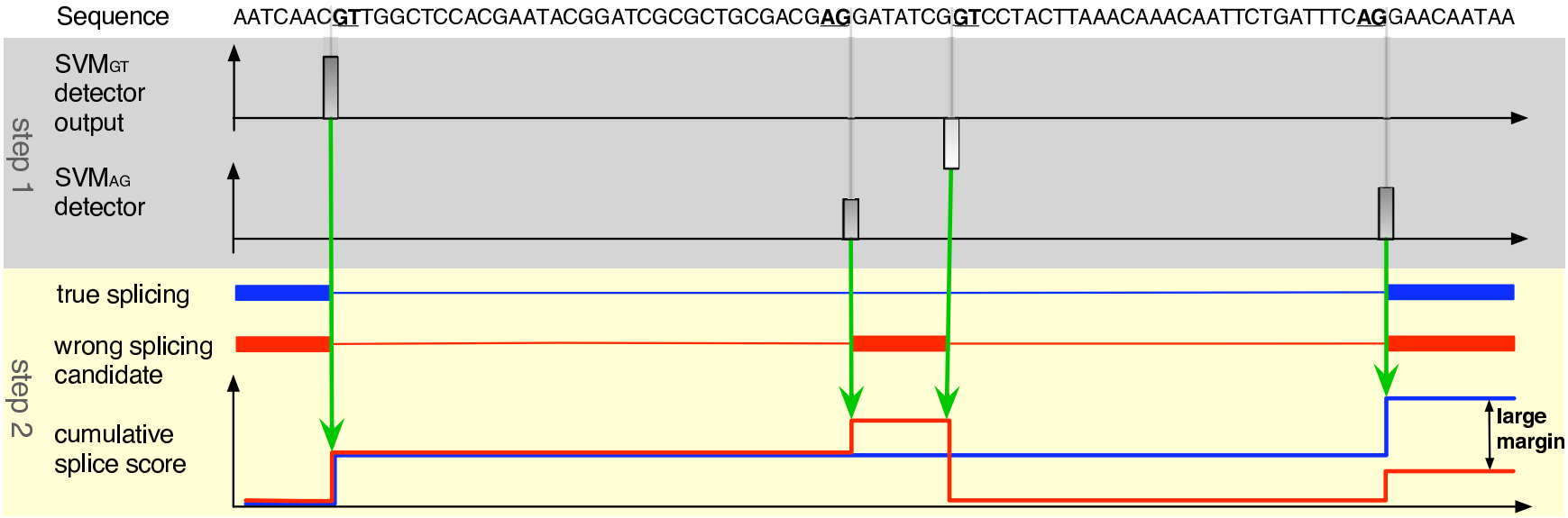
- true sites: fixed window around a true splice site
- decoys sites: generated by shifting the window



Design of new Support Vector kernel that allows appropriate processing:



Large Margin Splice Form Pred.

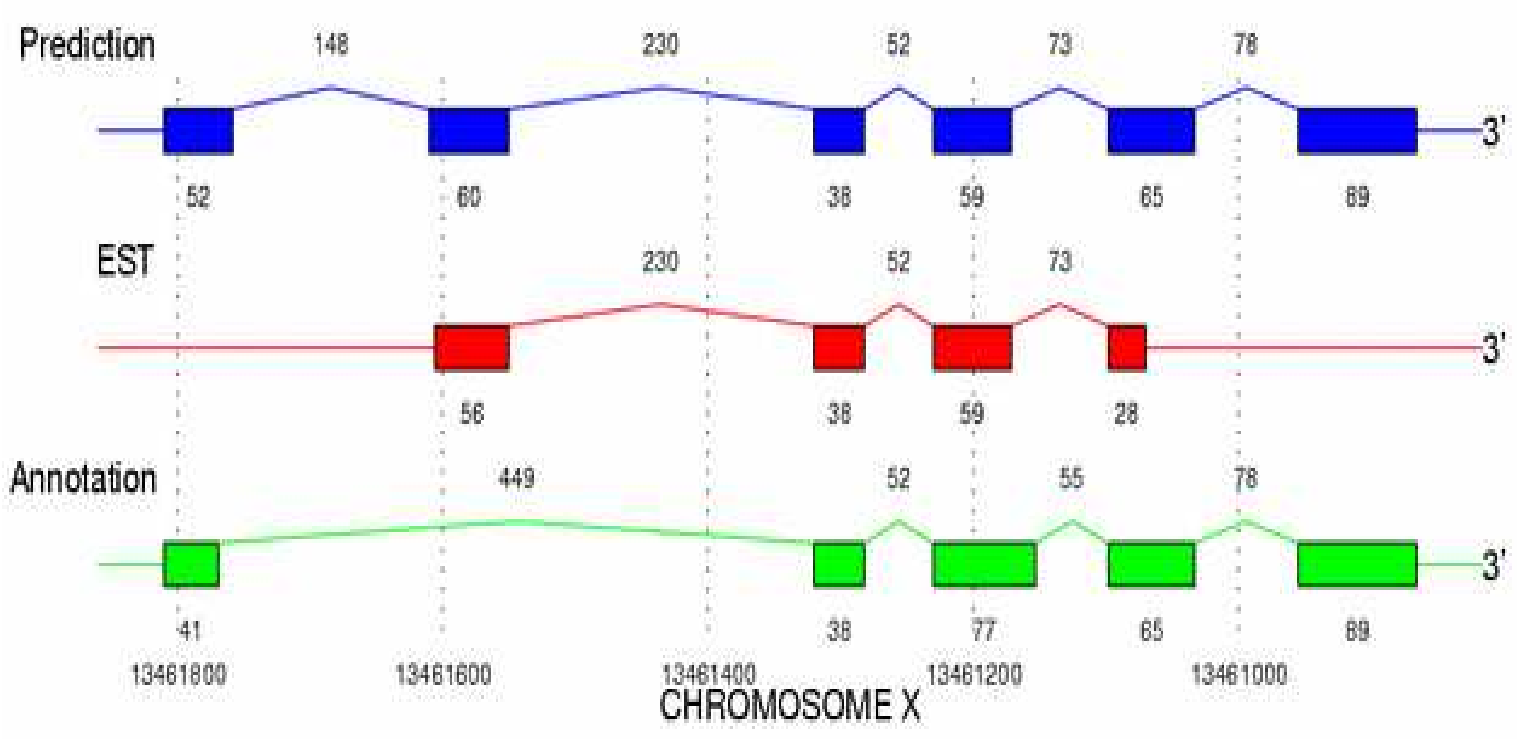


$$S := S_{L_{E,f}}(p_1^{\mathrm{GT}} - p_s) + S_E(s_{[p_s,p_1^{\mathrm{GT}}]}) + S_{L_{E,l}}(p_e - p_n^{\mathrm{AG}}) + S_E(s_{[p_n^{\mathrm{AG}},p_e]})$$

$$+ \sum_{i=1}^{n} \left[S_{L_I}(p_i^{\mathrm{AG}} - p_i^{\mathrm{GT}}) + S_I(s_{[p_i^{\mathrm{GT}},p_i^{\mathrm{AG}}]}) + S_{\mathrm{AG}}(p_i^{\mathrm{AG}}) + S_{\mathrm{GT}}(p_i^{\mathrm{GT}}) \right]$$

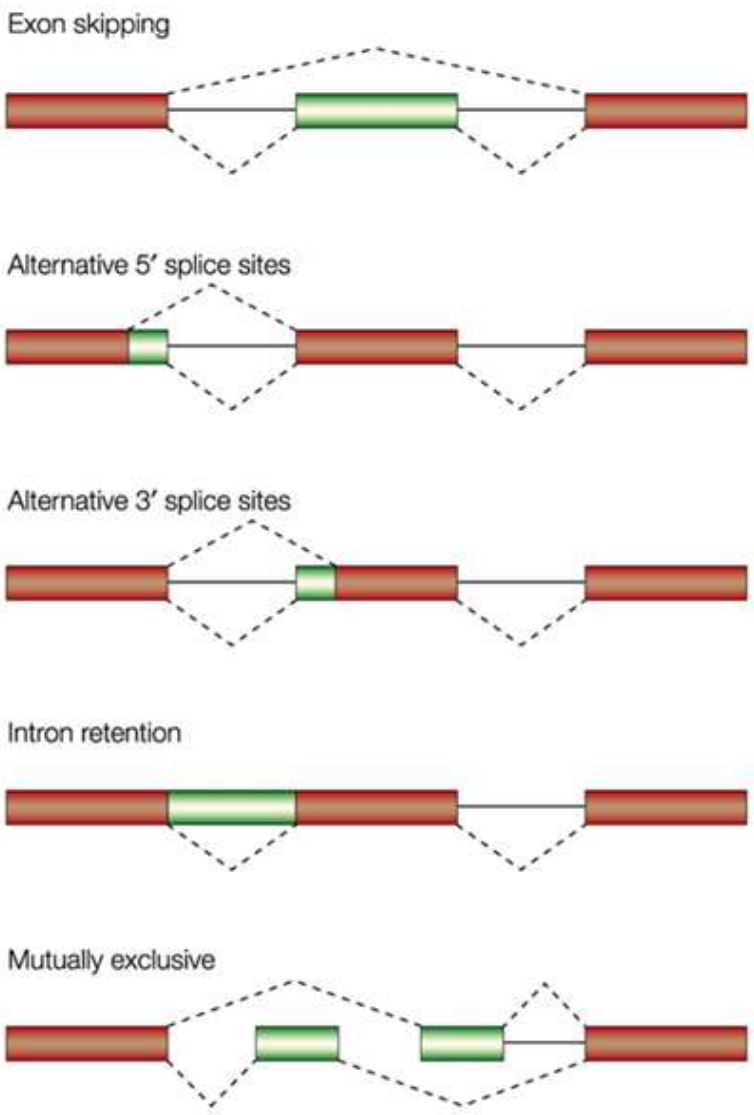
$$+ \sum_{i=1}^{n-1} \left[S_E(s_{[p_i^{\mathrm{AG}},p_{i+1}^{\mathrm{GT}}]}) + S_{L_E}(p_i^{\mathrm{AG}} - p_{i+1}^{\mathrm{GT}}) \right]$$

Results



⇒ **submitted to PLoS computational biology**

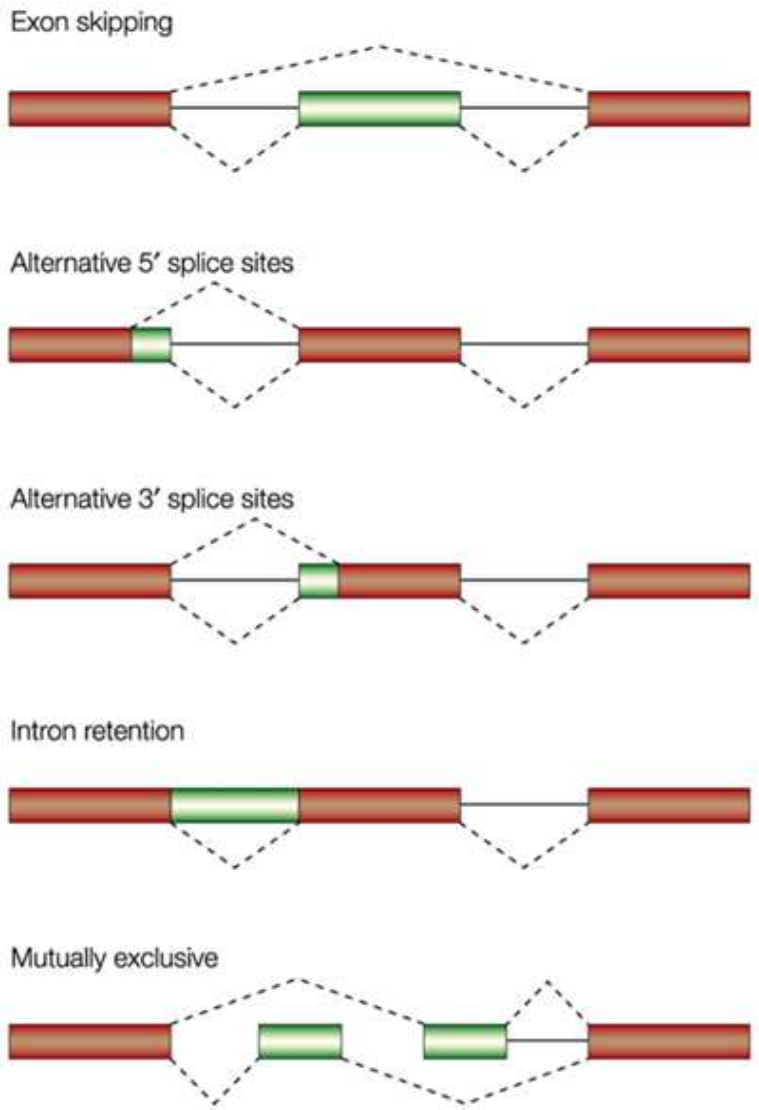
Alternative Splicing

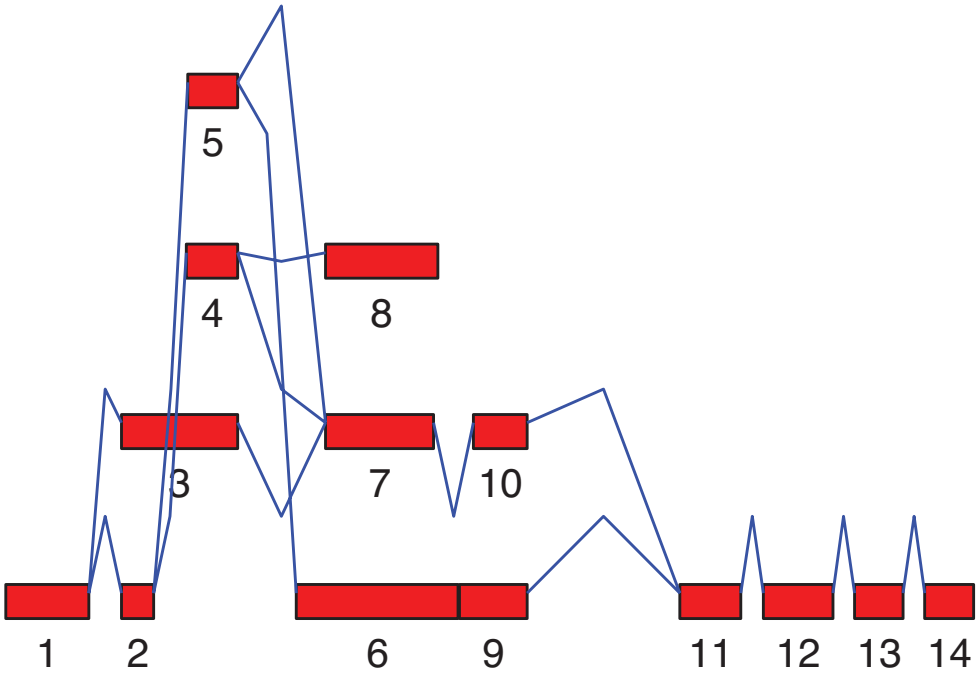


Idea: Use *Machine Learning* to

- analyze sequences near splice sites
- understand differences between alternative and constitutive splicing
- exploit and identify regulative splicing elements
- predict yet unknown alternative splicing events

Forms of Alternative Splicing

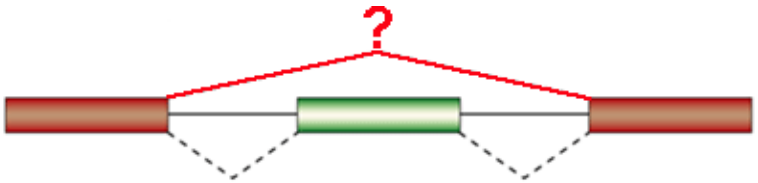




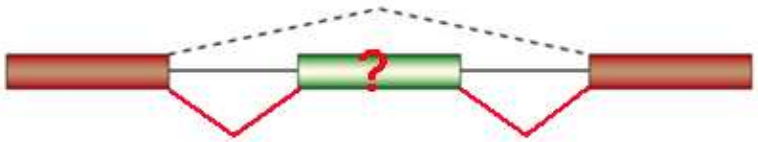
(*C. elegans* gene T08B2.5)

Prediction of Alt. Spliced Exons

- Two different Tasks:
  - Exon is known
  - Can it be skipped?



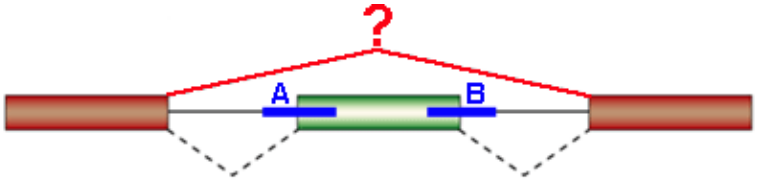
  - Intron is known
  - Does it contain an exon?



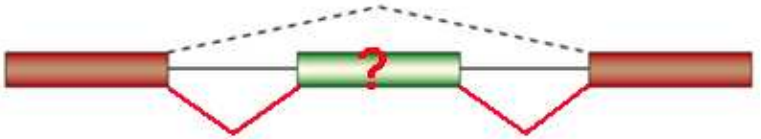
Prediction of Alt. Spliced Exons

- Two different Tasks:

  - Exon is known
  - Can it be skipped?



  - Intron is known
  - Does it contain an exon?



Solution to Task 1

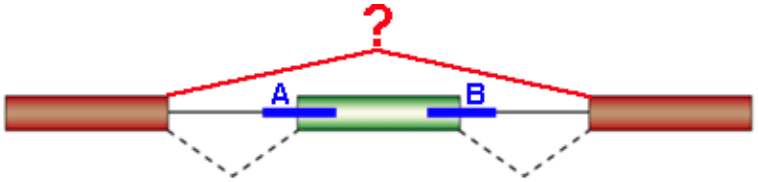
- Two-class Classification Problem

$\Rightarrow$ Use Support Vector Machines (SVMs) on

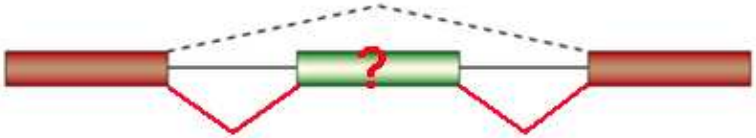
  - sequences A & B ($\pm$ 100nt of splice sites)
  - exon & intron lengths

Prediction of Alt. Spliced Exons

- Two different Tasks:
  - Exon is known
  - Can it be skipped?



  - Intron is known
  - Does it contain an exon?

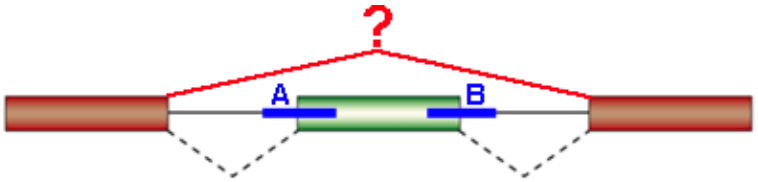


Solution to Task 2

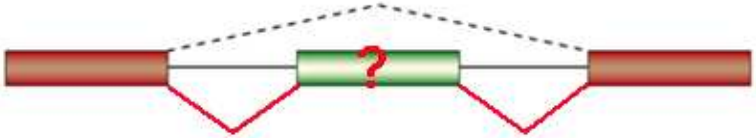
- *Problem:* We do not know yet the exon boundaries!
- *Solution:* Consider all possible exons within the intron.

Prediction of Alt. Spliced Exons

- Two different Tasks:
  - Exon is known
  - Can it be skipped?



  - Intron is known
  - Does it contain an exon?



Solution to Task 2

- *Problem:* We do not know yet the exon boundaries!
- *Solution:* Consider all possible exons within the intron.
- Classify true exons *vs.* wrong exons

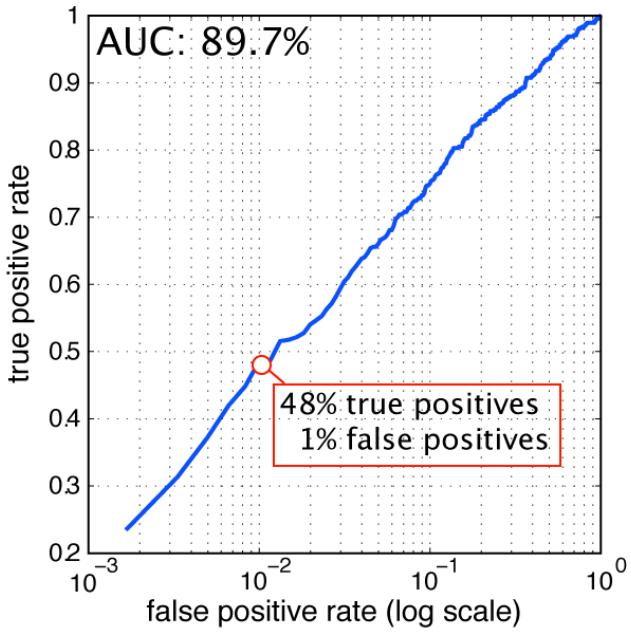
$\Rightarrow$ Use SVM-like algorithm using the

- sequences A & B ($\pm$ 100nt of splice sites)
- exon & intron lengths and
- *splice site scores* (SVM based)

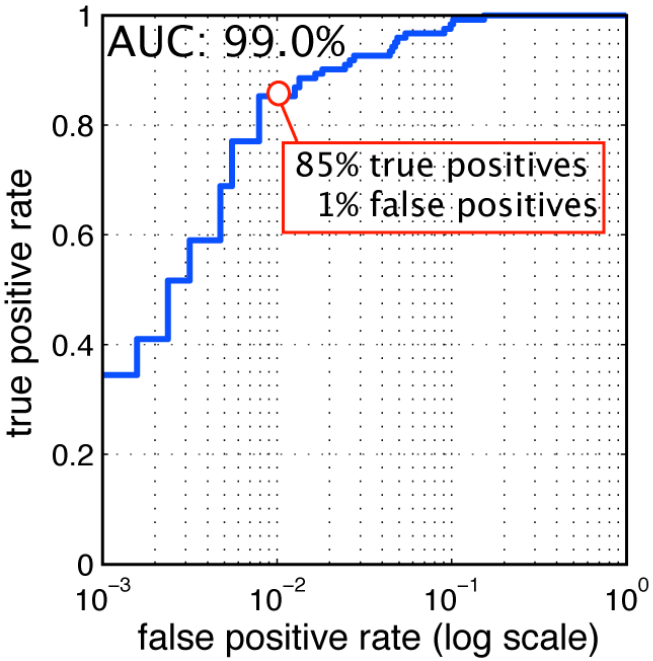
Computational Results

- 487 alternatively and 2531 constitutively spliced exons

Task 1 (exon known)



Task 2 (intron known)



As accurate as for conserved alternatively spliced exons in human.

$\Rightarrow$ Conservation not needed!

$\Rightarrow$ **presented at ISMB 2005**

TFBS Module Discovery

project just started. . .

- given *genome wide* binding sites locations for MEF2 binding genes
- task1: find binding sites that form modules with MEF2 associated BS
- task2: genes are active in different development stages (early,late,always) - which are the discriminating TFBS ?

Future

- Genefinding
- Alternative Promoters
- Quantitave Predictions of Alternative Splicing
- Mass Spectrography
- . . .

Questions ?