Detection of Alternative Splicing Events Using Machine Learning

G. Rätsch[1], S. Sonnenburg[2], B. Schölkopf[3], R. Bohnert[1], C.S. Ong[1,3] and H. Shin[1]

[1] Friedrich Miescher Laboratory, Tübingen, Germany

[2] Fraunhofer FIRST, Berlin, Germany

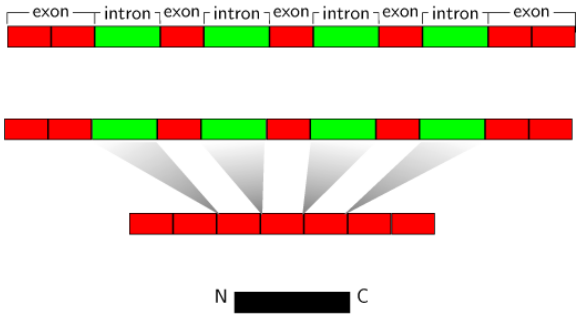
[3] Max Planck Institute for Biol. Cybernetics, Tübingen, Germany

http://www.fml.tuebingen.mpg.de/raetsch

Background & Motivation

From DNA to protein

- genes organized in exons & introns
- transcribe DNA to pre-mRNA
- Splicing removes introns $\Rightarrow$ mRNA
- mRNA is translated into protein



Alternative Splicing (AS)

- can produce several mRNA transcript per gene
  (sometimes leading to $\gg$100 slightly different proteins)
- is highly regulated
- greatly increases the proteome diversity in eukaryotes

$\geq$70% of human genes are alternatively spliced!

Motivation

Alternative Splicing (AS)

- can produce several mRNA transcript per gene
  (sometimes leading to $\gg$100 slightly different proteins)
- is highly regulated
- greatly increases the proteome diversity in eukaryotes

$\geq$70% of human genes are alternatively spliced!

Motivation

Alternative Splicing (AS)

- can produce several mRNA transcript per gene
  (sometimes leading to $\gg$100 slightly different proteins)
- is highly regulated
- greatly increases the proteome diversity in eukaryotes

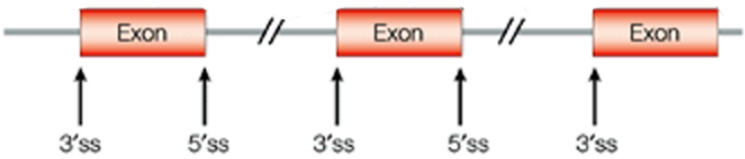
$\geq$70% of human genes are alternatively spliced!

Methods for identifying alternative splicing

- usually need many EST sequences or
- exploit conservation between several organisms

Novel AS prediction method only using the pre-mRNA

Splicing



Splice sites are

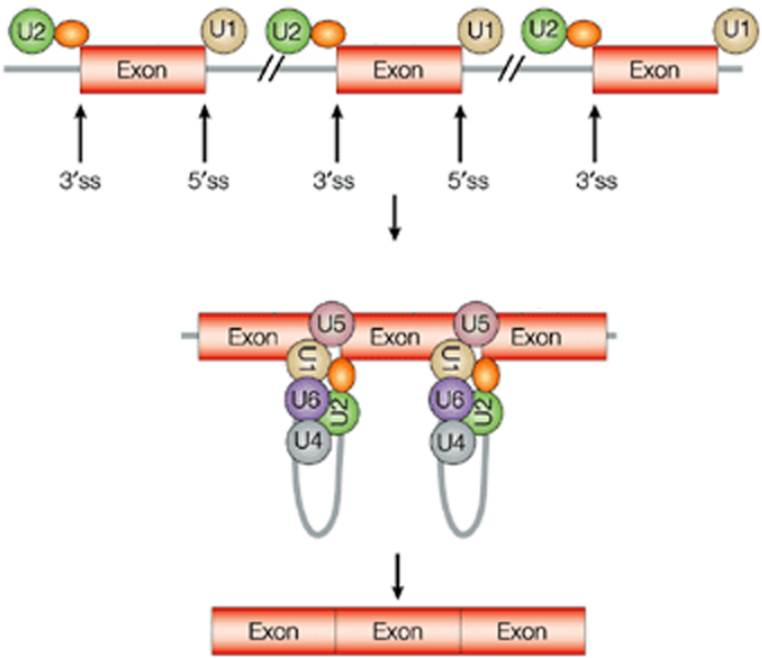
- the exon/intron boundaries

Splicing

Splice sites are

- the exon/intron boundaries
- recognized by five snRNAs assembled in snRNPs

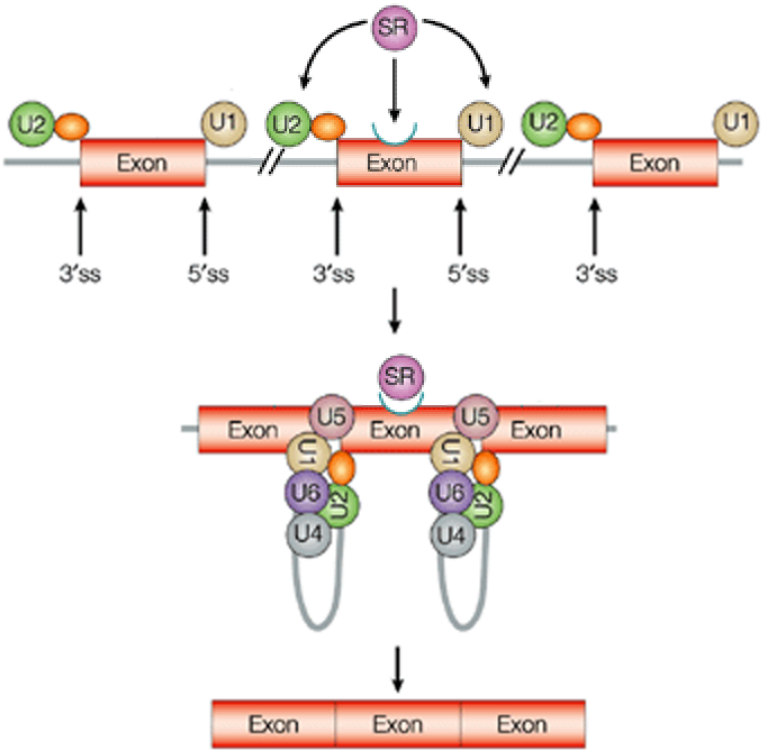
Splicing



Splice sites are

- the exon/intron boundaries
- recognized by five snRNAs assembled in snRNPs

Splicing



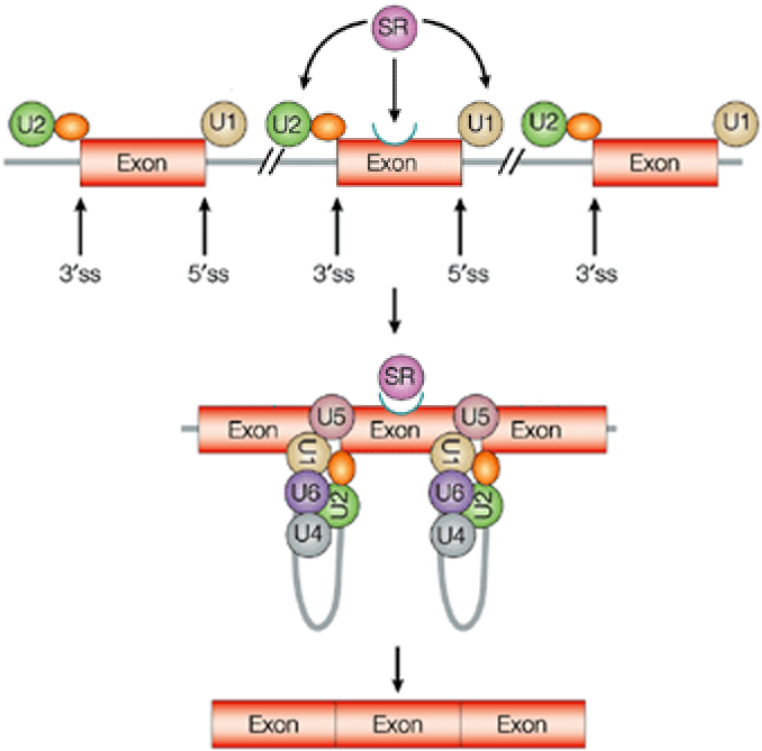
Splice sites are

- the exon/intron boundaries
- recognized by five snRNAs assembled in snRNPs
- flanked by regulatory elements

Spliceosomal Proteins

- interact with snRNPs and mRNA
- regulate recognition of splice sites

Splicing



Splice sites are

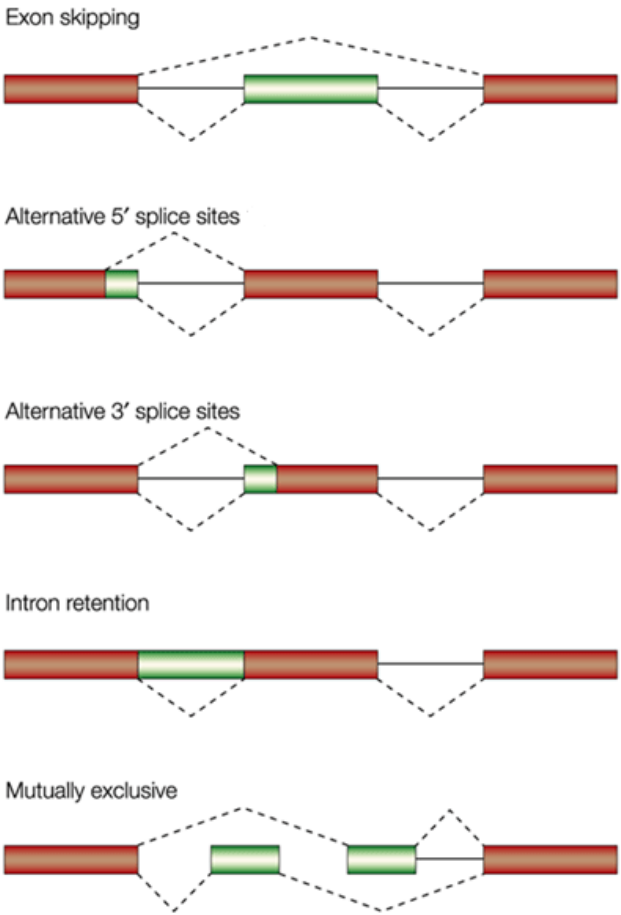
- the exon/intron boundaries
- recognized by five snRNAs assembled in snRNPs
- flanked by regulatory elements

Spliceosomal Proteins

- interact with snRNPs and mRNA
- regulate recognition of splice sites
- can lead to alternative transcripts

One gene may correspond to several transcripts/proteins

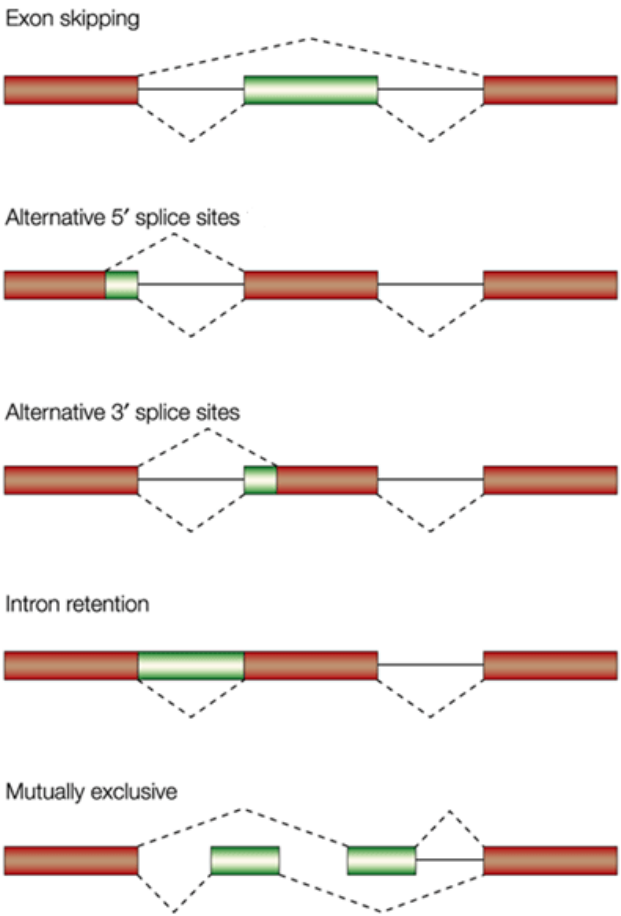
Forms of Alternative Splicing

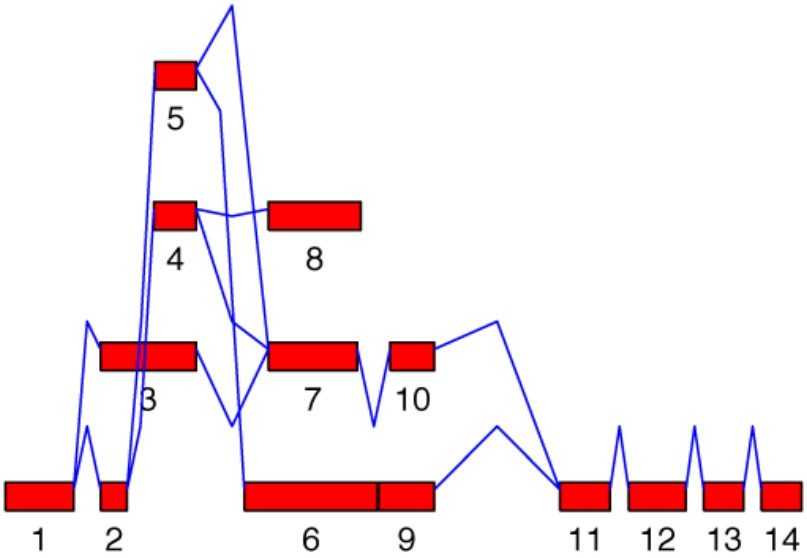


Idea: Use Machine Learning to

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

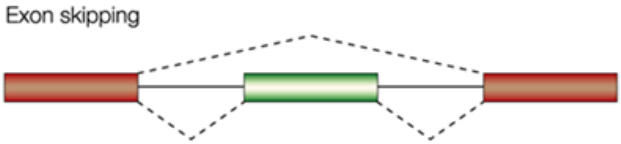
Forms of Alternative Splicing





(*C. elegans* gene T08B2.5)

Alternatively Spliced Exons



Idea: Use Machine Learning to

- understand differences between alternative and constitutive splicing
- exploit and identify regulative elements
- predict unknown alternative splicing events

Previous work:

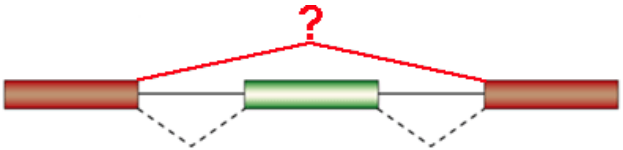
- Analysis of *conserved alternatively spliced exons*
  ⇒ Sorek et al., Yeo et al. and others
  - consider *conserved* alternative spliced exons (ACE)
  - exploit that ACE and flanking introns are more conserved between mouse and human
- Problem: only works for conserved exons

Difference to our approach:

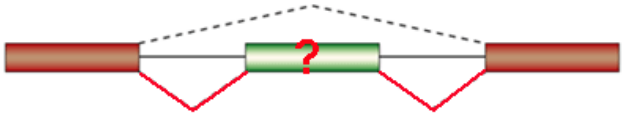
- we only use features derived from the pre-mRNA

Prediction of Alt. Spliced Exons

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

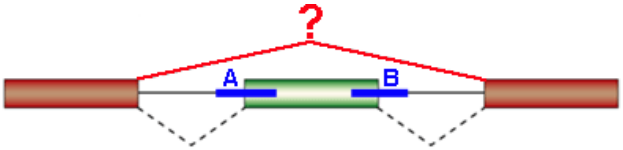


  - Intron is known
  - Does it contain an exon?

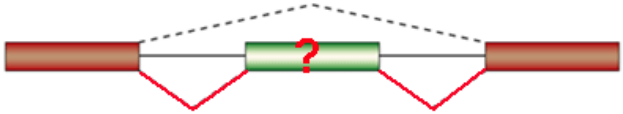


Prediction of Alt. Spliced Exons

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



  - Intron is known
  - Does it contain an exon?



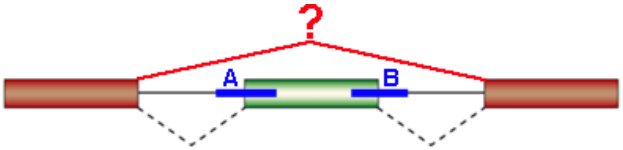
Solution to Task 1

- Two-class Classification Problem

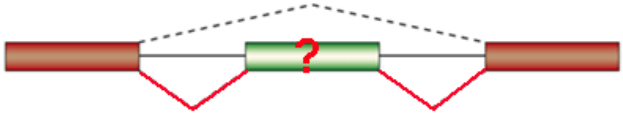
$\Rightarrow$ Use Support Vector Machines (SVMs) on
  - sequences A & B ($\pm$ 100nt of splice sites)
  - exon & intron lengths

Prediction of Alt. Spliced Exons

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



- Intron is known
- Does it contain an exon?

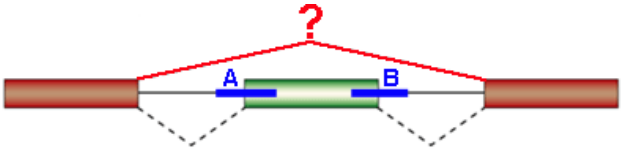


Solution to Task 2

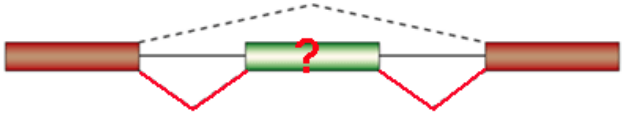
- Problem: We do not know yet the exon boundaries!
- Solution: Consider all possible exons within the intron.

Prediction of Alt. Spliced Exons

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



  - Intron is known
  - Does it contain an exon?



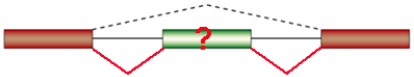
Solution to Task 2

- 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)

Generating Decoys on the Fly



- find function $f(\mathbf{e})$ that scores exons $\mathbf{e}$:
- Training examples:
  - Introns that contain exon $\mathbf{e}_i^+$ and many decoys $\mathbf{e}_{i,j}^-$

$$f(\mathbf{e}_i^+) \geq 1 - \xi_i$$

$$f(\mathbf{e}_{i,j}^-) \leq -1 + \xi_i$$

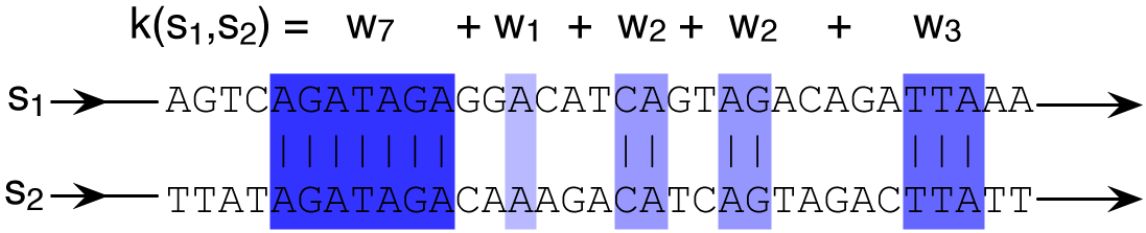
  - Introns without exons with many decoys $\mathbf{e}_{i,j}^-$

$$f(\mathbf{e}_{i,j}^-) \leq -1 + \xi_i$$

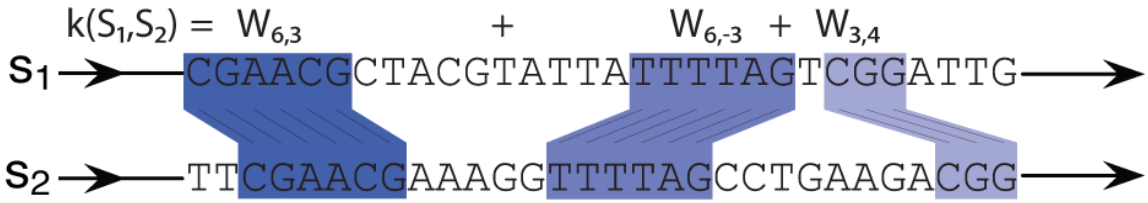
- minimize $\sum_i \xi_i + C\mathbf{P}(f)$ (Linear Program)
- Too many decoys! Use *Column Generation* technique:
  - iteratively include decoys with violated constraints
  - fast & guaranteed convergence

A novel kernel for sequences

- *Weighted Degree Kernel* (Sonnenburg & Rätsch, 2002)



  - finds motifs at specific positions
  - fails if motif positions vary
- new kernel shifts sequences against each other:



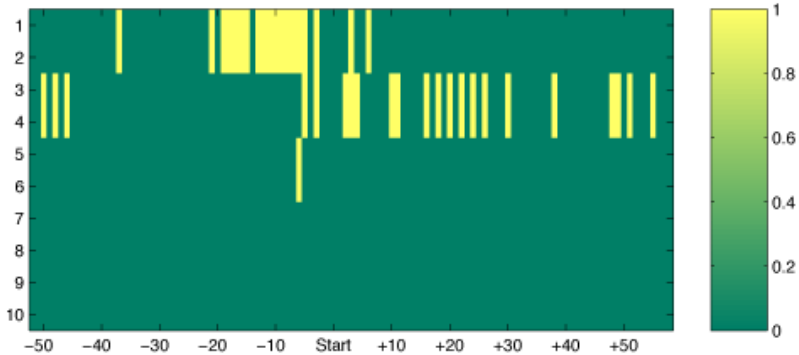
  - improved recognition of motifs at nearby positions
  - additionally: information about exon & intron lengths

Method for Interpreting SVMs

- Weighted Degree kernel: linear comb. of LD kernels

$$k(\mathbf{x}, \mathbf{x}') = \sum_{d=1}^{D} \sum_{l=1}^{L-d+1} \gamma_{l,d} \mathbf{I}(\mathbf{u}_{l,d}(\mathbf{x}) = \mathbf{u}_{l,d}(\mathbf{x}'))$$

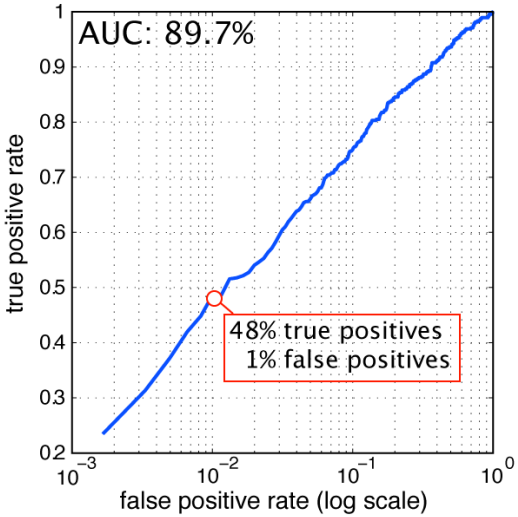
- Designed a new method for optimizing the coefficients γ
  $\Rightarrow$ Multiple Kernel Learning
  $\Rightarrow$ Wrapper algorithm based on column-generation
- For instance result for classifying splice sites:



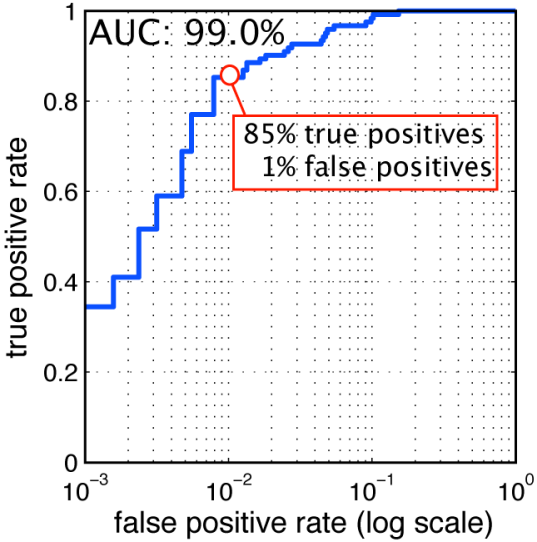
Computational Results

- 487 alternatively and 2531 constitutively spliced exons
  ...derived from EST data base (*C. elegans*)
- model selection and testing *via* 5-fold cross-validation

Task 1 (exon known)



Task 2 (intron known)



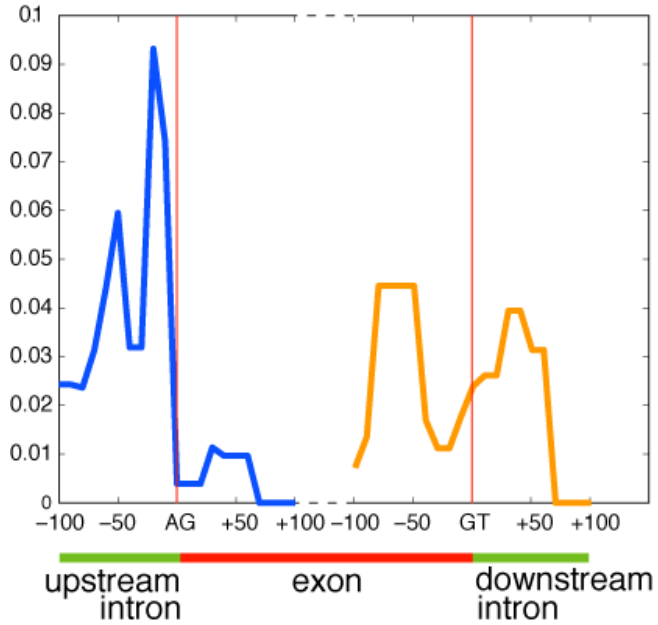
As accurate as for conserved alternatively spliced exons in human.

$\Rightarrow$ Conservation not needed!

Understanding the Classifier

- What does the algorithm use for discrimination?
- Apply a Multiple Kernel Learning algorithm

  ⇒ optimizes a combination of kernels (Sonnenburg, Rätsch & Schäfer, 2005)
- Which positions are important?

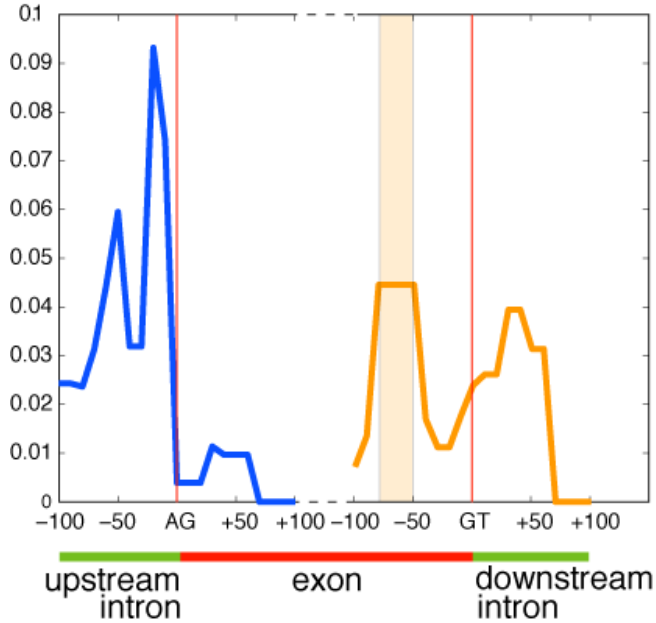


Understanding the Classifier

- What does the algorithm use for discrimination?
- Apply a Multiple Kernel Learning algorithm

  ⇒ optimizes a combination of kernels (Sonnenburg, Rätsch & Schäfer, 2005)

- Which positions are important?



- Which motifs are important?

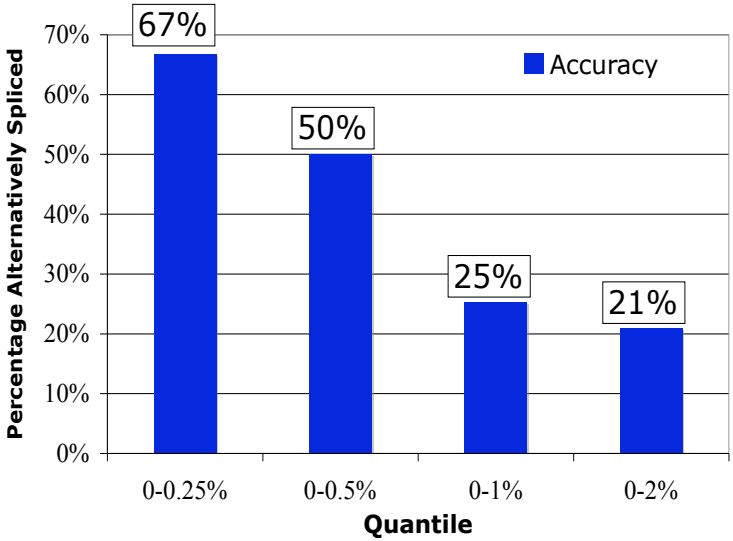
Hexamer	E-value
TTTAAA	1.8e-12
AATTTT	2.2e-10
ATTTTA	2.9e-09
CAGCAG	1.2e-08
TAATTT	8.3e-08
TTCCCC	2.1e-07
TTTTTT	5.2e-07
ATATAT	7.8e-07
ATTTAA	1.3e-06
TAAAAA	1.5e-06
GCTAGC	5.1e-06
AGGCGG	5.9e-06

Red: 5-mers found in Yeo et al., 2005, for human ($p < 1.2\%$)

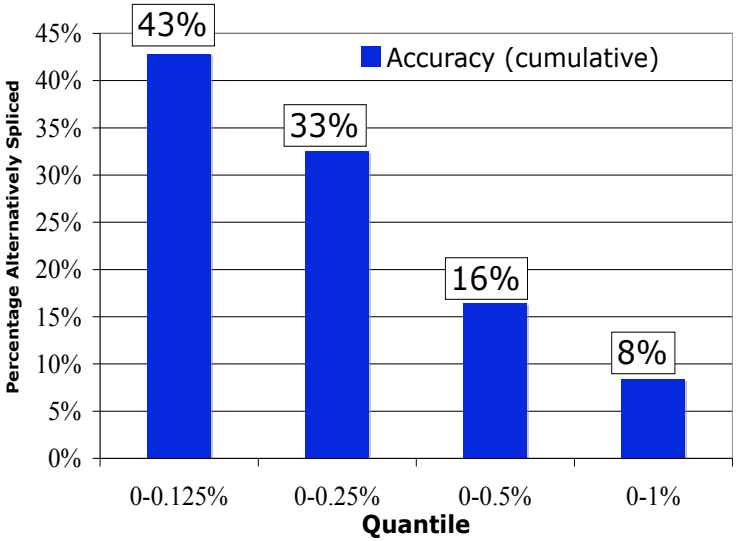
Wetlab Results

- 21,000 exons and 28,000 introns (single EST confirmed)
- 25 random exons & introns from 1-2% top ranks
- RT-PCR with primers in flanking exons 
- Gel separation & direct sequencing for verification

Task 1 (exon known)



Task 2 (intron known)



Conclusions

Based on wetlab experiments and accuracy estimates:
(in our test set)

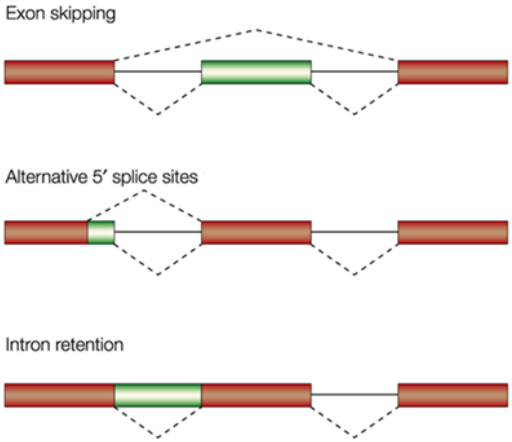
- $\approx$ 1% of known exons are alternatively spliced (AS)
- $\approx$ 0.25% of AS exons are yet completely unknown
- $\approx$ 280 AS spliced exons (total)
  - 13 confirmed by RT-PCR
  - additional $\approx$ 80 AS exons can be found with less than 200 additional RT-PCRs

Genome-wide: around 4x more (some are known already)

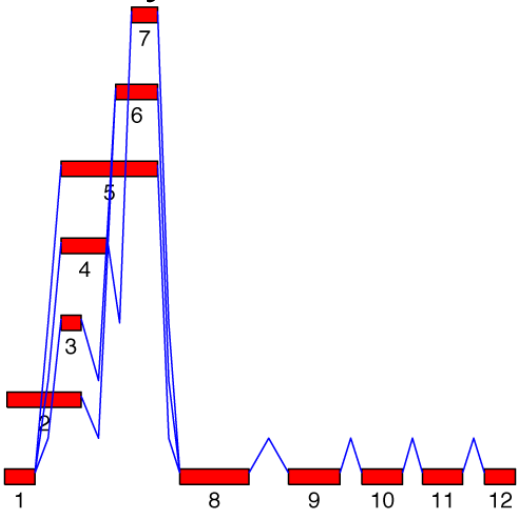
Predictions for *C. elegans* are available at
http://www.fml.tuebingen.mpg.de/raetsch/projects/RASE

Empirical Inference Challenges

Simple classes:



Reality:



- Predicting the simple cases is not enough

$\Rightarrow$ need to predict the gene structure

- Difficult learning setting:
  - Input: DNA sequence
  - Output: Splicegraph (vertices & edges unknown)

Summary

- Solved two tasks using SVM-like algorithms:
  - Classification of known exons (AS vs. CS)
  - Finding yet unknown AS exons
- Accurate predictions of AS exons is possible . . .
  - even without assuming conservation
- Wetlab experiments support computational results
- A few more experiments will reveal many more AS exons
- Future work
  - human and other organisms
  - other alternative splicing variants . . .
  - . . . in combination with *ab initio* gene-finding

Thanks for your attention!

MAX-PLANCK-GESELLSCHAFT

Details, Data sets & Predictions:

- http://www.fml.tuebingen.mpg.de/raetsch/projects/RASE

Acknowledgments:

- Uwe Ohler, Gene Yeo & Klaus R. Müller for discussions
- Sommer Lab for providing *C. elegans* mRNA

Postdoc & PhD student positions/scholarships available

- Please contact:

Gunnar Rätsch (Gunnar.Raetsch@tuebingen.mpg.de)
Friedrich Miescher Laboratory, Tübingen, Germany

http://www.fml.tuebingen.mpg.de/raetsch/jobs

Links

- Homepage
  http://www.fml.tuebingen.mpg.de/~raetsch
- Details, Data sets & Predictions for alt. splicing
  http://www.fml.tuebingen.mpg.de/raetsch/projects/RASE
- Work on Splice site detection:
  http://www.fml.tuebingen.mpg.de/raetsch/projects/AnuSplice
- Work on Large Scale Multiple Kernel Learning
  http://www.fml.tuebingen.mpg.de/raetsch/projects/mkl_splice
- Workshop on “New Problems and Methods in Computational Biology”
  http://www.fml.tuebingen.mpg.de/nipscompbio