Discriminative Learning Methods for Gene Finding and Polymorphism Identification

Gunnar Rätsch

Friedrich Miescher Laboratory
Max Planck Society
Tübingen, Germany

University of Liège, Belgium

May 21, 2008

Slides available at `http://www.fml.mpg.de/raetsch/lectures`

Research Topics

Group established in 2005 focusing on:

1. Machine learning methods
   $\Rightarrow$ *Develop fast, accurate and interpretable learning methods*
2. Genome annotation
   $\Rightarrow$ *Predict features encoded on DNA*
3. Analysis of polymorphisms
   $\Rightarrow$ *Discover polymorphisms and associate them with phenotypes*
4. Biological networks
   $\Rightarrow$ *Understand interactions between gene products*

Research Topics

Group established in 2005 focusing on:

1. Machine learning methods
   $\Rightarrow$ *Develop fast, accurate and interpretable learning methods*
2. Genome annotation
   $\Rightarrow$ *Predict features encoded on DNA*
3. Analysis of polymorphisms
   $\Rightarrow$ *Discover polymorphisms and associate them with phenotypes*
4. Biological networks
   $\Rightarrow$ *Understand interactions between gene products*

Research Topics

Group established in 2005 focusing on:

1. Machine learning methods
   $\Rightarrow$ *Develop fast, accurate and interpretable learning methods*
2. Genome annotation
   $\Rightarrow$ *Predict features encoded on DNA*
3. Analysis of polymorphisms
   $\Rightarrow$ *Discover polymorphisms and associate them with phenotypes*
4. Biological networks
   $\Rightarrow$ *Understand interactions between gene products*

Research Topics

Group established in 2005 focusing on:

1. Machine learning methods
   $\Rightarrow$ *Develop fast, accurate and interpretable learning methods*
2. Genome annotation
   $\Rightarrow$ *Predict features encoded on DNA*
3. Analysis of polymorphisms
   $\Rightarrow$ *Discover polymorphisms and associate them with phenotypes*
4. Biological networks
   $\Rightarrow$ *Understand interactions between gene products*

Part I

mGene: A Novel Discriminative Gene Finding System

$\frac{1}{150\,000}$ of the DNA of *C. elegans*

Friedrich Miescher Laboratory
of the Max Planck Society

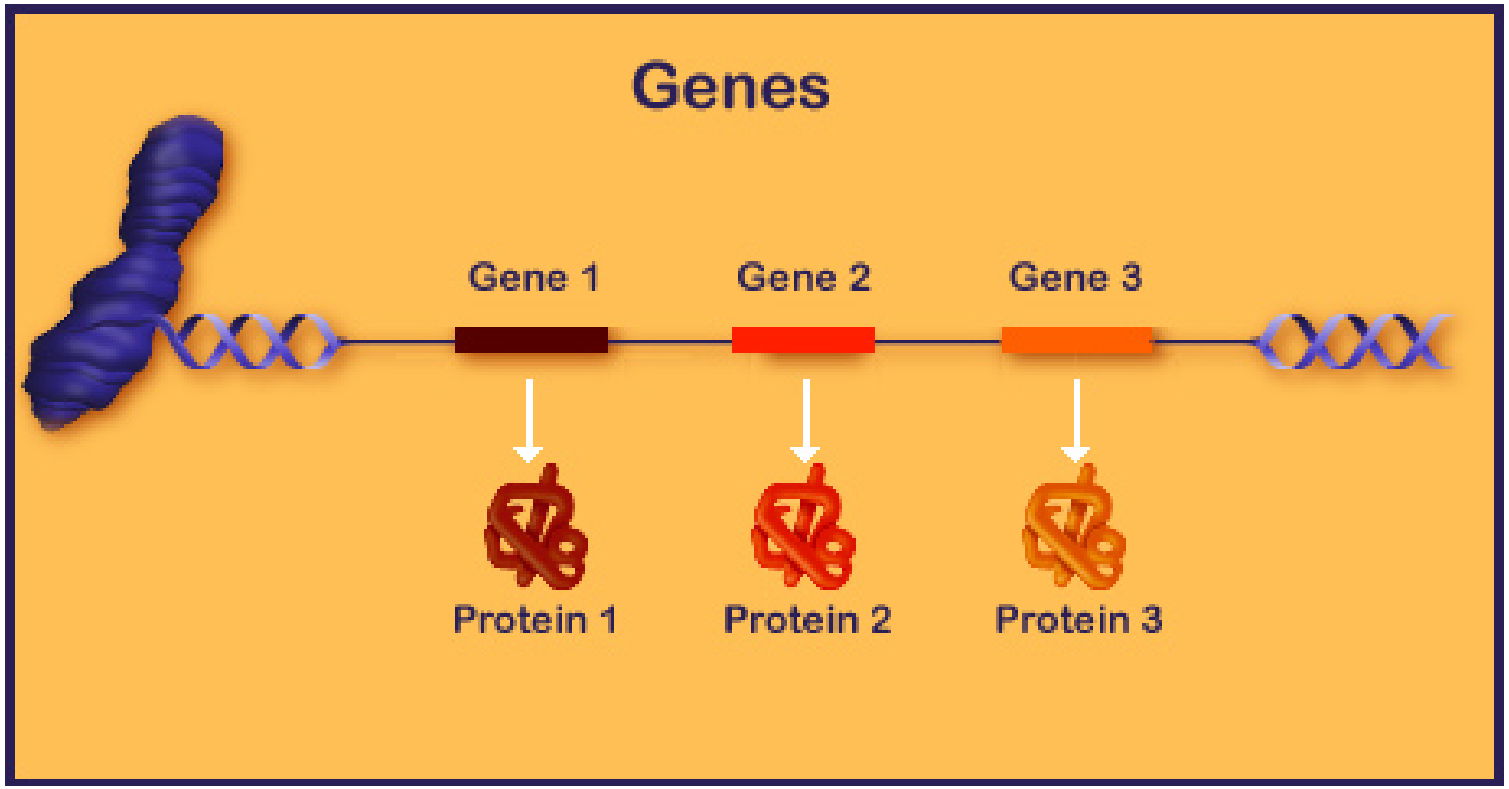
>CHROMOSOME_I
GCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGC
CTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCT
AAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAA
GCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGC
CTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCT
AAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAA
GCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGC
CTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCT
AAGCCTAAGCCTAAGCCTAAGCCTAAAAAATTGAGATAAGAAAA
CATTTTACTTTTTCAAAATTGTTTTCATGCTAAATTCAAAACGTTTTTTT
TTTAGTGAAGCTTCTAGATATTTGGCGGGTACCTCTAATTTTGCCTGCCT
GCCAACCTATATGCTCCTGTGTTTAGGCCTAATACTAAGCCTAAGCCTAA
GCCTAATACTAAGCCTAAGCCTAAGACTAAGCCTAATACTAAGCCTAAGC
CTAAGACTAAGCCTAAGACTAAGCCTAAGACTAAGCCTAATACTAAGCCT
...







Ab initio Gene Prediction



- Find stretches in the DNA coding for proteins
- 70-99% of the DNA sequence is non-coding

Genome Annotation

- Identification of Genes
  - ⇒ Basis for subsequent biological analyses
- Detailed description of genes
  - ⇒ Function
  - ⇒ Cellular localization
  - ⇒ Interactions

First steps:

- Alignment of known genes
  - ⇒ e.g. known mRNA or peptide sequences . . .
    . . . possibly from related organisms
- Prediction of new genes
  - ⇒ Recognition of various signals on DNA
  - ⇒ Prediction of mRNA & gene products

Genome Annotation

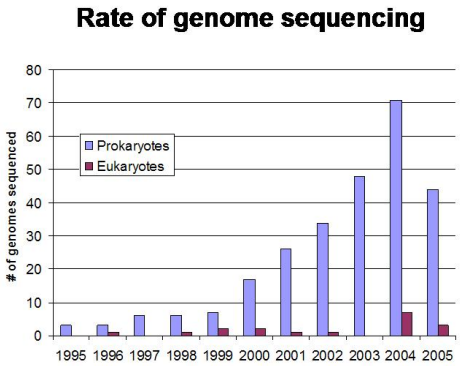
- Identification of Genes
  - ⇒ Basis for subsequent biological analyses
- Detailed description of genes
  - ⇒ Function
  - ⇒ Cellular localization
  - ⇒ Interactions

First steps:

- Alignment of known genes
  - ⇒ e.g. known mRNA or peptide sequences . . .
    . . . possibly from related organisms
- Prediction of new genes
  - ⇒ Recognition of various signals on DNA
  - ⇒ Prediction of mRNA & gene products

Genome Annotation Challenge

- Since 1995 already 180 genomes sequenced
- Many computational tools are available for gene prediction (usually HMM based)
- Yet, gene predictions are very inaccurate!



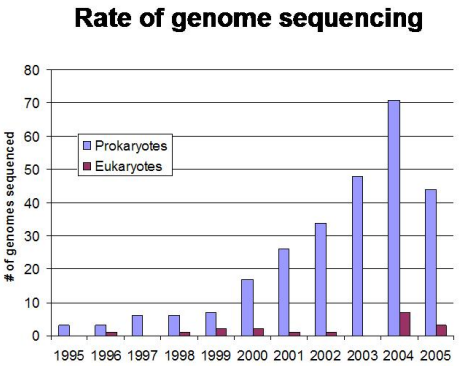
Example: Nematodes:

- *C. elegans* sequenced in 1998, *C. briggsae* in 2002
- Several more nematode genomes have recently been sequenced (*C. remanei*, *C. japonica*, *P. pacificus*, . . .)
- Genomes need to be annotated! Which gene finder?
- nGASP competition:
  - Controlled competition conditions
  - Identify the most accurate method

Our contribution: A novel accurate gene finding system

Genome Annotation Challenge

- Since 1995 already 180 genomes sequenced
- Many computational tools are available for gene prediction (usually HMM based)
- Yet, gene predictions are very inaccurate!



Example: Nematodes:

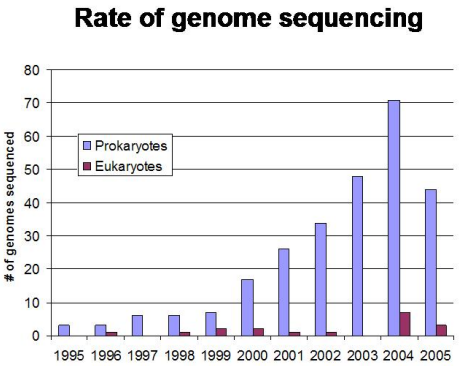


- *C. elegans* sequenced in 1998, *C. briggsae* in 2002
- Several more nematode genomes have recently been sequenced (*C. remanei*, *C. japonica*, *P. pacificus*, ...)
- Genomes need to be annotated! Which gene finder?
- nGASP competition:
  - Controlled competition conditions
  - Identify the most accurate method

Our contribution: A novel accurate gene finding system

Genome Annotation Challenge

- Since 1995 already 180 genomes sequenced
- Many computational tools are available for gene prediction (usually HMM based)
- Yet, gene predictions are very inaccurate!



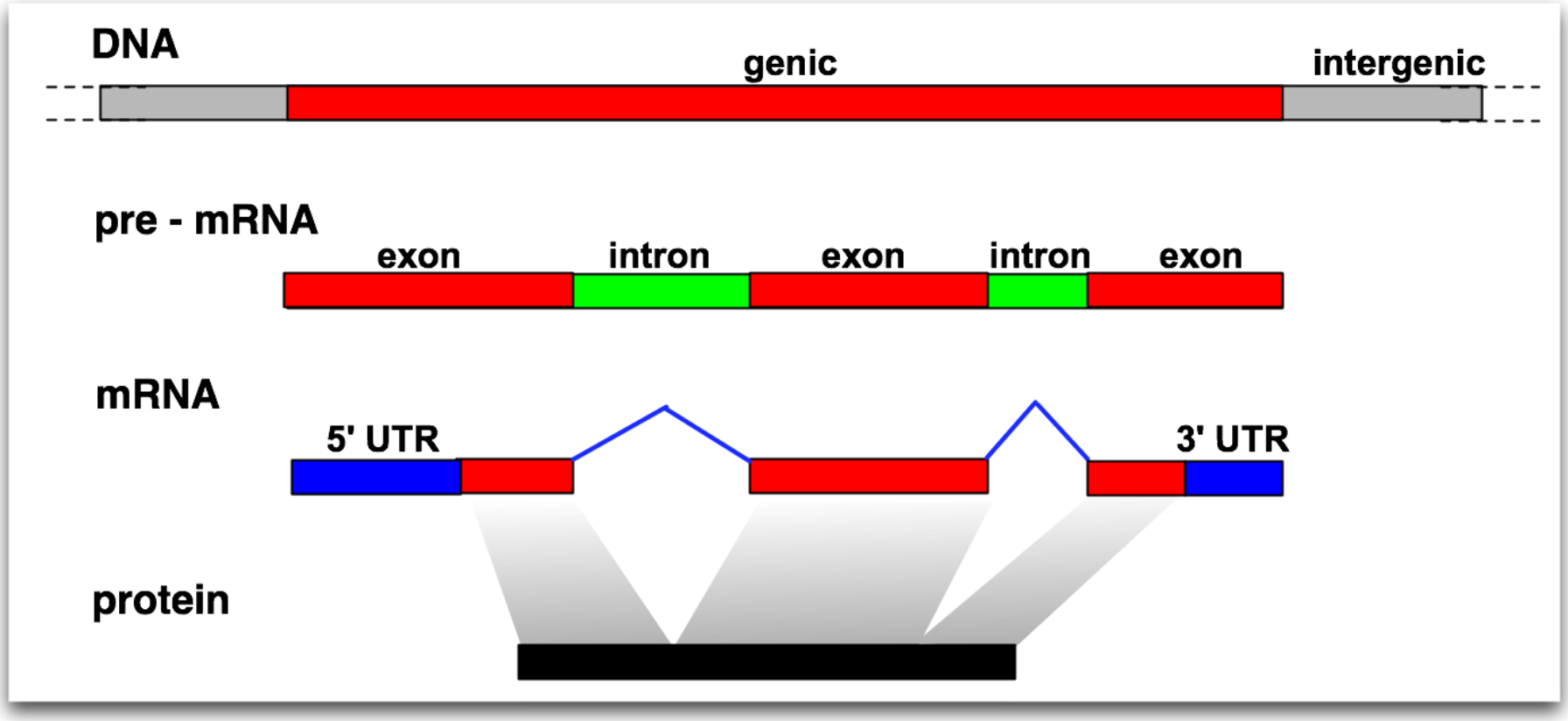
Example: Nematodes:



- *C. elegans* sequenced in 1998, *C. briggsae* in 2002
- Several more nematode genomes have recently been sequenced (*C. remanei*, *C. japonica*, *P. pacificus*, ...)
- Genomes need to be annotated! Which gene finder?
- nGASP competition:
  - Controlled competition conditions
  - Identify the most accurate method

Our contribution: A novel accurate gene finding system

A Label Sequence Learning Problem



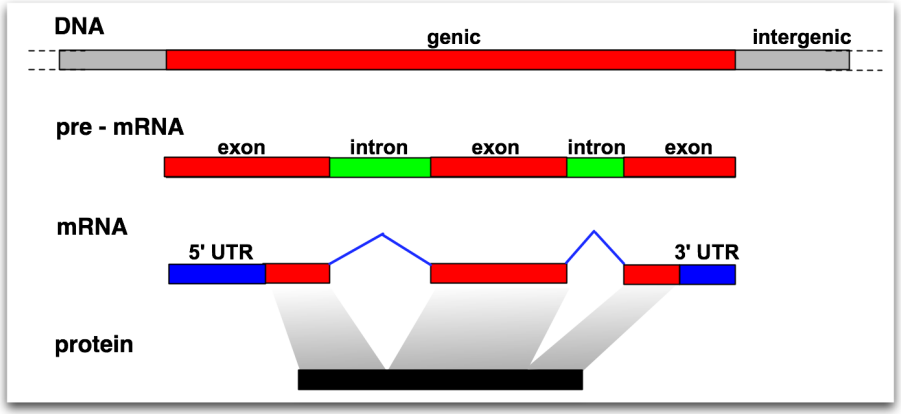
Problem is slightly more difficult: genes have structure

- Given a DNA sequence $\mathbf{x} \in \{\text{'A','C','G','T'}\}^L$
- Find the correct **label sequence y** $= y_1 y_2 \ldots y_L$
  ($y_i \in Y = \{$'intergenic', 'exon', 'intron', $\ldots\}$)

Hidden Markov Models

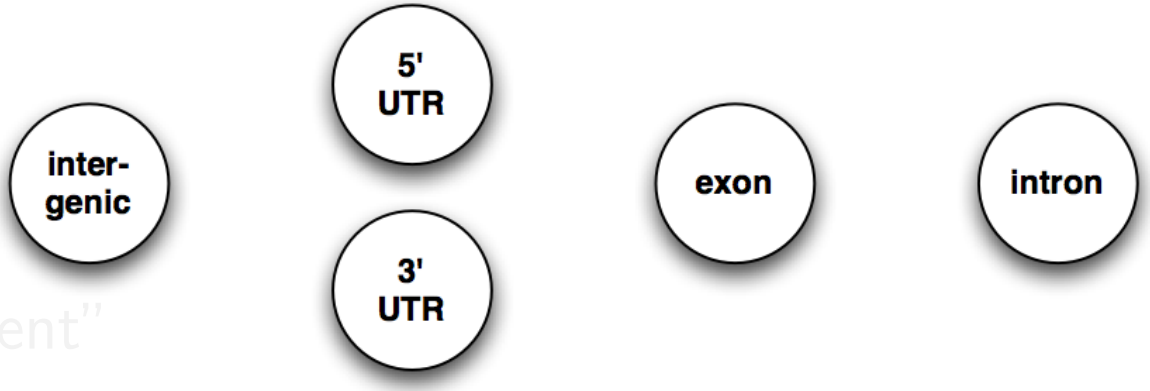
Model sequence content:

- One state per segment type
- Allow only plausible transitions
- Content statistics at each state
  - Derived from known genes



Prediction:

- Given DNA, find most likely state sequences

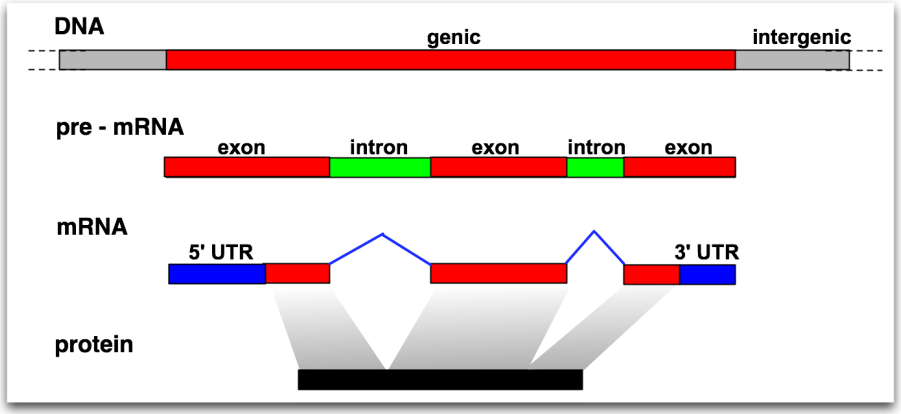


- Focuses on "content"
- "Signals" often poorly modeled

Hidden Markov Models

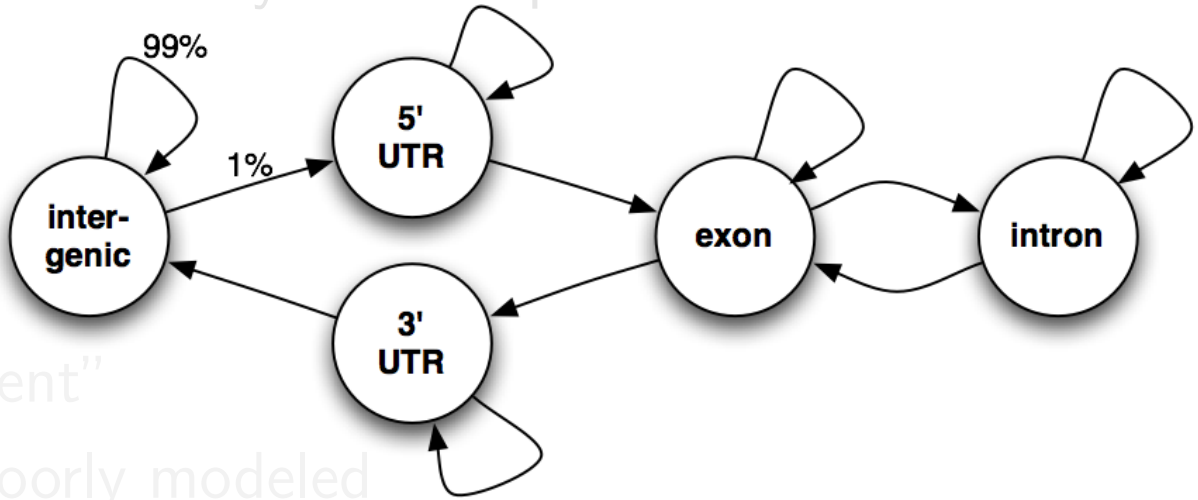
Model sequence content:

- One state per segment type
- Allow only plausible transitions
- Content statistics at each state
  - Derived from known genes



Prediction:

- Given DNA, find most likely state sequences

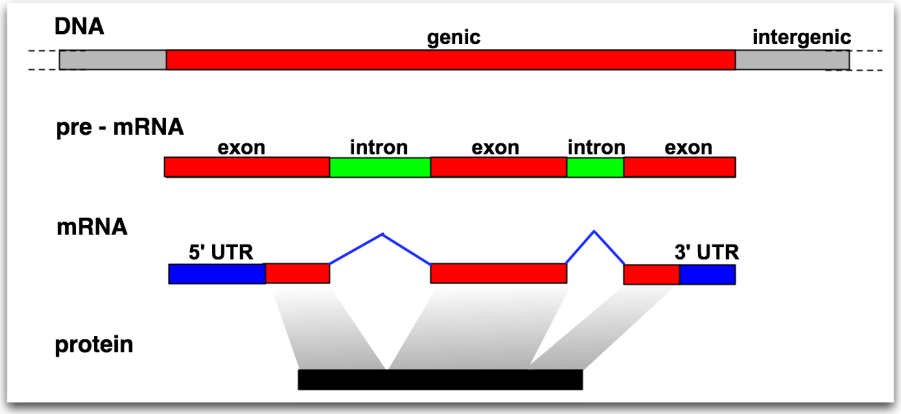


- Focuses on "content"
- "Signals" often poorly modeled

Hidden Markov Models

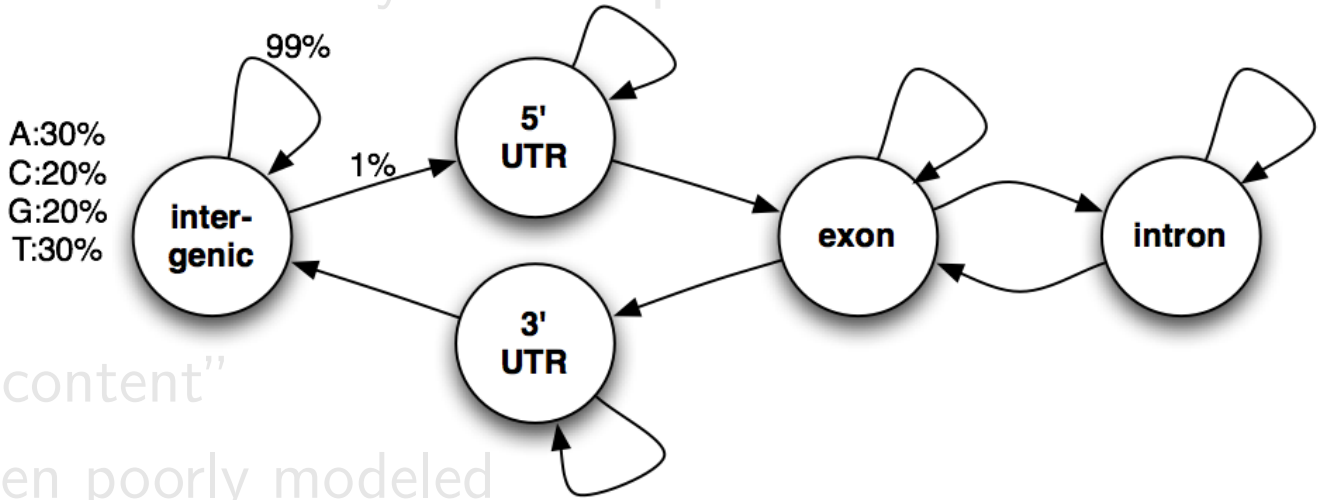
Model sequence content:

- One state per segment type
- Allow only plausible transitions
- Content statistics at each state
  - Derived from known genes



Prediction:

- Given DNA, find most likely state sequences

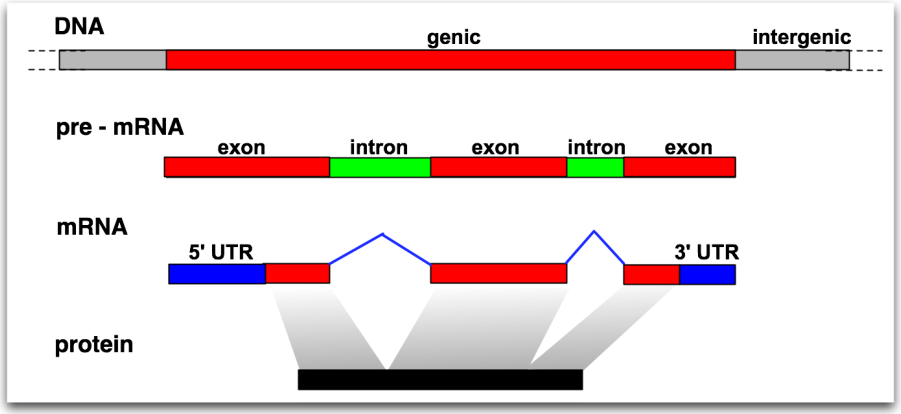


- Focuses on "content"
- "Signals" often poorly modeled

Hidden Markov Models

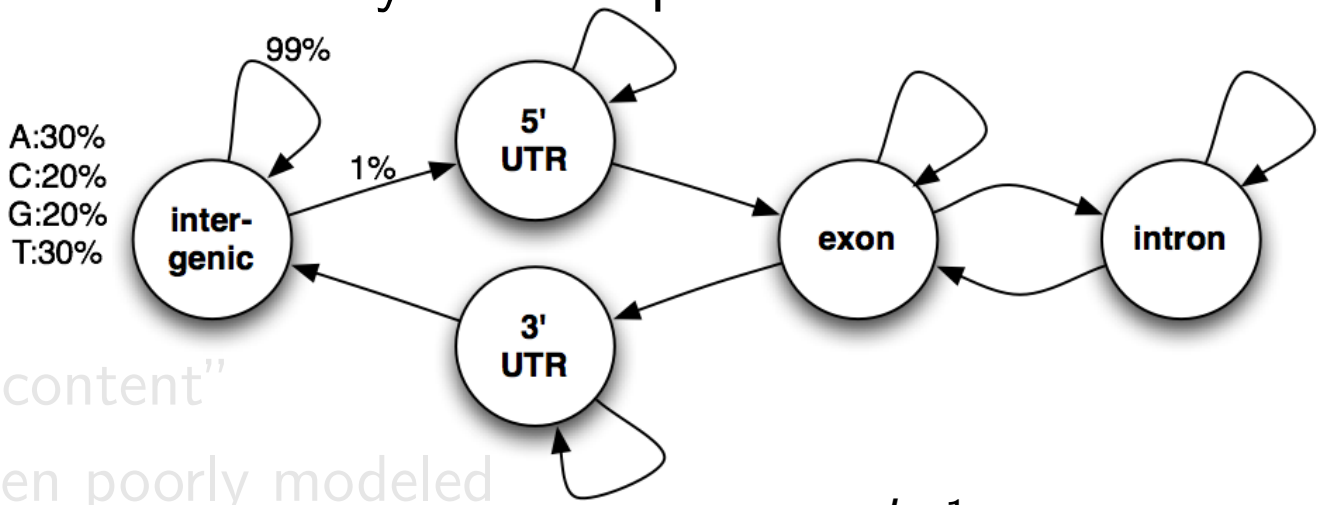
Model sequence content:

- One state per segment type
- Allow only plausible transitions
- Content statistics at each state
  - Derived from known genes



Prediction:

- Given DNA, find most likely state sequences



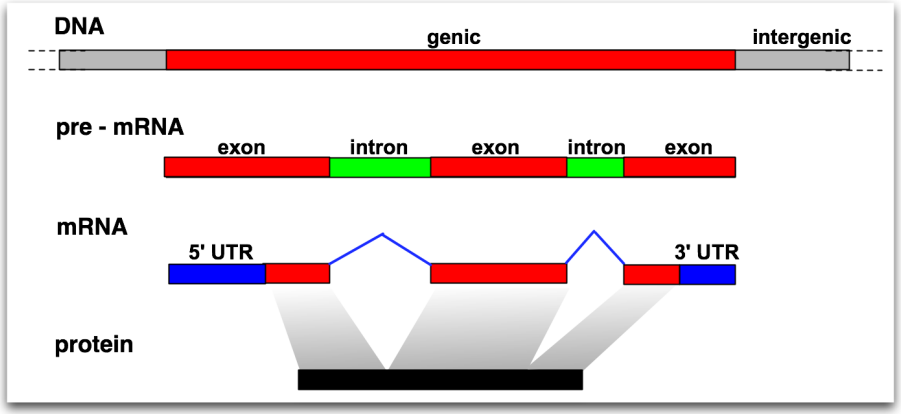
- Focuses on "content"
- "Signals" often poorly modeled

$$p(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{L-1} p(x_i|y_i)p(y_{i+1}|y_i)$$

Hidden Markov Models

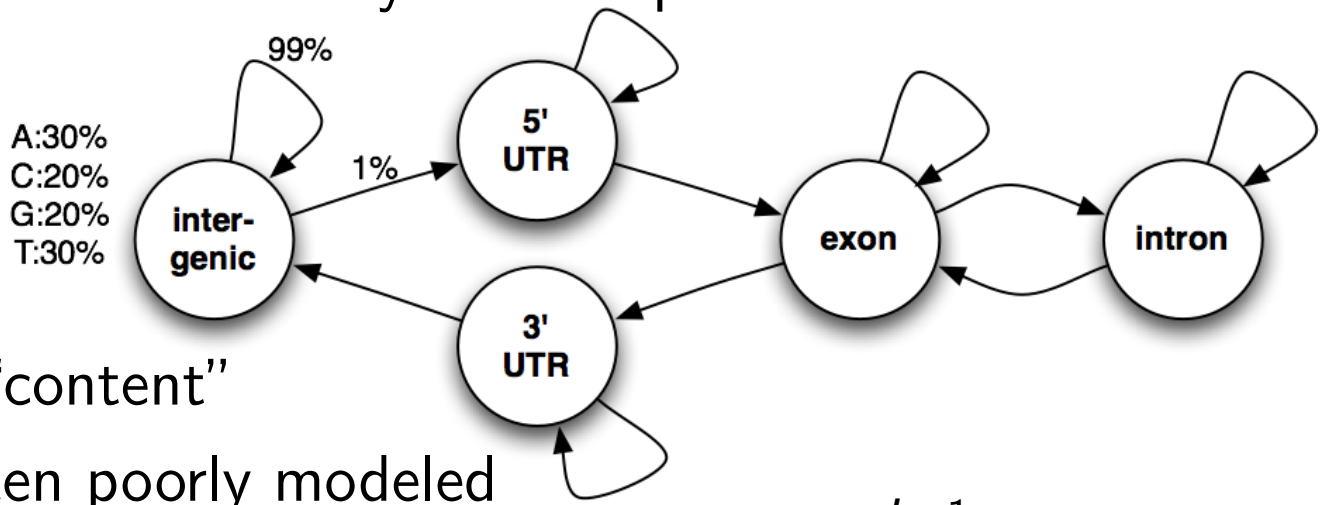
Model sequence content:

- One state per segment type
- Allow only plausible transitions
- Content statistics at each state
  - Derived from known genes



Prediction:

- Given DNA, find most likely state sequences

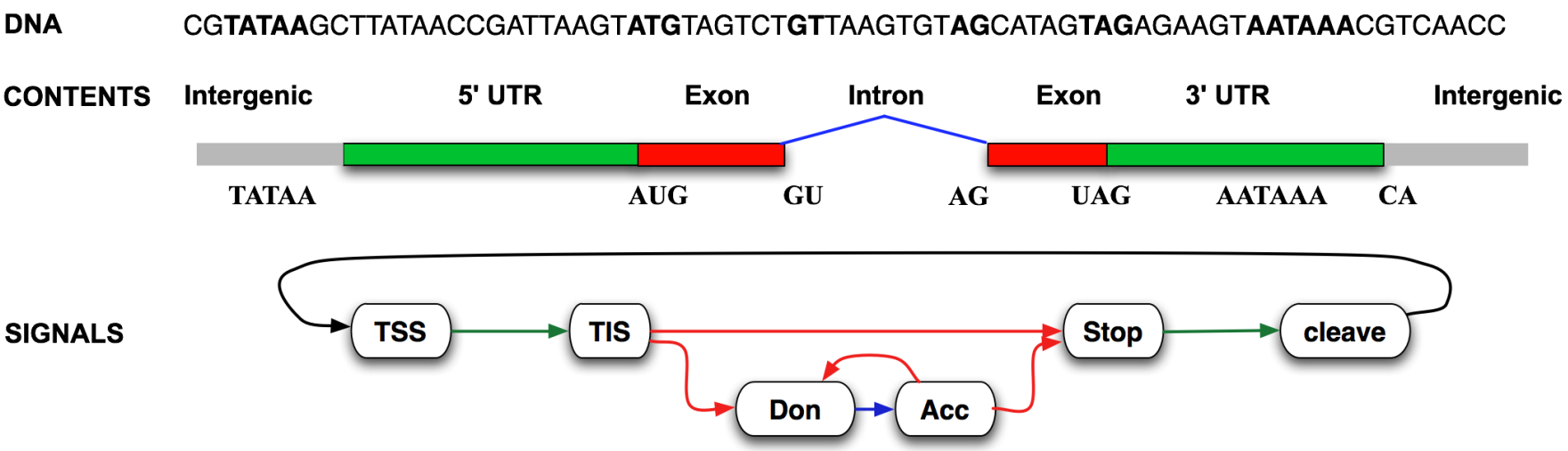


- Focuses on "content"
- "Signals" often poorly modeled

$$p(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{L-1} p(x_i|y_i)p(y_{i+1}|y_i)$$

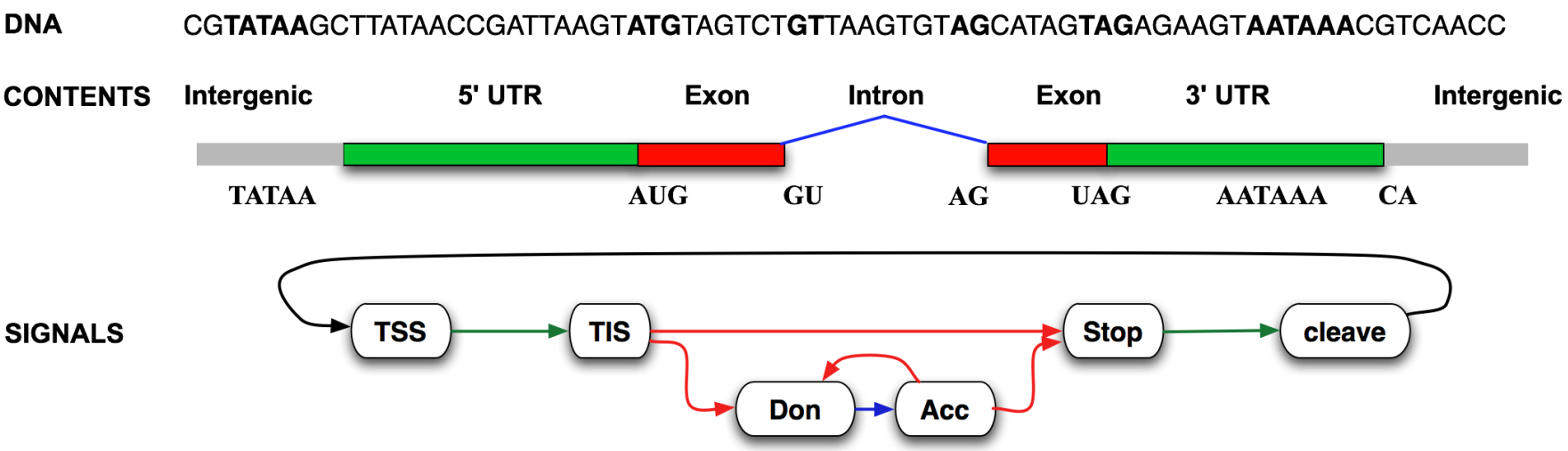
Our Approach: Modeling Signals & Content

- States correspond to sequence signals
  - Depends on recognition of signals on the DNA
- Transitions correspond to segments
  - Model length and content of segment
- Markovian on segment level, non-Markovian within segments
  - Allows efficient decoding and modeling of segment lengths



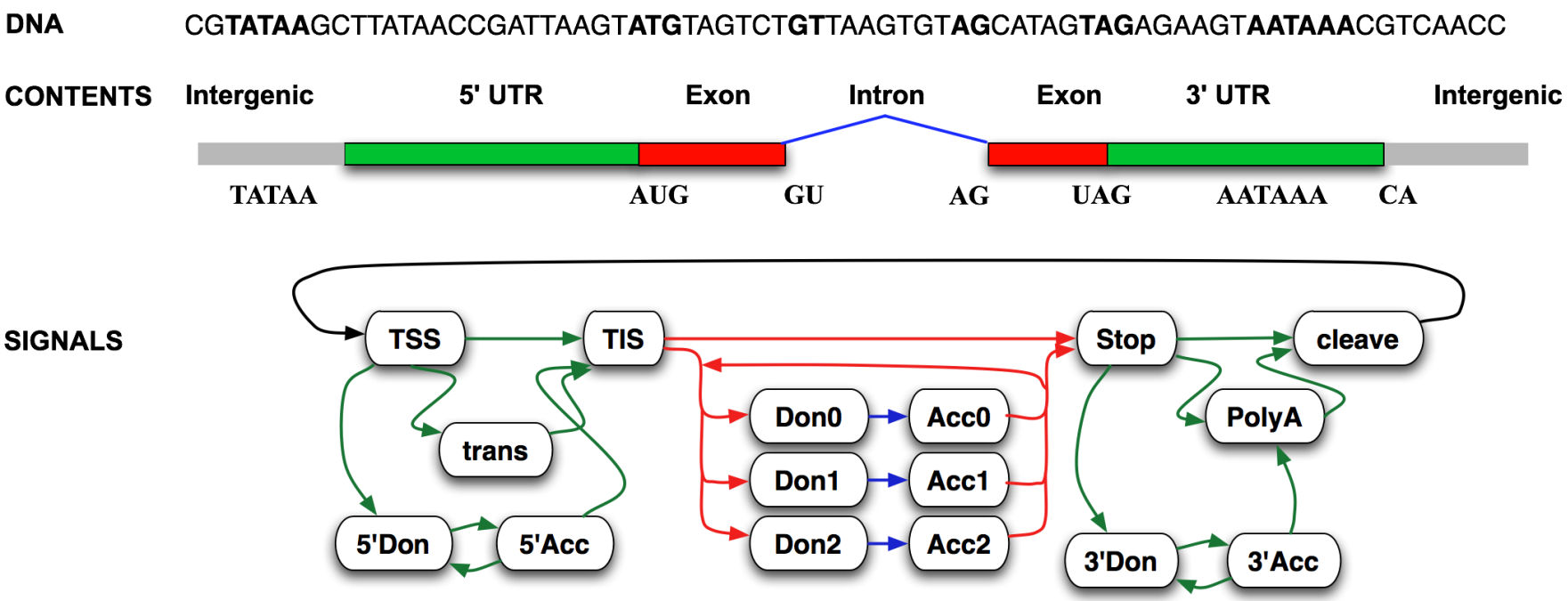
Our Approach: Modeling Signals & Content

- States correspond to sequence signals
  - Depends on recognition of signals on the DNA
- Transitions correspond to segments
  - Model length and content of segment
- Markovian on segment level, non-Markovian within segments
  - Allows efficient decoding and modeling of segment lengths



Our Approach: Modeling Signals & Content

- States correspond to sequence signals
  - Depends on recognition of signals on the DNA
- Transitions correspond to segments
  - Model length and content of segment
- Markovian on segment level, non-Markovian within segments
  - Allows efficient decoding and modeling of segment lengths



Recognition of Signals and Content

Sensors to recognize signals:

- Transcription start and cleavage site, polyA site
- Translation initiation site and stop codon
- Donor and acceptor splice sites

Discriminate true signal positions against all other positions

Sensors to recognize contents:

- Exons
- Introns
- Intergenic

Distinguish one content type from all others

Typical approach: PSSMs or higher order Markov chains

Here: Support Vector Machines

Recognition of Signals and Content

Sensors to recognize signals:

- Transcription start and cleavage site, polyA site
- Translation initiation site and stop codon
- Donor and acceptor splice sites

Discriminate true signal positions against all other positions

Sensors to recognize contents:

- Exons
- Introns
- Intergenic

Distinguish one content type from all others

Typical approach: PSSMs or higher order Markov chains

Here: Support Vector Machines

Recognition of Signals and Content

Sensors to recognize signals:

- Transcription start and cleavage site, polyA site
- Translation initiation site and stop codon
- Donor and acceptor splice sites

Discriminate true signal positions against all other positions

Sensors to recognize contents:

- Exons
- Introns
- Intergenic

Distinguish one content type from all others

Typical approach: PSSMs or higher order Markov chains

Here: Support Vector Machines

Recognition of Signals and Content

Sensors to recognize signals:

- Transcription start and cleavage site, polyA site
- Translation initiation site and stop codon
- Donor and acceptor splice sites

Discriminate true signal positions against all other positions

Sensors to recognize contents:

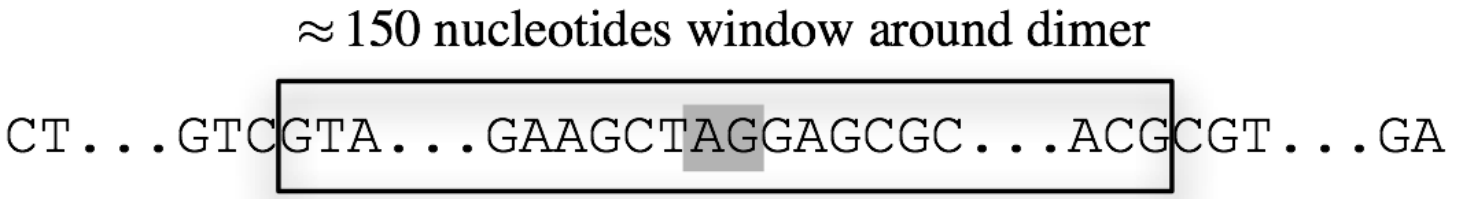
- Exons
- Introns
- Intergenic

Distinguish one content type from all others

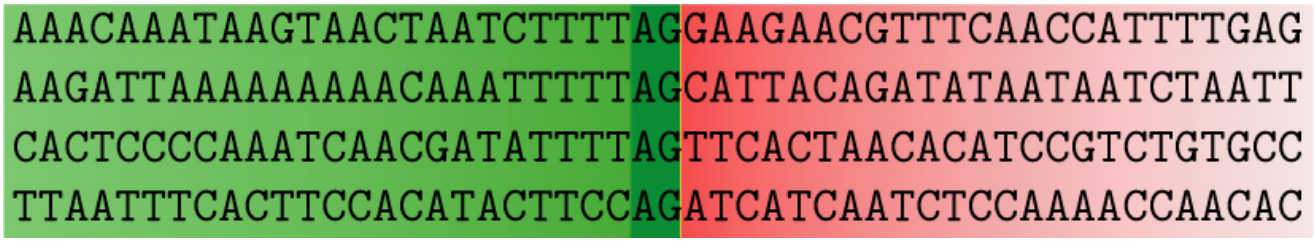
Typical approach: PSSMs or higher order Markov chains

Here: Support Vector Machines

Example: SVMs for Splice Site Recognition



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



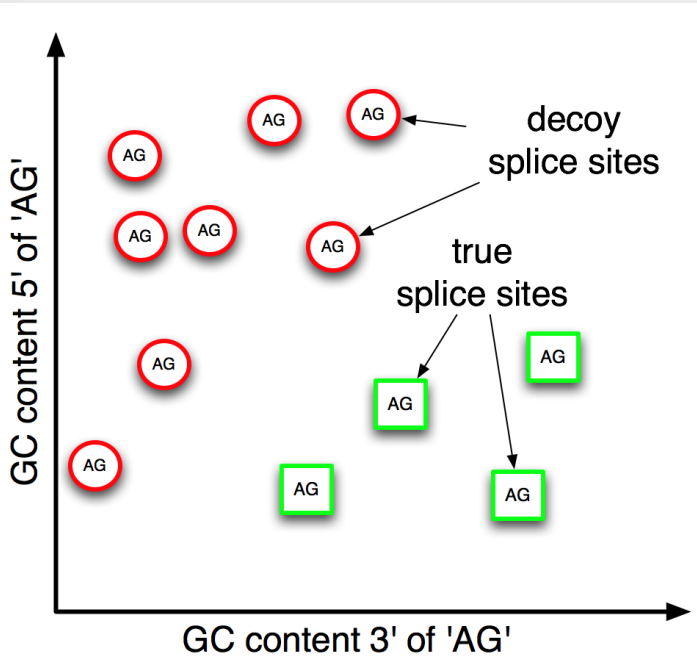
⇒ Very unbalanced problem (1 positive against ≈ 200 negatives)
⇒ Millions of points from EST databases
⇒ Large scale methods necessary

Example: SVMs for Splice Site Recognition

≈ 150 nucleotides window around dimer

CT...GTCGTA...GAAGCTAGGAGCGC...ACGCGT...GA

Basic idea:



For instance, exploit that exons have higher GC content

or

that certain motifs appear near splice sites.

In practice: Use one feature per possible substring (e.g. ≤ 20) at all positions

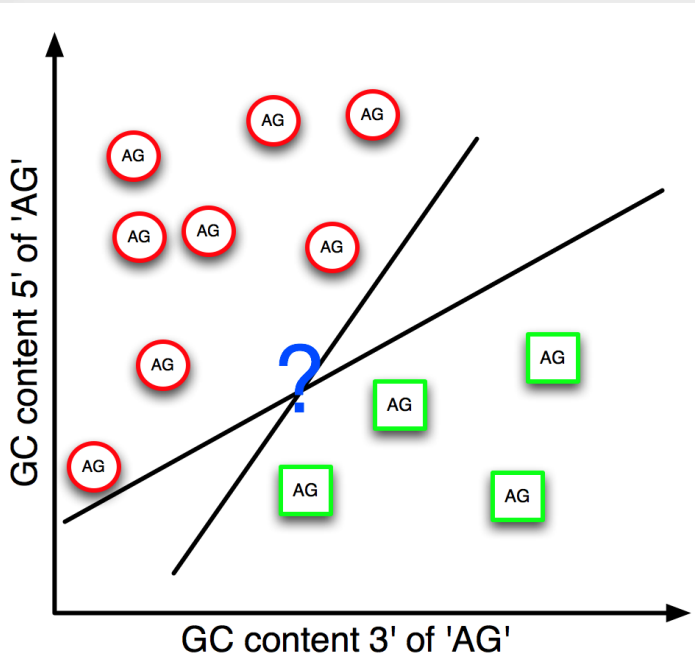
$$150 \cdot (4^1 + 4^2 + \ldots + 4^{20}) \approx 2 \cdot 10^{14} \text{ features}$$

Example: SVMs for Splice Site Recognition

≈ 150 nucleotides window around dimer

CT...GTCGTA...GAAGCTAGGAGCGC...ACGCGT...GA

Basic idea:



For instance, exploit that exons have higher GC content

or

that certain motifs appear near splice sites.

In practice: Use one feature per possible substring (e.g. ≤ 20) at all positions

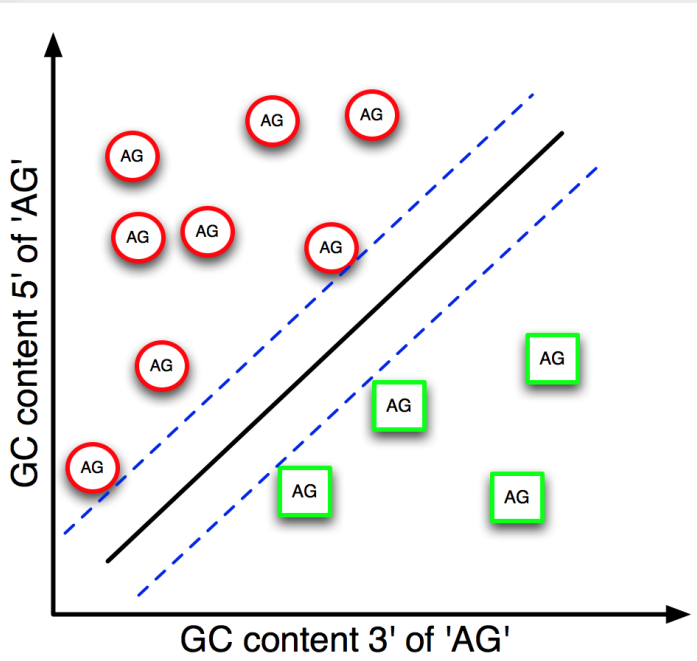
$$150 \cdot (4^1 + 4^2 + \ldots + 4^{20}) \approx 2 \cdot 10^{14} \text{ features}$$

Example: SVMs for Splice Site Recognition

≈ 150 nucleotides window around dimer

CT...GTCGTA...GAAGCTAGGAGCGC...ACGCGT...GA

Basic idea:



For instance, exploit that exons have higher GC content

or

that certain motifs appear near splice sites.

In practice: Use one feature per possible substring (e.g. ≤ 20) at all positions

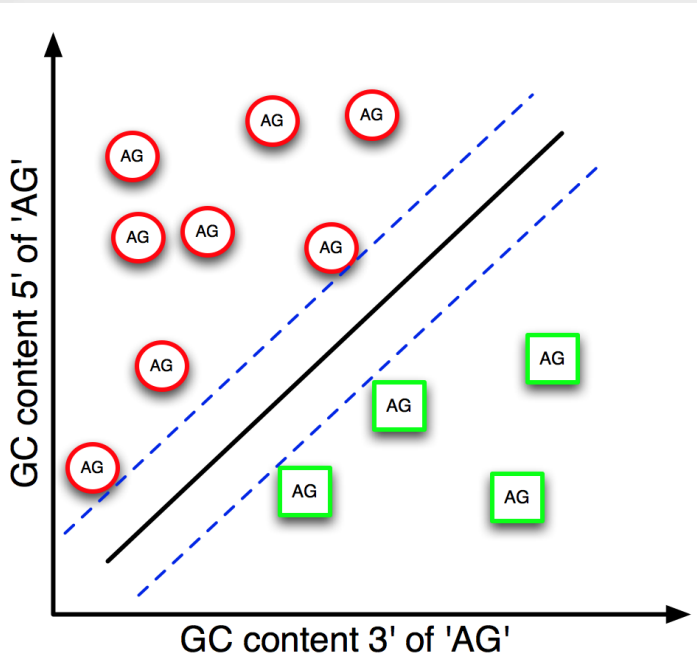
$$150 \cdot (4^1 + 4^2 + \ldots + 4^{20}) \approx 2 \cdot 10^{14} \text{ features}$$

Example: SVMs for Splice Site Recognition

≈ 150 nucleotides window around dimer

CT...GTCGTA...GAAGCTAGGAGCGC...ACGCGT...GA

Basic idea:



For instance, exploit that exons have higher GC content

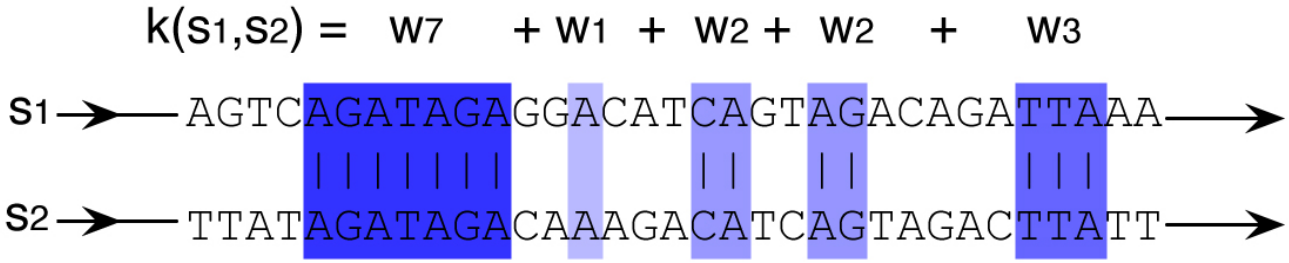
or

that certain motifs appear near splice sites.

In practice: Use one feature per possible substring (e.g. ≤ 20) at all positions

$$150 \cdot (4^1 + 4^2 + \ldots + 4^{20}) \approx 2 \cdot 10^{14} \text{ features}$$

Good news: Can use **kernel** to implicitly consider features:



[Rätsch, Sonnenburg, Srinivasan, Witte, Müller, Sommer, Schölkopf, 2007]

Bad news: Need to compute kernel between all pairs of sequences

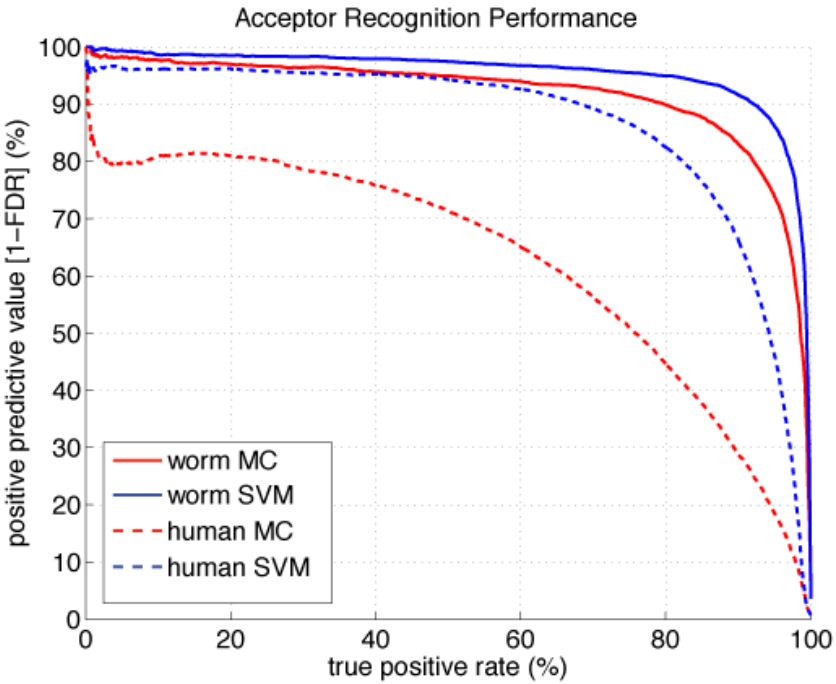
- $1,000,000^2$ kernel elements :-(
- Need more sophisticated algorithms [Schölkopf & Smola, 2006]
  - Iteratively consider smaller subproblems
    - SMO type optimization
  - Exploit that features are sparse
    - String indexing data structures

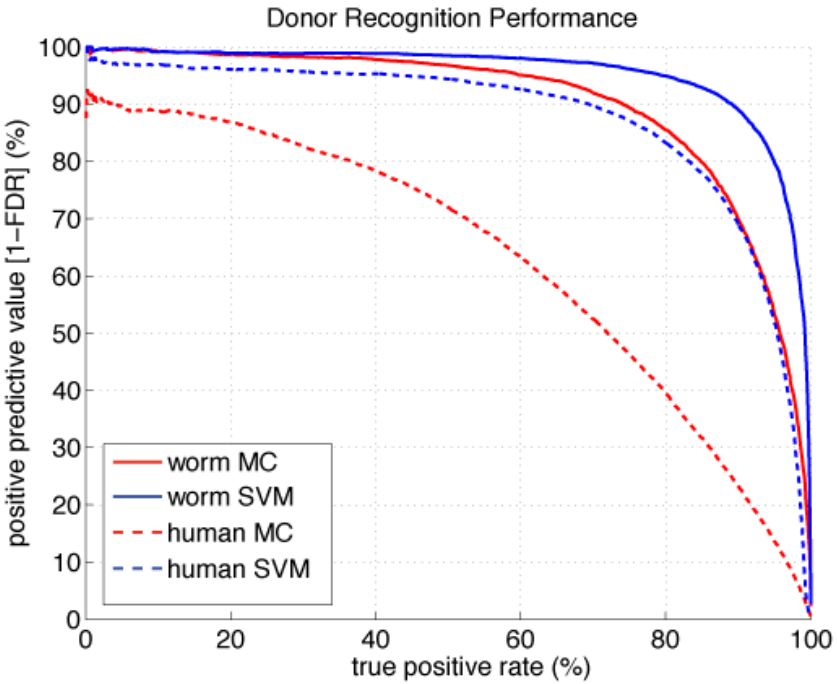
Implemented in a freely available software package

`http://www.shogun-toolbox.org` [Sonnenburg, Rieck, Rätsch, 2007]

Example 1: Splice Site Recognition

	Worm		Fly		Cress		Fish		Human	
	Acc	**Don**	**Acc**	**Don**	**Acc**	**Don**	**Acc**	**Don**	**Acc**	**Don**
Markov Chain auPRC(%)	92.1	90.0	80.3	78.5	87.4	88.2	63.6	62.9	16.2	26.0
SVM auPRC(%)	95.9	95.3	86.7	87.5	92.2	92.9	86.6	86.9	54.4	56.9



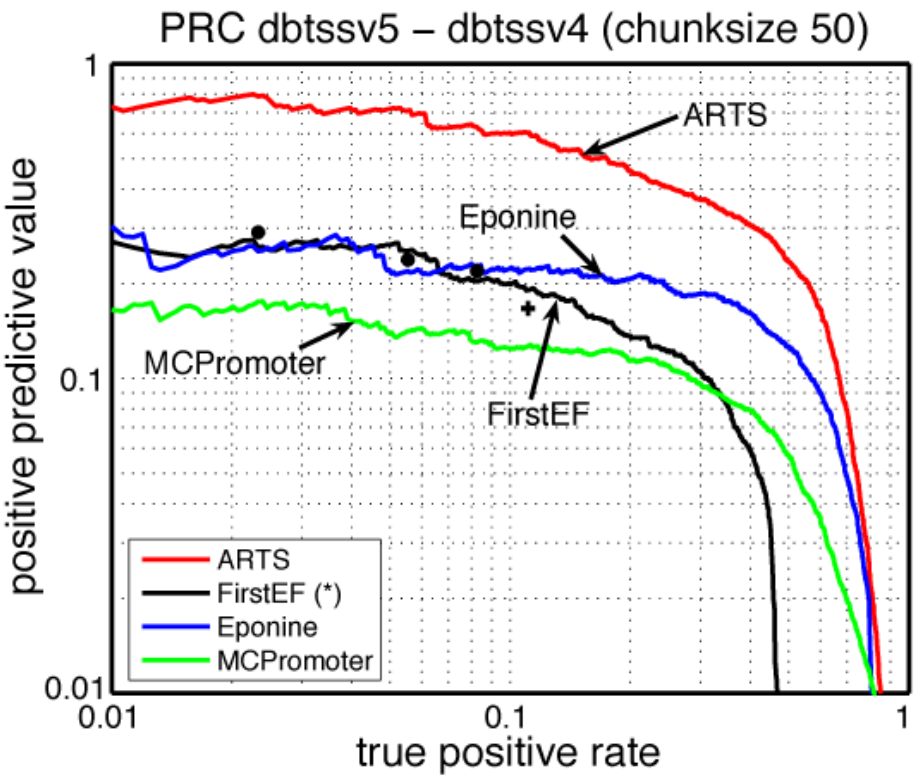


[Sonnenburg, Schweikert, Philips, Behr, Rätsch, 2007]

Example 2: Transcription Start Prediction

- Important to detect gene boundaries
- Depends on many factors
- Weakly defined signal
- Use combination of features
  - TSS Signal
  - Physical properties of DNA

Performance for human:



$\Rightarrow$ 35% **true positives at a false positive rate of** $1/1000$
(best other method finds about half (18%**))**

[Sonnenburg, Zien, Rätsch, 2006]

SVMs vs. Markov Chains

Pro SVMs:

- Much more accurate than Markov chains (and many other methods)
- Continue to improve with more data

[Sonnenburg, Zien, Rätsch, 2006, Sonnenburg et al., 2007]

Contra SVMs:

- (Originally) not large scale
  - New developments: can deal with 10 million examples

[Sonnenburg, Rieck, Rätsch, 2007]

- Difficult to interpret
  - New visualization technique: POIMs

[Sonnenburg, Zien, Philips, Rätsch, ISMB, 2008]

- No probabilistic outputs
  - Prerequisite for use in generative models (e.g. HMM)
  - Need new approach for combining predictions

[Sonnenburg, Rätsch, Jagota, Müller, 2009]

SVMs vs. Markov Chains

Pro SVMs:

- Much more accurate than Markov chains (and many other methods)
- Continue to improve with more data

[Sonnenburg, Zien, Rätsch, 2006, Sonnenburg et al., 2007]

Contra SVMs:

- (Originally) not large scale
  - New developments: can deal with 10 million examples

[Sonnenburg, Rieck, Rätsch, 2007]

- Difficult to interpret
  - New visualization technique: POIMs

[Sonnenburg, Zien, Philips, Rätsch, ISMB, 2008]

- No probabilistic outputs
  - Prerequisite for use in generative models like HMMs
  - Need new approach for combining predictions!

[Tsochantaridis et al., 2005, Rätsch & Sonnenburg 2006]

SVMs vs. Markov Chains

Pro SVMs:

- Much more accurate than Markov chains (and many other methods)
- Continue to improve with more data

[Sonnenburg, Zien, Rätsch, 2006, Sonnenburg et al., 2007]

Contra SVMs:

- (Originally) not large scale
  - New developments: can deal with 10 million examples

[Sonnenburg, Rieck, Rätsch, 2007]

- Difficult to interpret
  - New visualization technique: POIMs

[Sonnenburg, Zien, Philips, Rätsch, ISMB, 2008]

- No probabilistic outputs
  - Prerequisite for use in generative models like HMMs
  - Need new approach for combining predictions!

[Tsochantaridis et al., 2005, Rätsch & Sonnenburg 2006]

SVMs vs. Markov Chains

Pro SVMs:

- Much more accurate than Markov chains (and many other methods)
- Continue to improve with more data

[Sonnenburg, Zien, Rätsch, 2006, Sonnenburg et al., 2007]

Contra SVMs:

- (Originally) not large scale
  - New developments: can deal with 10 million examples

[Sonnenburg, Rieck, Rätsch, 2007]

- Difficult to interpret
  - New visualization technique: POIMs

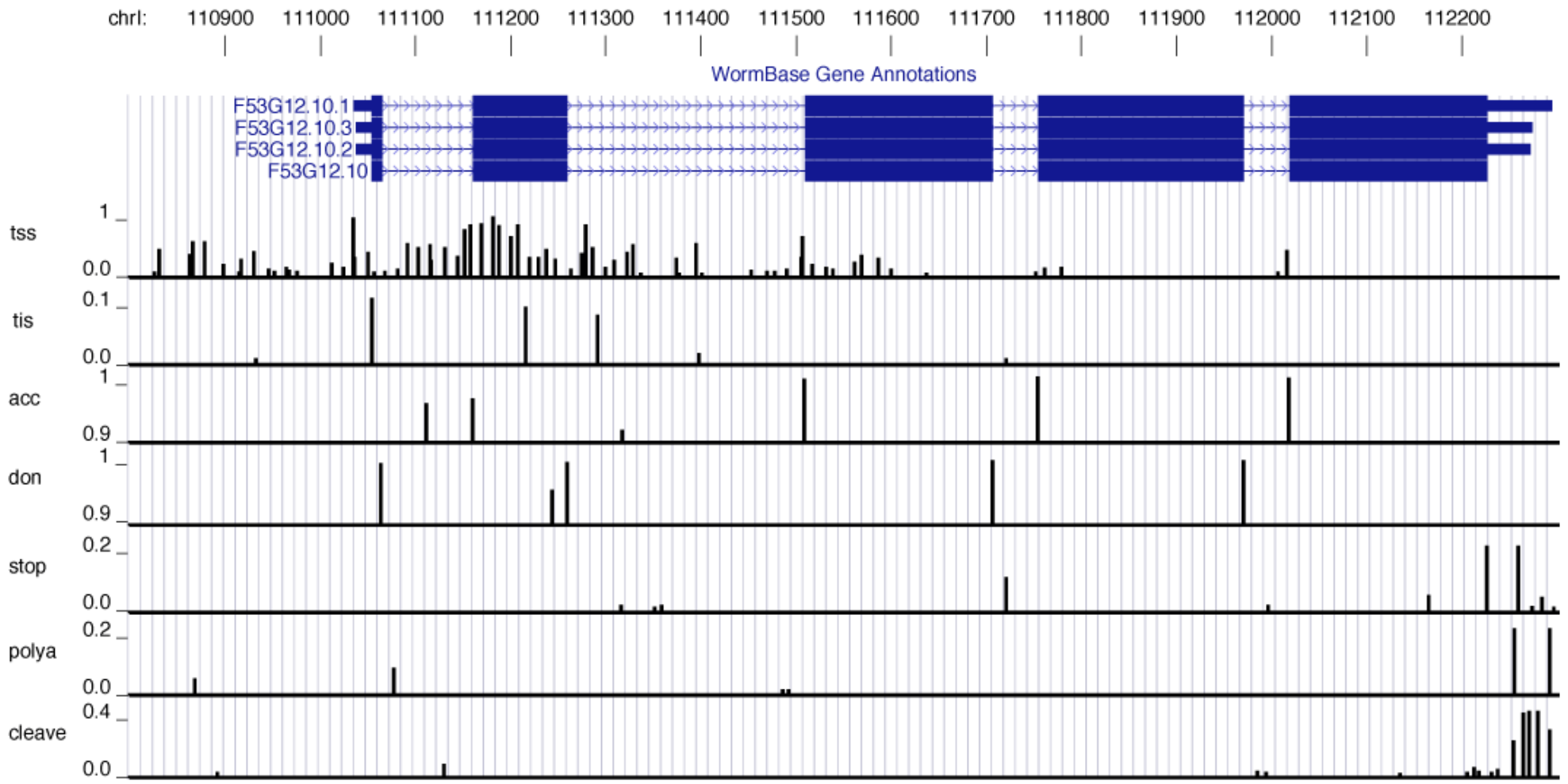
[Sonnenburg, Zien, Philips, Rätsch, ISMB, 2008]

- No probabilistic outputs
  - Prerequisite for use in generative models like HMMs
  - Need new approach for combining predictions!

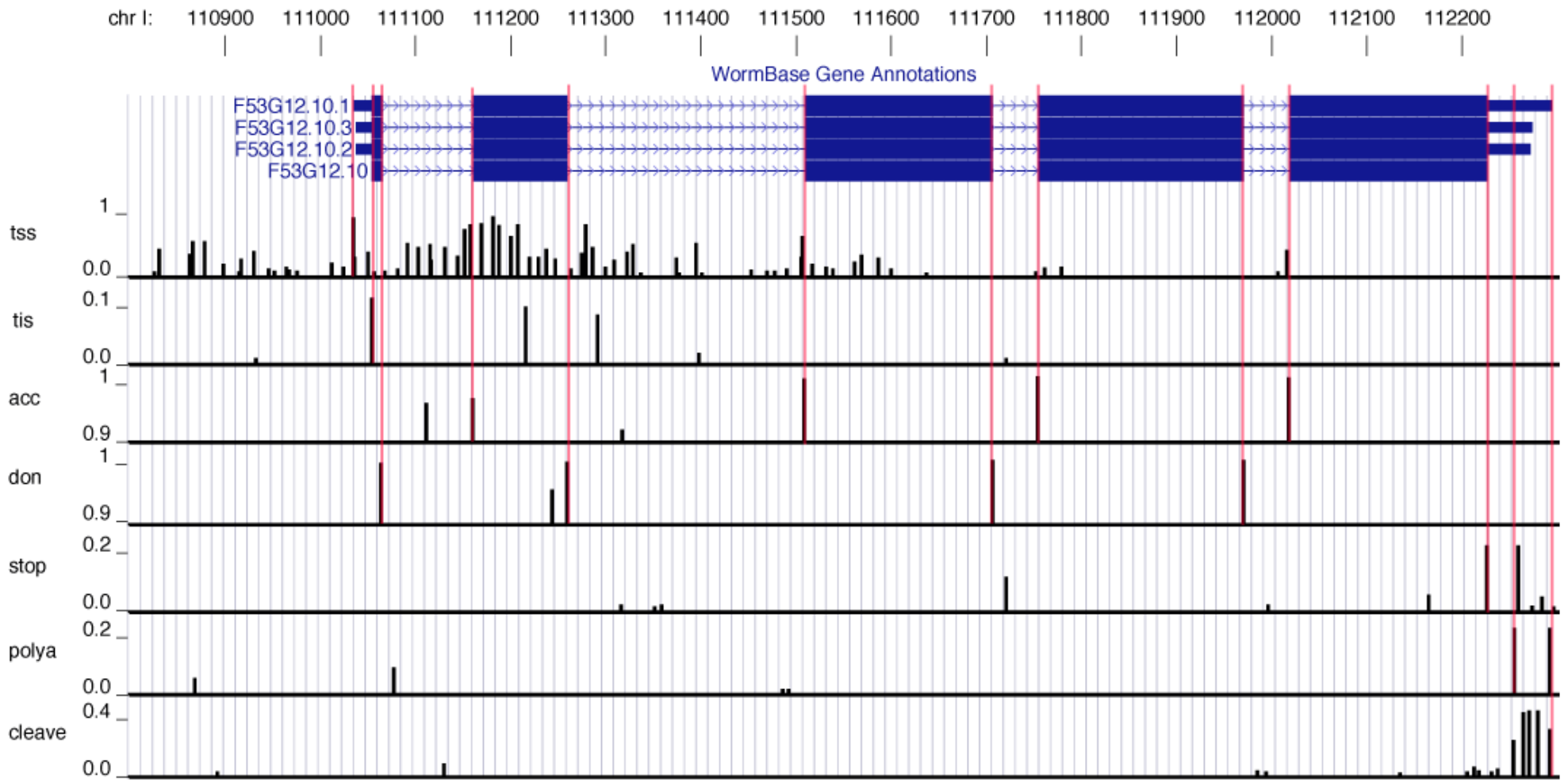
[Tsochantaridis et al., 2005, Rätsch & Sonnenburg 2006]

Example: Predictions in UCSC Browser

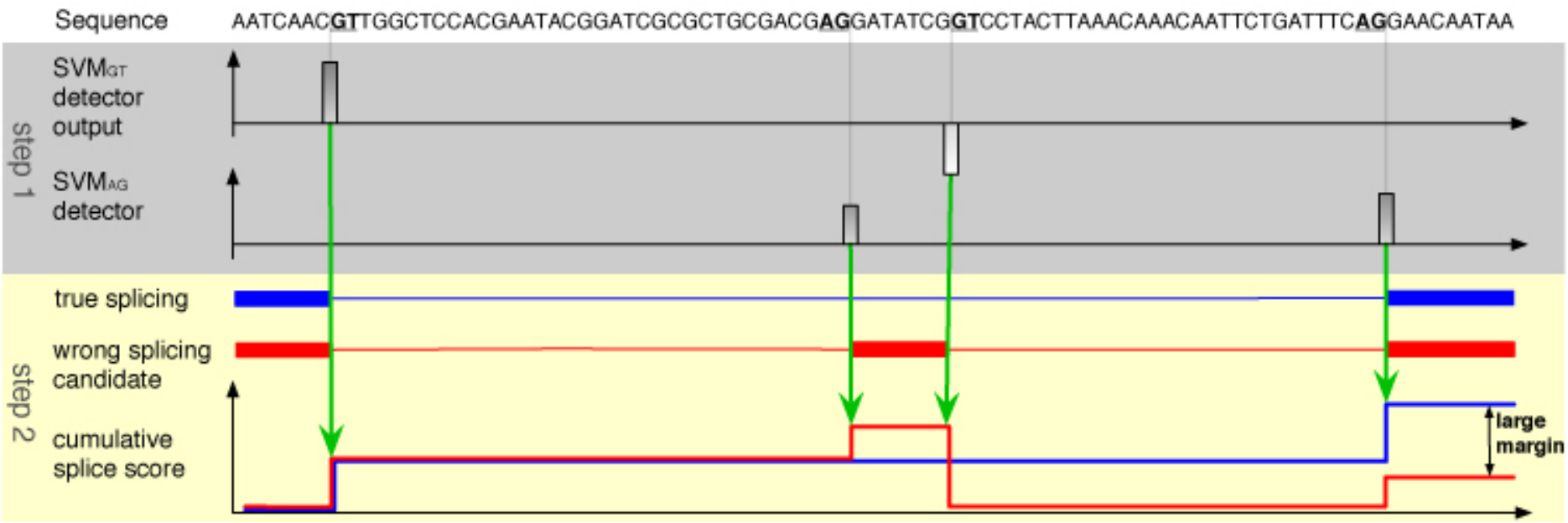
Friedrich Miescher Laboratory
of the Max Planck Society



Example: Predictions in UCSC Browser



Large Margin Learning (simplified)

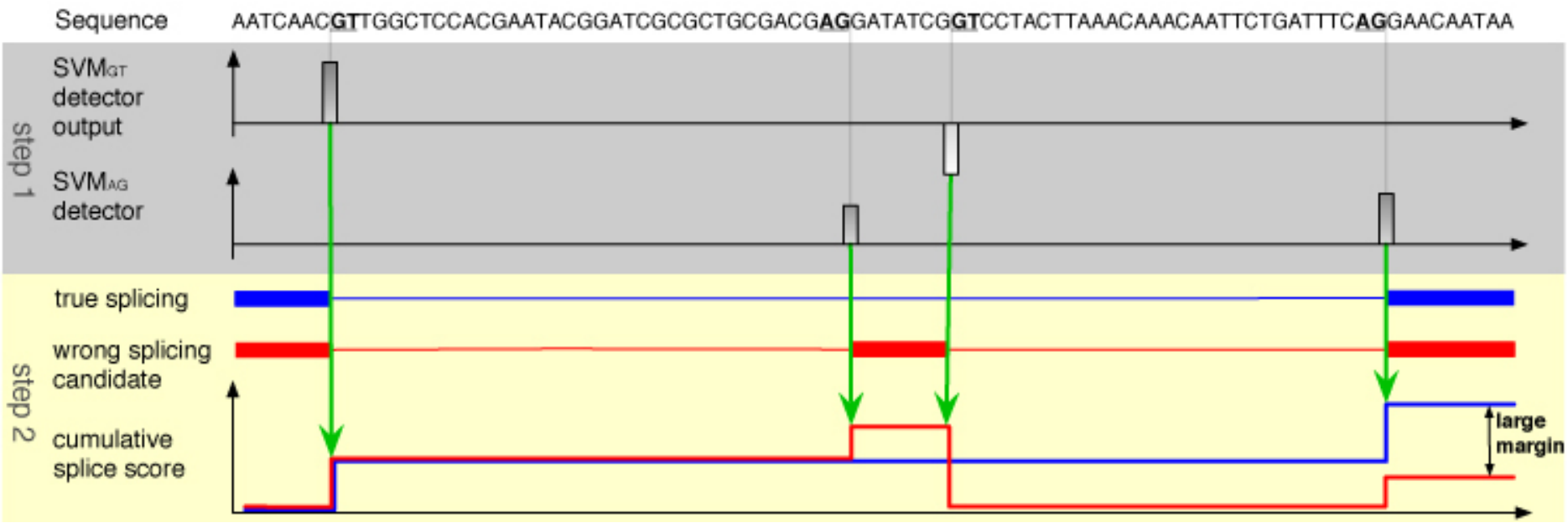


[Rätsch, Sonnenburg, Srinivasan, Witte, Müller, Sommer, Schölkopf, 2007]

- Simplified Model: Score for splice form $\mathbf{y} = \{(p_j, q_j)\}_{j=1}^{J}$:

$$f(\mathbf{y}) := \underbrace{\sum_{j=1}^{J-1} S_{GT}(f_j^{GT}) + \sum_{j=2}^{J} S_{AG}(f_j^{AG})}_{\text{Splice signals}} + \underbrace{\sum_{j=1}^{J-1} S_{L_I}(p_{j+1} - q_j) + \sum_{j=1}^{J} S_{L_E}(q_j - p_j)}_{\text{Segment lengths}}$$

Large Margin Learning (simplified)



[Rätsch, Sonnenburg, Srinivasan, Witte, Müller, Sommer, Schölkopf, 2007]

- Simplified Model: Score for splice form $\mathbf{y} = \{(p_j, q_j)\}_{j=1}^{J}$:

$$f(\mathbf{y}) := \underbrace{\sum_{j=1}^{J-1} S_{GT}(f_j^{GT}) + \sum_{j=2}^{J} S_{AG}(f_j^{AG})}_{\text{Splice signals}} + \underbrace{\sum_{j=1}^{J-1} S_{L_I}(p_{j+1} - q_j) + \sum_{j=1}^{J} S_{L_E}(q_j - p_j)}_{\text{Segment lengths}}$$

The Learning Problem

Given a sequence of observations (features) $x \in X$
We want to learn a function:

$$f : X \to Y$$

which yields a **label sequence (or path)** $\pi \in Y$
(of equal length: $|x| = |y|$).

Employ a function

$$F_\Theta : X \times Y \to \mathbb{R} \qquad \text{(path scoring)}$$

with which

$$f(x) = \arg\max_{\pi \in Y} F(x, \pi) \qquad \text{(Viterbi decoding)}$$

The Learning Problem

The discriminant function F_Θ (simplified)

$$F_\Theta(x,\pi) = \sum_{p=1}^{L} \sum_{j,k} [[\pi_p = k]]\, g_{j,k}(x_{j,p}) + \phi(\pi_{p-1}, \pi_p)$$

- $g_{j,k}$: feature scoring contributions, for each pair of features $j \in 1,...,M$ and states $k \in 1,...,S$
- ϕ: transition scores

Large margin separation

$$F_\Theta(x^i,\pi^i) - F_\Theta(x^i,\overline{\pi}) \gg 0 \quad \forall \overline{\pi} \neq \pi^i \quad \forall i$$

between the correct labeling π^i
and any other (wrong) labeling $\overline{\pi}$.

The Learning Problem

The discriminant function F_Θ (simplified)

$$F_\Theta(x,\pi) = \sum_{p=1}^{L} \sum_{j,k} [[\pi_p = k]]\, g_{j,k}(x_{j,p}) + \phi(\pi_{p-1}, \pi_p)$$

- $g_{j,k}$: feature scoring contributions, for each pair of features $j \in 1,...,M$ and states $k \in 1,...,S$
- ϕ: transition scores

Large margin separation

$$F_\Theta(x^i,\pi^i) - F_\Theta(x^i,\overline{\pi}) \gg 0 \quad \forall \overline{\pi} \neq \pi^i \quad \forall i$$

between the correct labeling π^i
and any other (wrong) labeling $\overline{\pi}$.

Training: Solve Optimization Problem

$$\min_{\boldsymbol{\Theta},\, \boldsymbol{\xi} \geq \mathbf{0}} \quad \frac{1}{n} \sum_{i=1}^{n} \xi^{(i)} + C\, \Omega(\boldsymbol{\Theta})$$

$$\text{s.t.} \quad F_{\Theta}(\mathbf{x}^{(\mathbf{i})}, \boldsymbol{\pi}^{(i)}) - F_{\Theta}(\mathbf{x}^{(i)}, \overline{\boldsymbol{\pi}}) \;\geq\; 1 - \xi^{(i)}$$

$$\forall \overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)} \quad \forall i = 1, \ldots, n,$$

- $\boldsymbol{\xi}$: slack variables
- Ω: linear regularization term

Training: Solve Optimization Problem

$$\min_{\boldsymbol{\Theta},\, \boldsymbol{\xi} \geq \mathbf{0}} \quad \frac{1}{n} \sum_{i=1}^{n} \xi^{(i)} + C\, \Omega(\boldsymbol{\Theta})$$

$$\text{s.t.} \quad F_{\Theta}(\mathbf{x}^{(i)}, \boldsymbol{\pi}^{(i)}) - F_{\Theta}(\mathbf{x}^{(i)}, \overline{\boldsymbol{\pi}}) \;\geq\; 1 - \xi^{(i)}$$

$$\forall \overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)} \quad \forall i = 1, \ldots, n,$$

- $\boldsymbol{\xi}$: slack variables
- Ω: linear regularization term

Exponentially number of margin constraints !

Training: Solve Optimization Problem

$$\min_{\boldsymbol{\Theta},\, \boldsymbol{\xi} \geq \mathbf{0}} \quad \frac{1}{n} \sum_{i=1}^{n} \xi^{(i)} + C\, \Omega(\boldsymbol{\Theta})$$

$$\text{s.t.} \quad F_{\Theta}(\mathbf{x}^{(\mathbf{i})}, \boldsymbol{\pi}^{(i)}) - F_{\Theta}(\mathbf{x}^{(i)}, \overline{\boldsymbol{\pi}}) \;\geq\; 1 - \xi^{(i)}$$

$$\forall \overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)} \quad \forall i = 1, \ldots, n,$$

- $\boldsymbol{\xi}$: slack variables
- Ω: linear regularization term

Employ "Column Generation" strategy

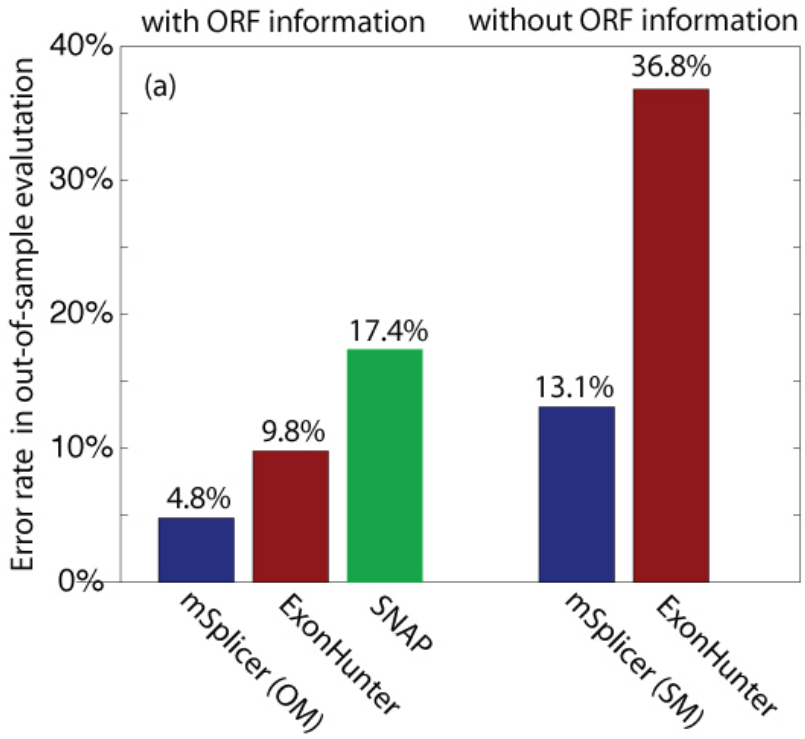
- Initialize with empty set of constraints, random Θ
- Compute paths that maximally violates the margin constraints
- Add constraint and solve intermediate LP
- Run until convergence

Results Summary

- Splice forms only [Rätsch et al., PLoS Comp. Biol., 2007]
  - Comparison with other methods
  - Analysis of a few disagreeing cases
  - Predictions available on `http://www.wormbase.org`
- Full gene predictions
  - Participation at nGASP competition
  - Preliminary evaluation
  - Predictions available on `http://www.wormbase.org`

Results I (Splice forms only)

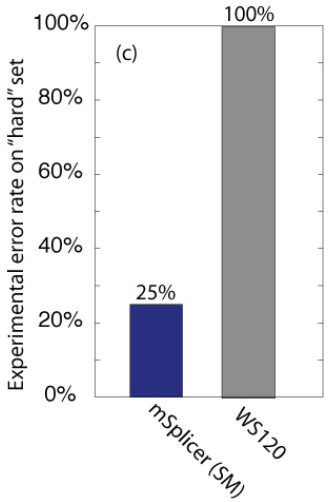
- ≈3,800 gene models derived from cDNAs and ESTs
  - 60% for training and validation
  - 40% for testing (exclude alt. spliced genes)
- Out-of-sample accuracy (≈1100 gene models):
  - Splice form error rate
    - 4.8% (coding)
    - 13.1% (mixed)
  - Much lower error rates than state-of-the-art
    - Exonhunter (Brejova et al., ISMB'05)
    - Snap (Korf, BMC Bioinformatics 2004)

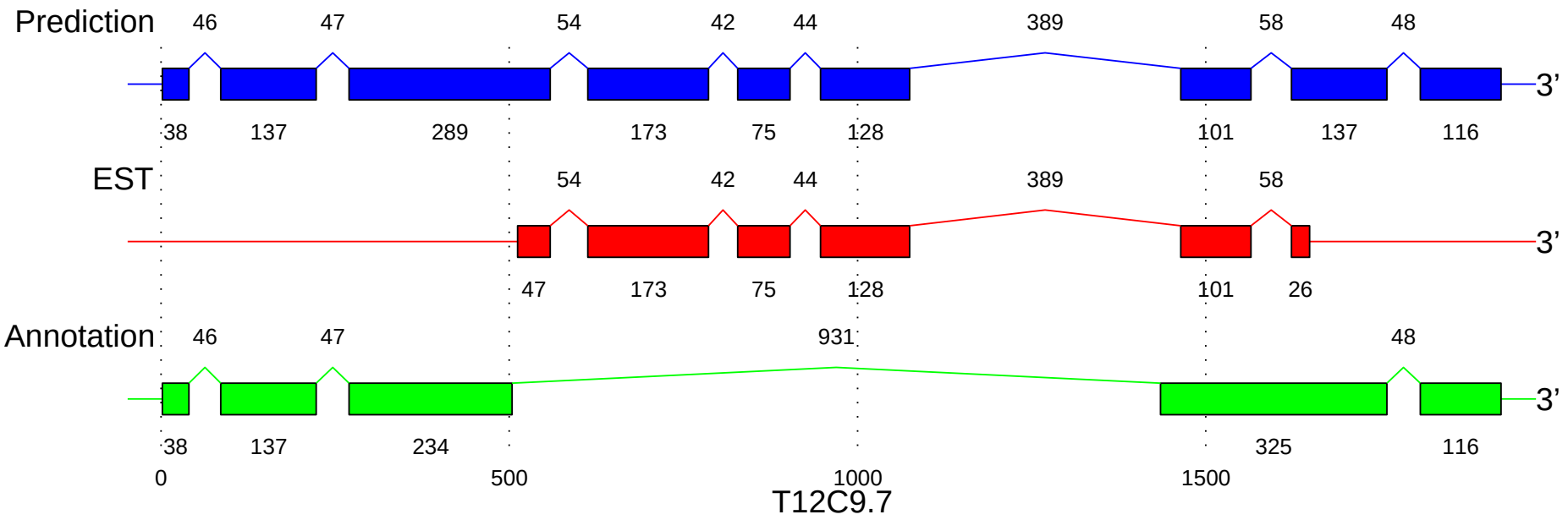


[Rätsch et al., PLoS Comp. Biology, 2007]

Results II (Splice form only)

- Validation by RT-PCR & direct sequencing
  - Consider 20 disagreeing cases
  - Annotation was never correct
  - 75% of our predictions were correct
    (In collaboration with the Sommer lab.)



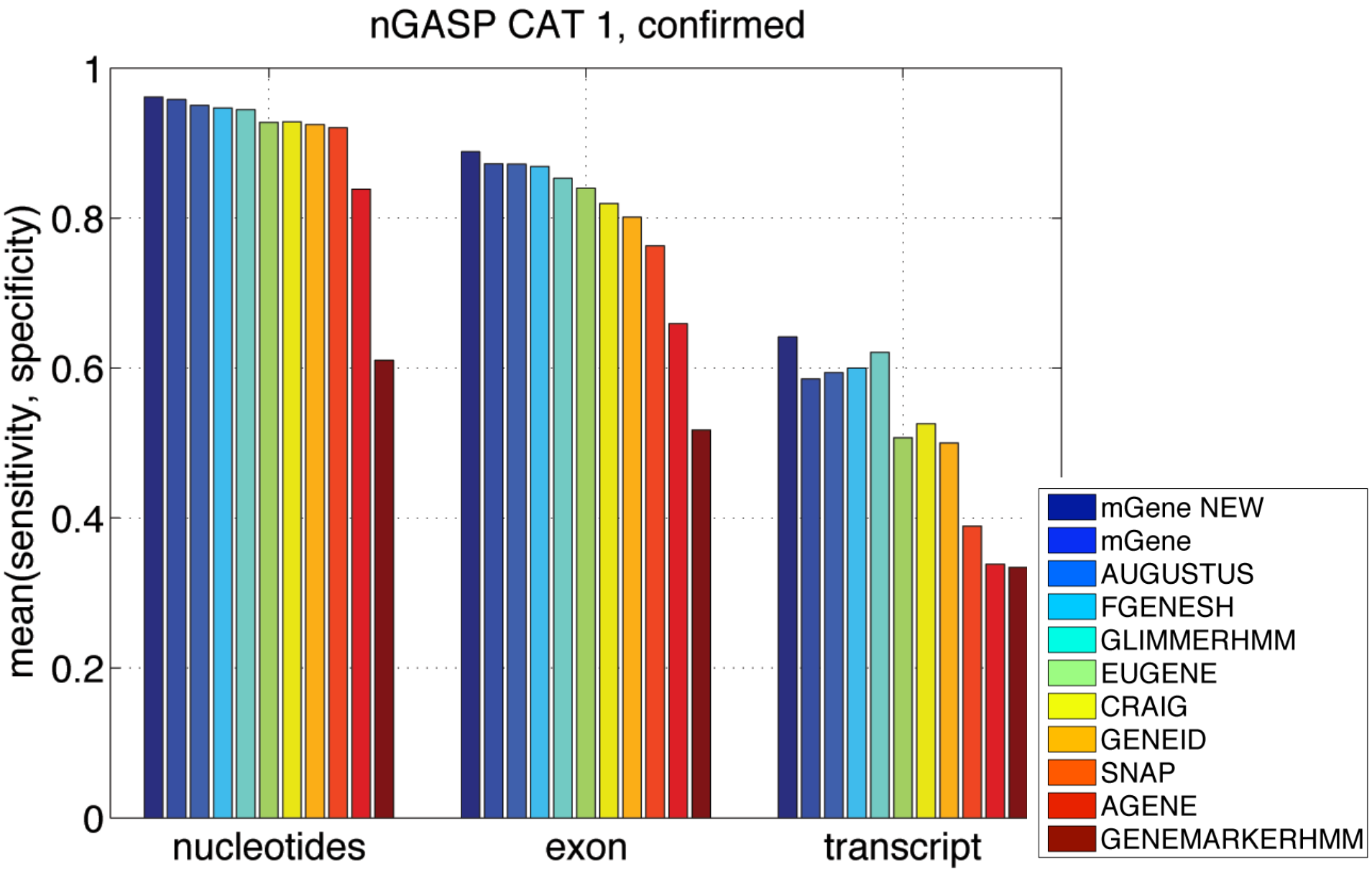


[Rätsch et al., PLoS Comp. Biology, 2007]

Results III: nGASP Competition

- Controlled competition conditions:
  - 10% for training methods
  - 10% for evaluation
- Phase I: single predictors
- Phase II: combining algorithms
- Four categories:
  - *Ab initio* gene finders
  - Dual/Multi-genome gene finders
  - Gene finders that use sequences alignments
  - Gene finders that combine any of above information
- Preliminary evaluation on WS160 fully confirmed genes

nGASP Cat. 1 evaluation (prelim.)



Discussion

- Very good on nucleotide and exon level
- Substantial improvements on transcript level
- Re-annotation of *C. elegans* in official wormbase annotation
  - total: 19532 genes predicted
  - 635 *new* genes
  - 350 unconfirmed genes are not predicted
- Wetlab confirmations (preliminary results)
  - Confirmed 55 of 103 newly predicted genes
  - Confirmed only 4 of 50 annotated genes that were not predicted
- Annotation of 6 other nematodes (`http://www.wormbase.org`)
- Annotations of *Arabidopsis* species is under way

Summary and Future Work

- High accuracy on constitutively spliced genes for nematodes
- First version of new gene finder system seems very promising
  - Further improvements seem possible
- Computational challenge: Include alternative splicing
  - Often incomplete information
  - Needs more data!?
    - New sequencing technologies
    - Tiling arrays
- Re-annotation of *Caenorhabditis* and other genomes
  - How can we learn from one organism for others?
- General method for sequence segmentation
  - mRNA to DNA alignments
    (Schulze et al., 2007, De Bona et al., 2008)Schulze et al. [2007], Bona et al. [2008]
  - Segmentation of Resequencing Data (Zeller et al., 2008)
  - Transcript Mapping from Tiling Arrays (Zeller et al., 2008)

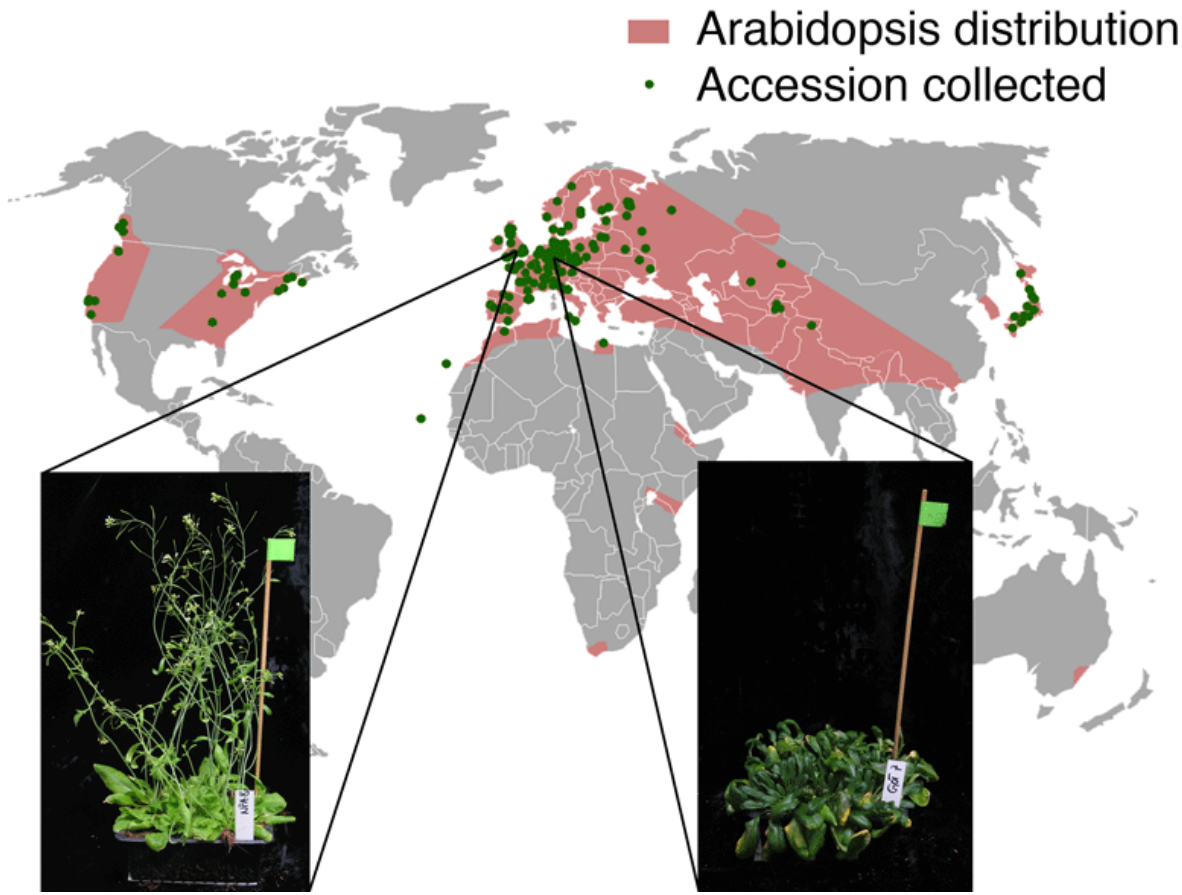
Acknowledgments & Links

- mSplicer contributors: Sören Sonnenburg, Jagan Srinivasan, Hanh Witte, Ralf Sommer, Klaus-Robert Müller, Bernhard Schölkopf
- mGene contributors: Gabriele Schweikert, Georg Zeller, Alexander Zien, Cheng Soon Ong, Sören Sonnenburg

- Group homepage:
  `http://www.fml.mpg.de/raetsch`
- Details, data sets & predictions (soon):
  `http://www.msplicer.org` and `http://www.mgene.org`
- A large scale open source ML toolbox:
  `http://www.shogun-toolbox.org`
- Other projects:
  `http://www.fml.mpg.de/raetsch/projects`

Part II

Machine Learning for Polymorphism Identification in *Arabidopsis thaliana*



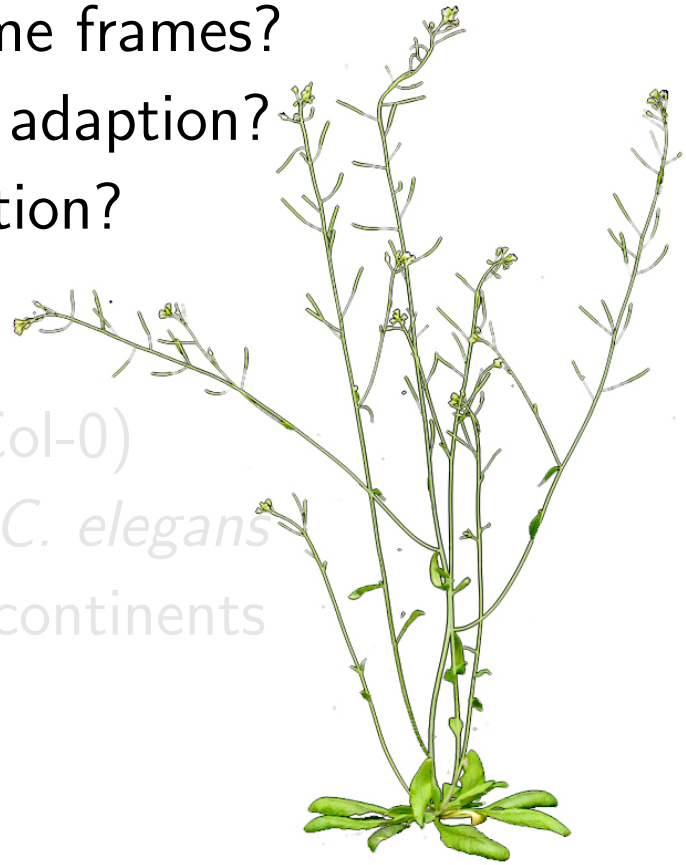
What is the genetic basis of variation?

Modified from Koornneef et al. 2004. *Ann. Rev. Plant Biology*, 55, 141-172.

Introduction

Questions:

- What sequence changes occur in short time frames?
- Which polymorphisms and genes underlie adaption?
- What are the consequences for gene function?



Arabidopsis thaliana:

- 119 Mb finished euchromatic sequence (Col-0)
- Resources comparable to *Drosophila* and *C. elegans*
- Collections of >1000 wild strains from 3 continents
- Strains are largely homozygous

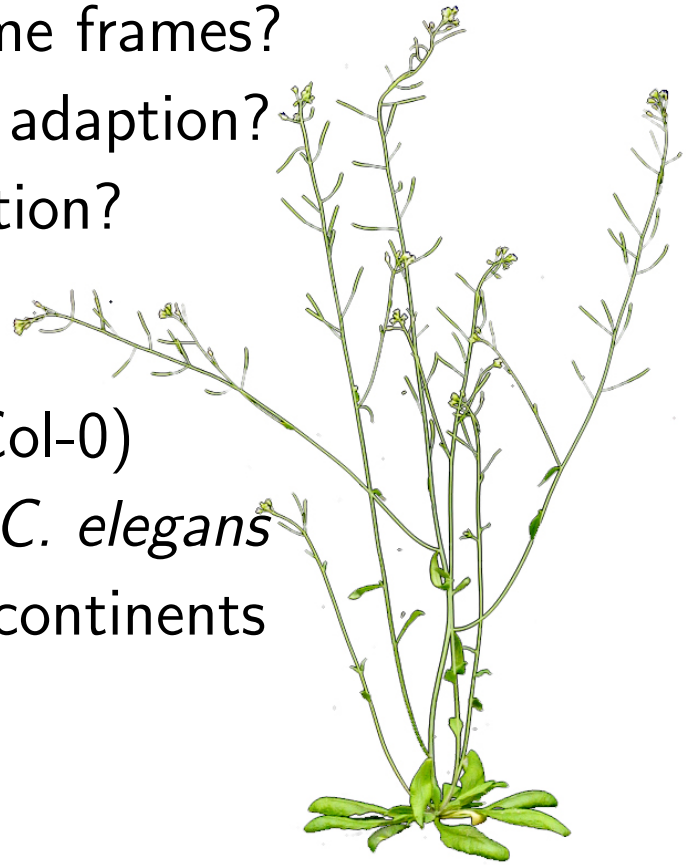
Resequencing of 20 wild strains

- Genome-wide identification of sequence polymorphisms
- High-density oligo-nucleotide arrays for high-throughput resequencing

Introduction

Questions:

- What sequence changes occur in short time frames?
- Which polymorphisms and genes underlie adaption?
- What are the consequences for gene function?



Arabidopsis thaliana:

- 119 Mb finished euchromatic sequence (Col-0)
- Resources comparable to *Drosophila* and *C. elegans*
- Collections of >1000 wild strains from 3 continents
- Strains are largely homozygous

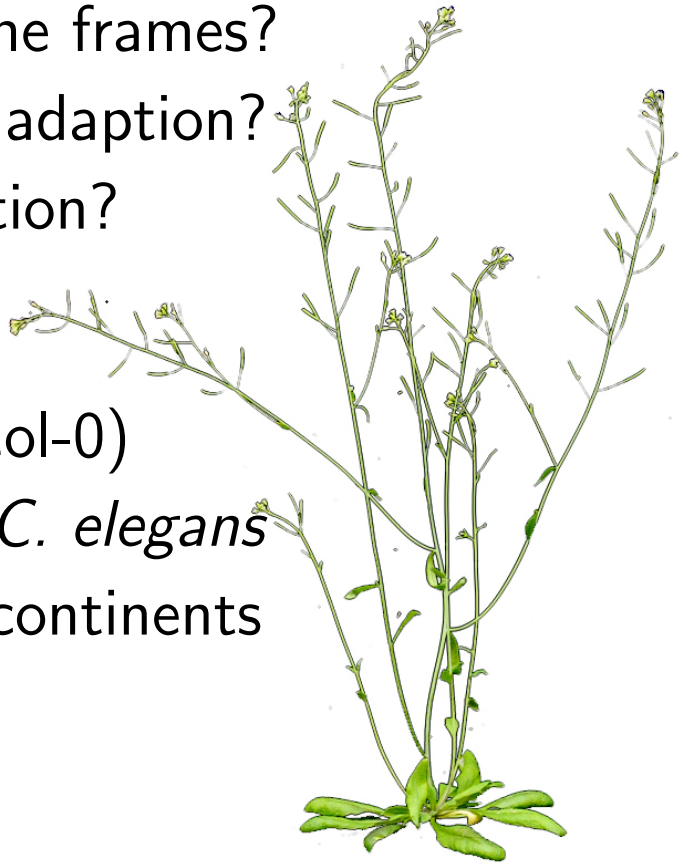
Resequencing of 20 wild strains

- Genome-wide identification of sequence polymorphisms
- High-density oligo-nucleotide arrays for high-throughput resequencing

Introduction

Questions:

- What sequence changes occur in short time frames?
- Which polymorphisms and genes underlie adaption?
- What are the consequences for gene function?



Arabidopsis thaliana:

- 119 Mb finished euchromatic sequence (Col-0)
- Resources comparable to *Drosophila* and *C. elegans*
- Collections of >1000 wild strains from 3 continents
- Strains are largely homozygous

Resequencing of 20 wild strains

- Genome-wide identification of sequence polymorphisms
- High-density oligo-nucleotide arrays for high-throughput resequencing

Resequencing Array Basics I



Resequencing Array Basics II

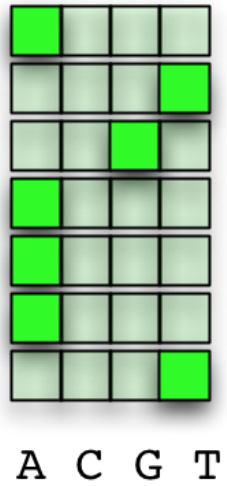
Reference DNA sequence



ACGTAAAGTCGAATGAAATGACCCCTTTTGAGAGCCCCGTT

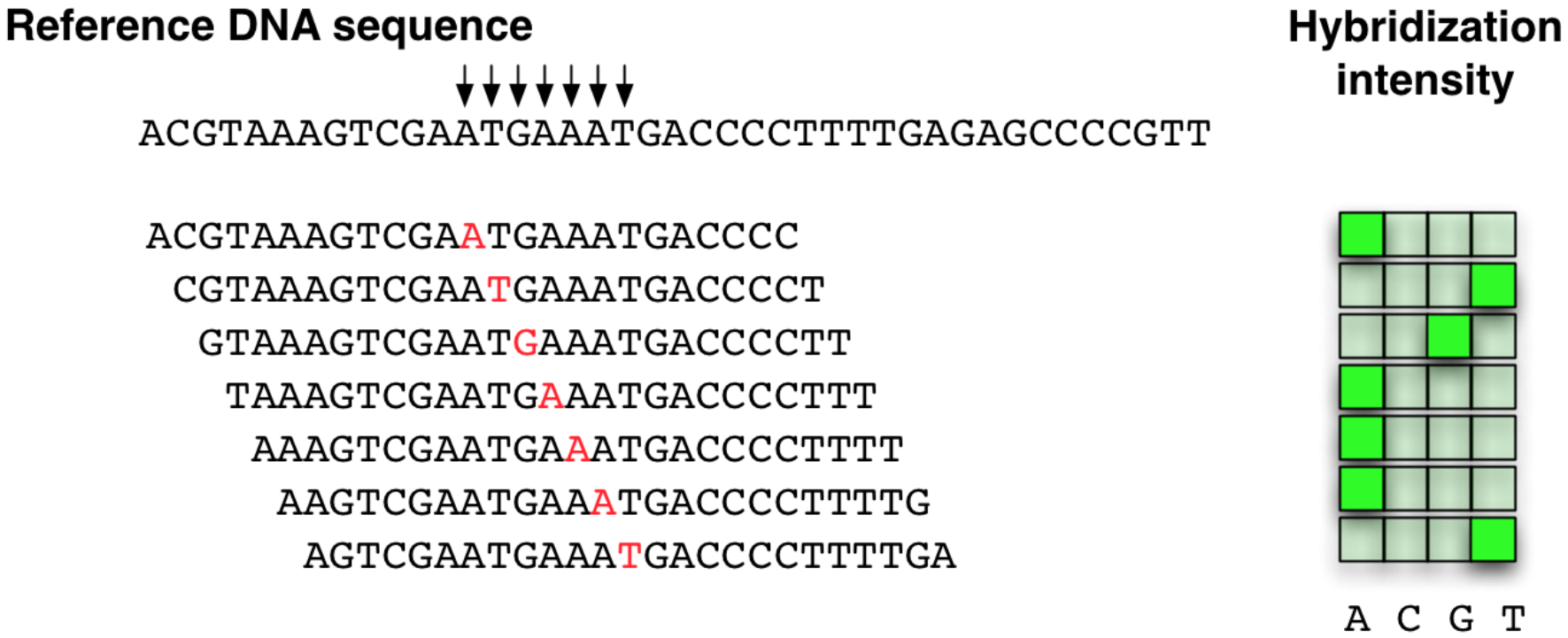
ACGTAAAGTCGAATGAAATGACCCC
CGTAAAGTCGAATGAAATGACCCCT
GTAAAGTCGAATGAAATGACCCCTT
TAAAGTCGAATGAAATGACCCCTTT
AAAGTCGAATGAAATGACCCCTTTT
AAGTCGAATGAAATGACCCCTTTTG
AGTCGAATGAAATGACCCCTTTTGA

Hybridization intensity



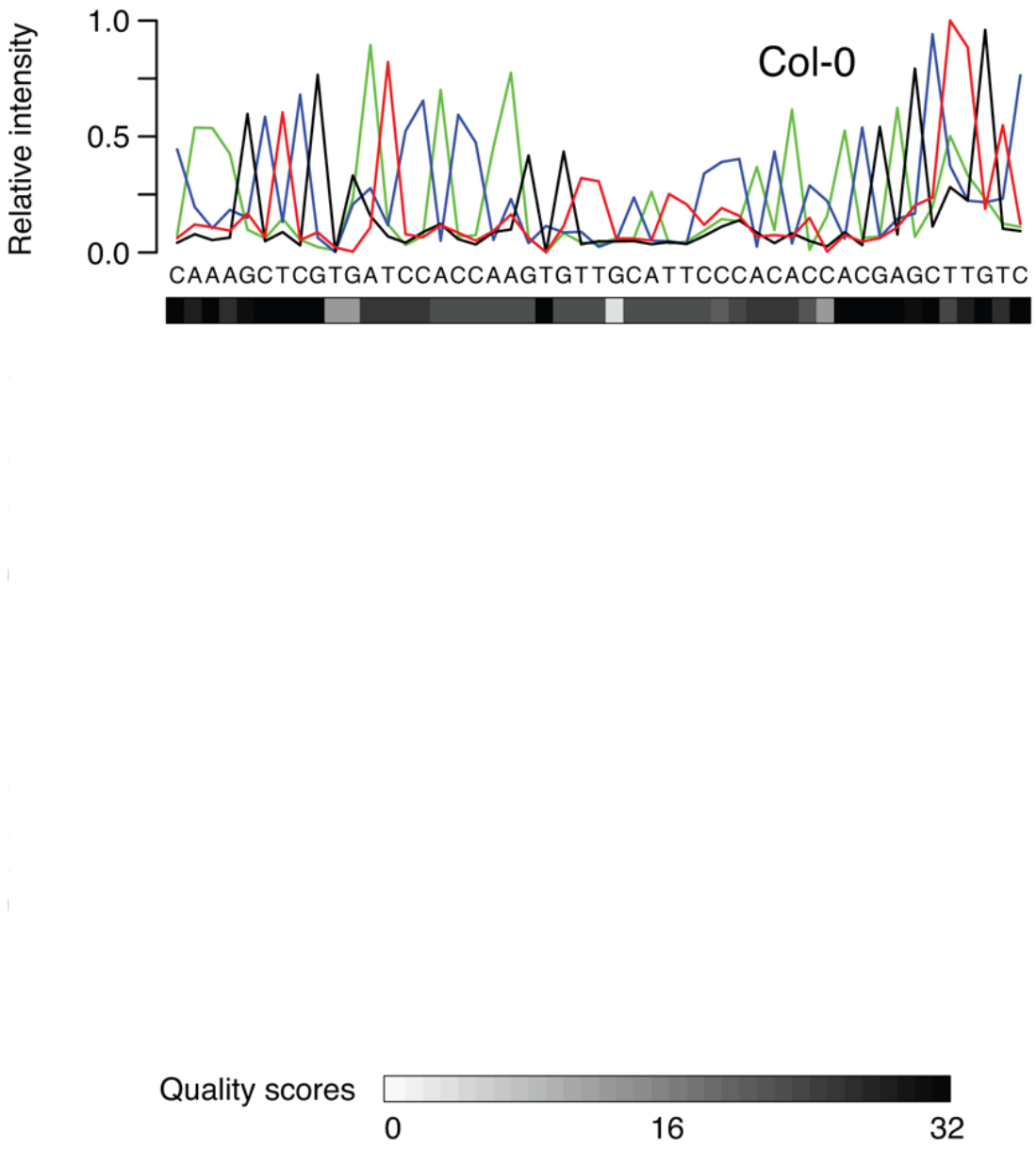
- >99.99% of bases represented
- Each base queried with forward and reverse strand probe quartets
- Nearly 1 billion oligos per ecotype
- 19+1 ecotypes surveyed representing worldwide distribution

Resequencing Array Basics II



- >99.99% of bases represented
- Each base queried with forward and reverse strand probe quartets
- Nearly 1 billion oligos per ecotype
- 19+1 ecotypes surveyed representing worldwide distribution

Resequencing Data



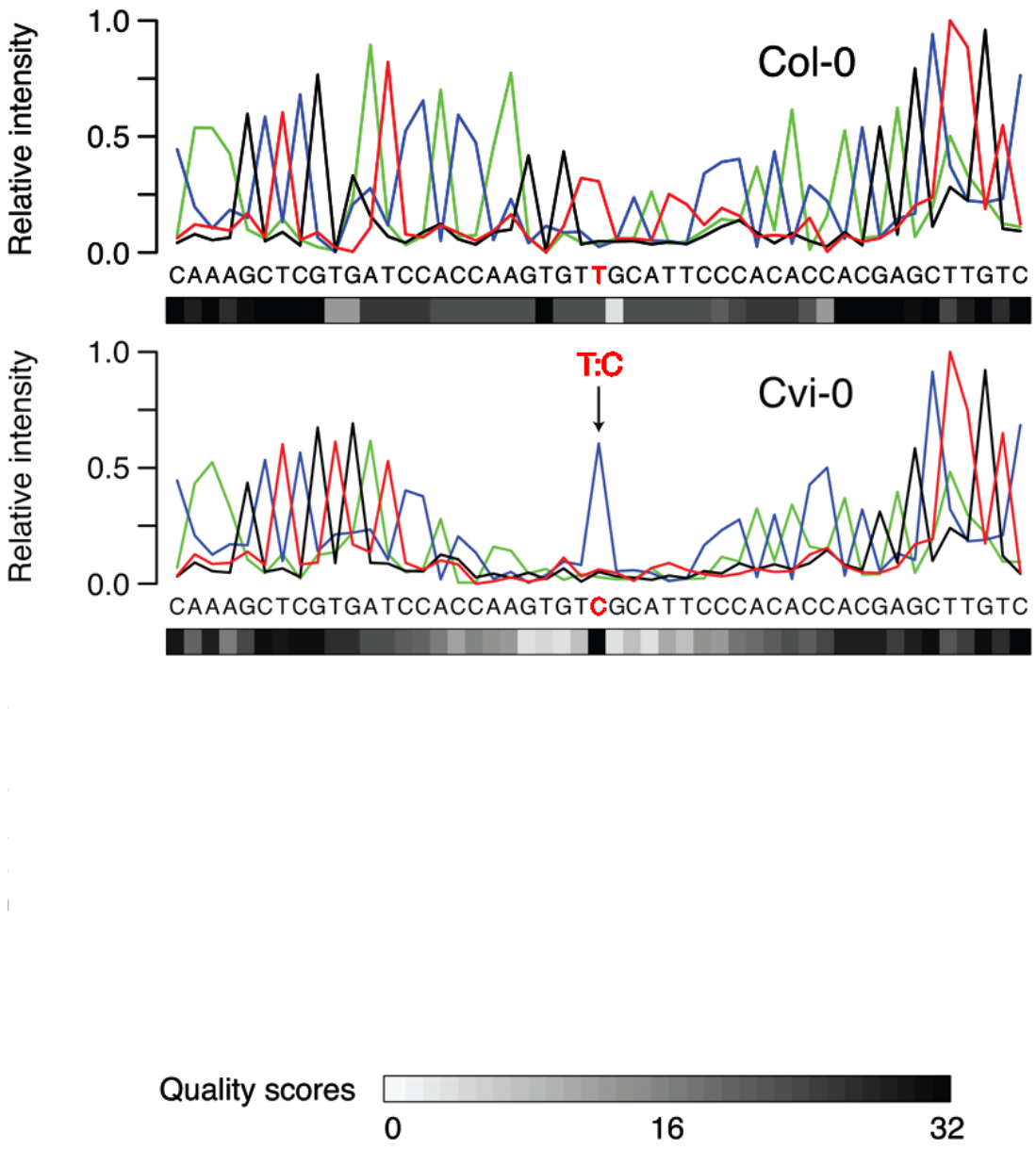
Data analysis challenge

- Hybridization intensities depend on
  - Oligomer
  - Ecotype
  - Repeats
- Measurement noise
- Identify SNPs

Problematic cases

- Highly polymorphic regions
- Deletions/insertions

Resequencing Data



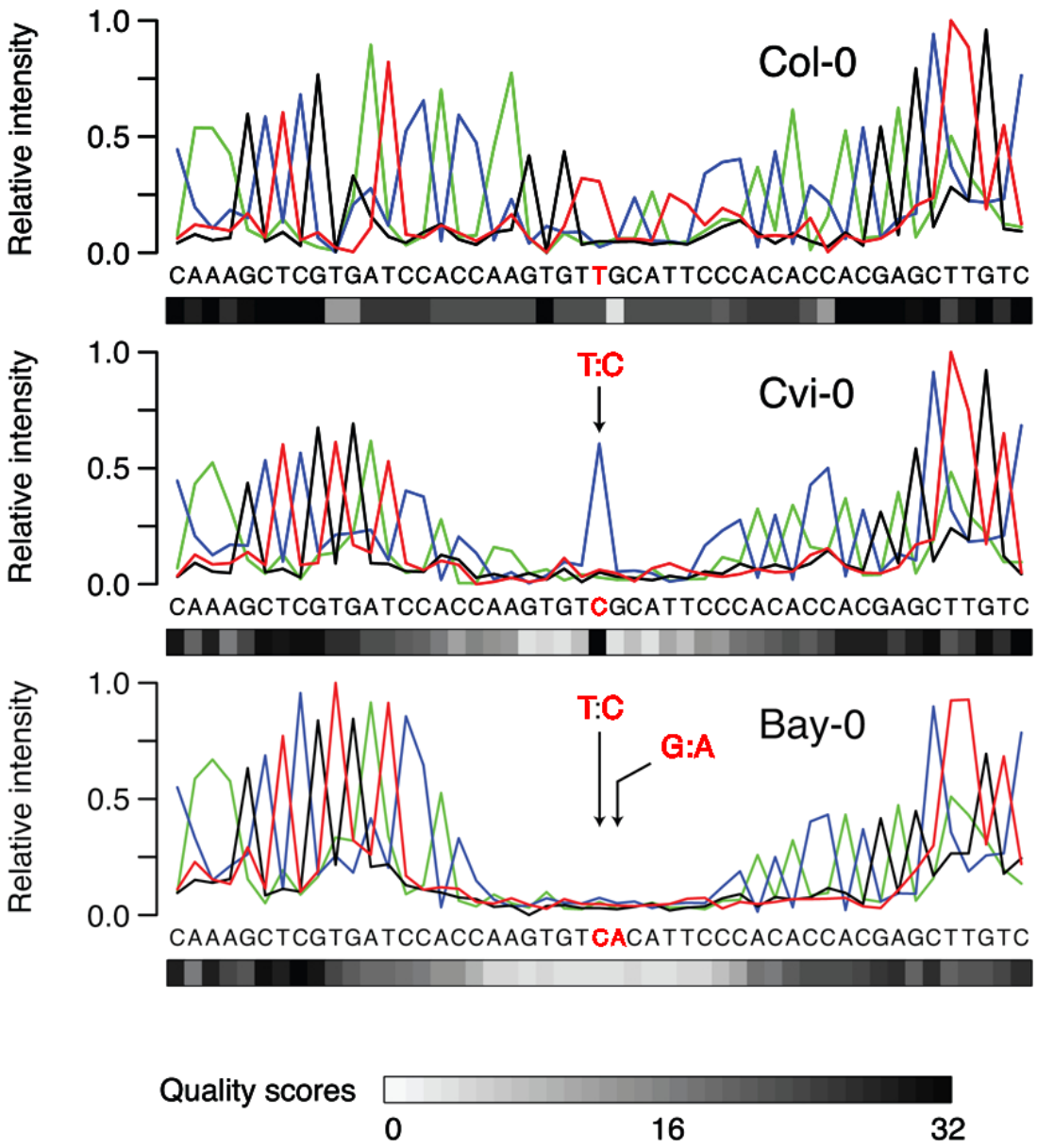
Data analysis challenge

- Hybridization intensities depend on
  - Oligomer
  - Ecotype
  - Repeats
- Measurement noise
- Identify SNPs

Problematic cases

- Highly polymorphic regions
- Deletions/insertions

Resequencing Data



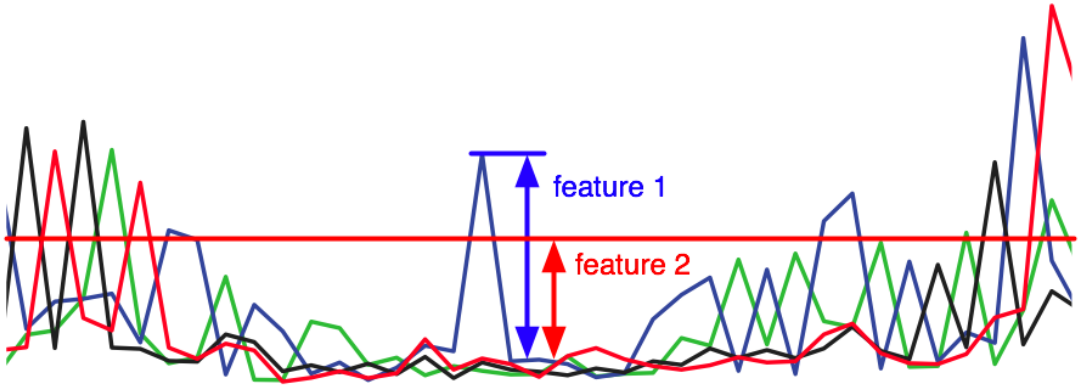
Data analysis challenge

- Hybridization intensities depend on
  - Oligomer
  - Ecotype
  - Repeats
- Measurement noise
- Identify SNPs

Problematic cases

- Highly polymorphic regions
- Deletions/insertions

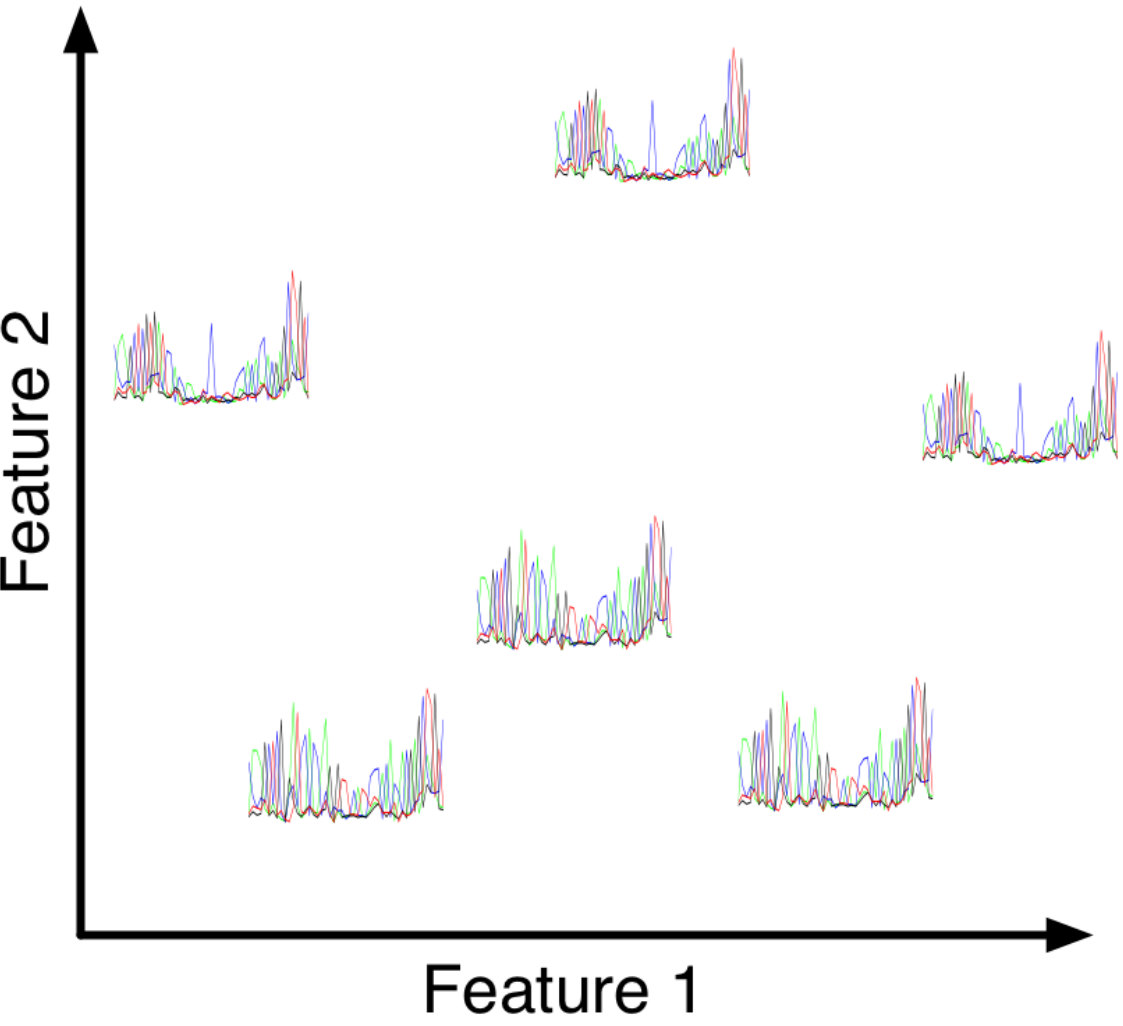
SVMs for SNP Identification



Large Margin Learning

- Extract 302 features
  - Hybridization
  - $\pm$4nt around pos.
- Use $\approx$600kbp/strain labeled positions
- Use SVMs to find non-linear separation
- Tricks:
  - Use filters to reduce number of negatives
  - Use information across strains

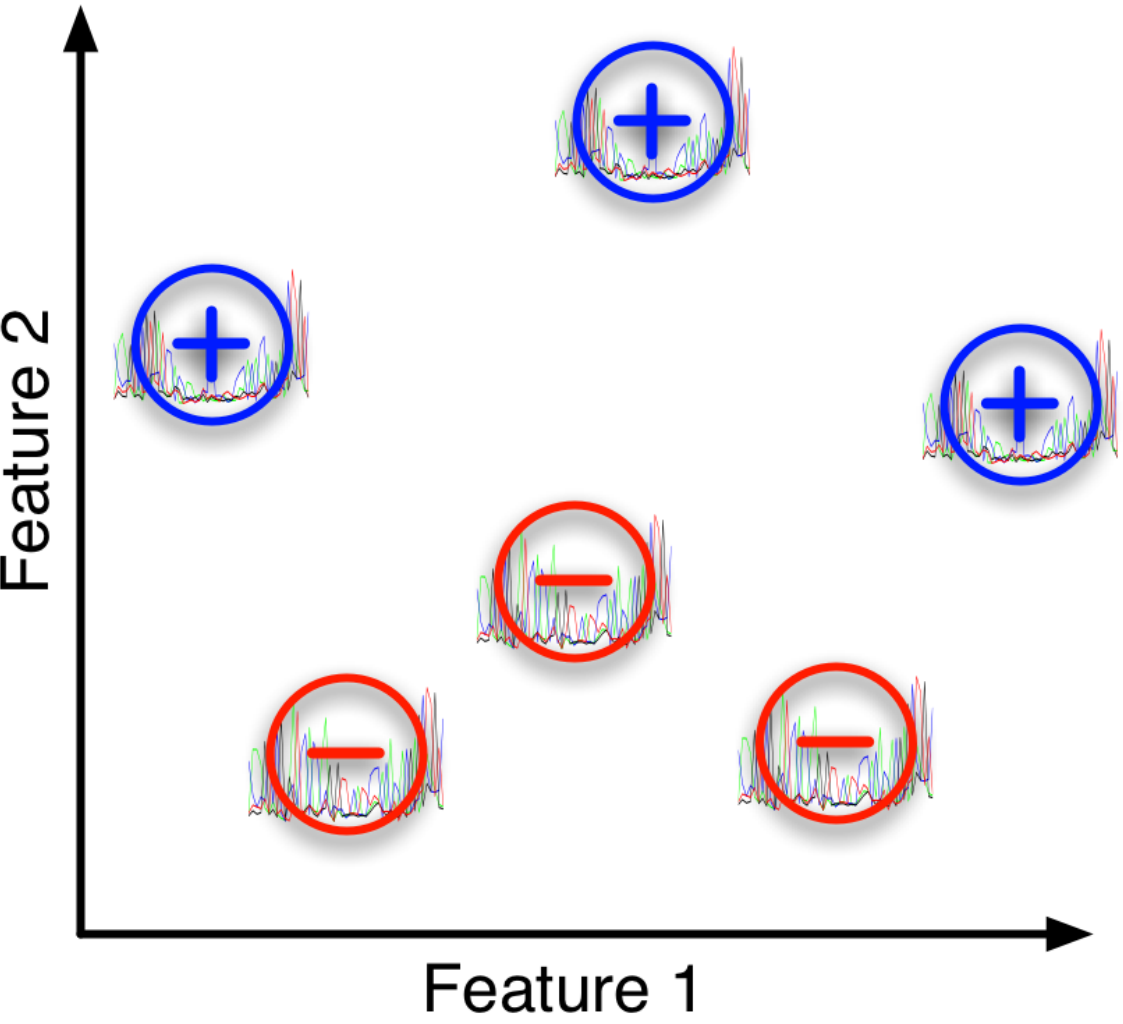
SVMs for SNP Identification



Large Margin Learning

- Extract 302 features
  - Hybridization
  - $\pm$4nt around pos.
- Use $\approx$600kbp/strain labeled positions
- Use SVMs to find non-linear separation
- Tricks:
  - Use filters to reduce number of negatives
  - Use information across strains

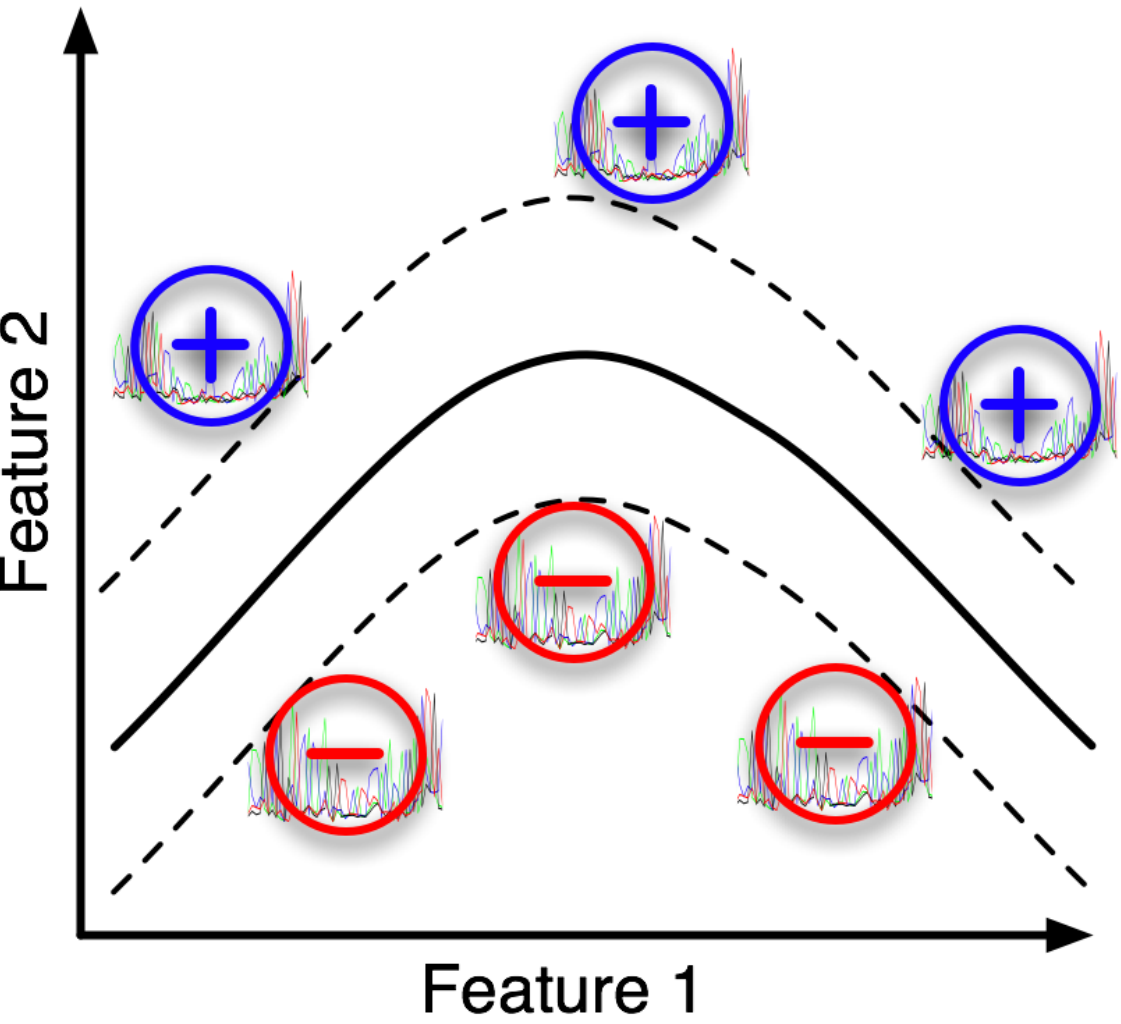
SVMs for SNP Identification



Large Margin Learning

- Extract 302 features
  - Hybridization
  - $\pm$4nt around pos.
- Use $\approx$600kbp/strain labeled positions
- Use SVMs to find non-linear separation
- Tricks:
  - Use filters to reduce number of negatives
  - Use information across strains

SVMs for SNP Identification

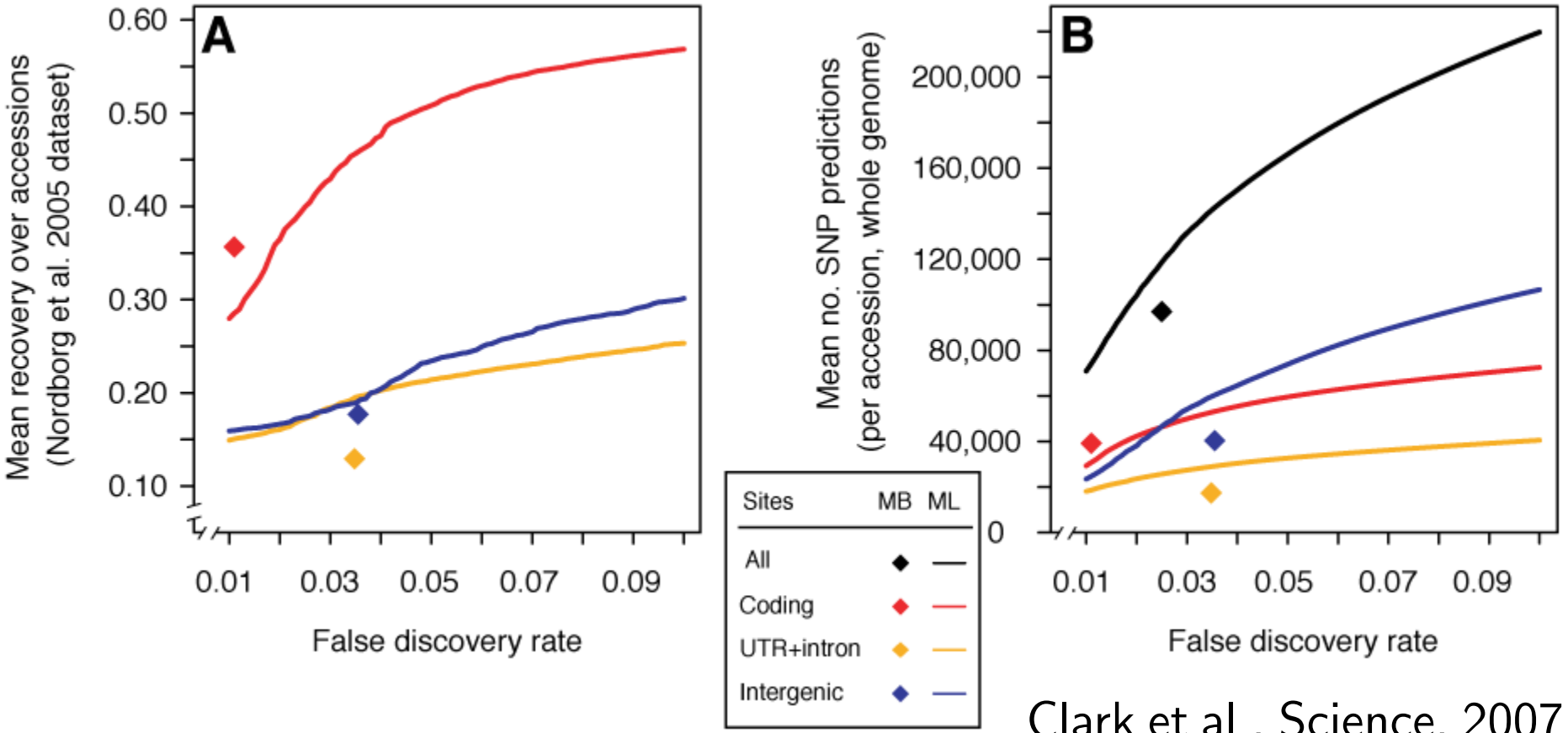


Large Margin Learning

- Extract 302 features
  - Hybridization
  - $\pm$4nt around pos.
- Use $\approx$600kbp/strain labeled positions
- Use SVMs to find non-linear separation
- Tricks:
  - Use filters to reduce number of negatives
  - Use information across strains

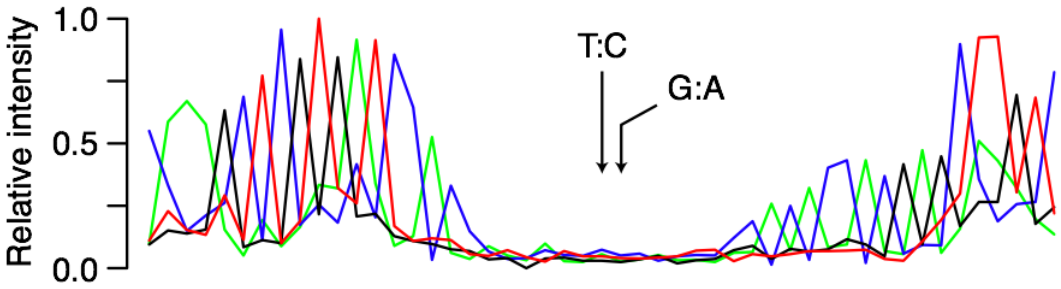
Application to SNP Discovery

- Classification using Support Vector Machines with 302 features
- ≈2,400 known SNPs/ecotype (Nordborg et al., PLoS Biol., 2005)
- Out-of-sample evaluation and prediction on whole genome
- Comparison with Perlegens model based method (Hinds et al., Science, 2005)



Clark et al., Science, 2007.

Limitations of the Technique



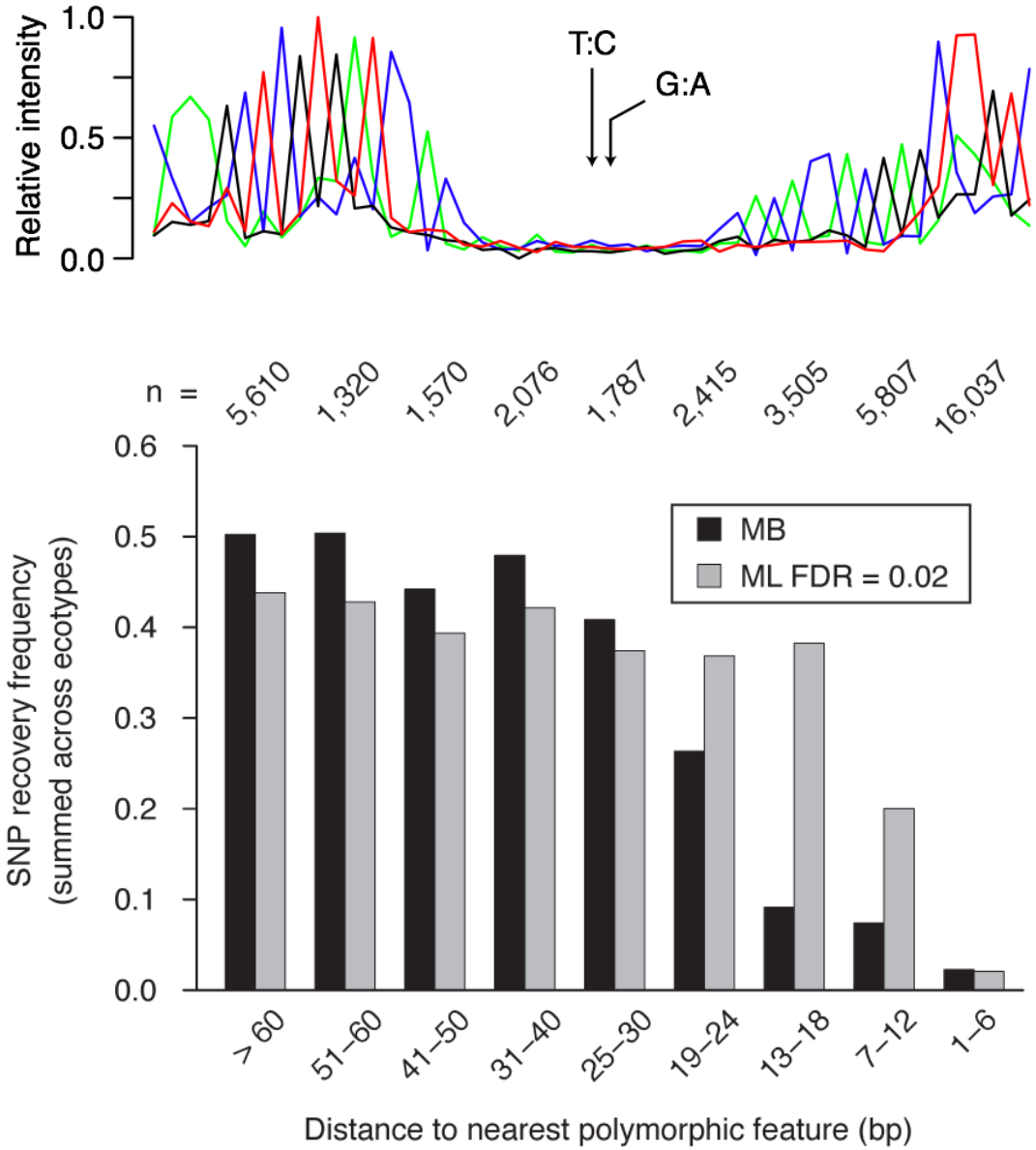
Results

- Performance drops, when other SNPs are in vicinity
- Least predicted SNPs in highly polymorphic regions!
- ML more sensitive

New Approach

- **P**olymorphic **R**egion **P**rediction (PRP)
- Novel ML techniques for predicting genomic segments

Limitations of the Technique



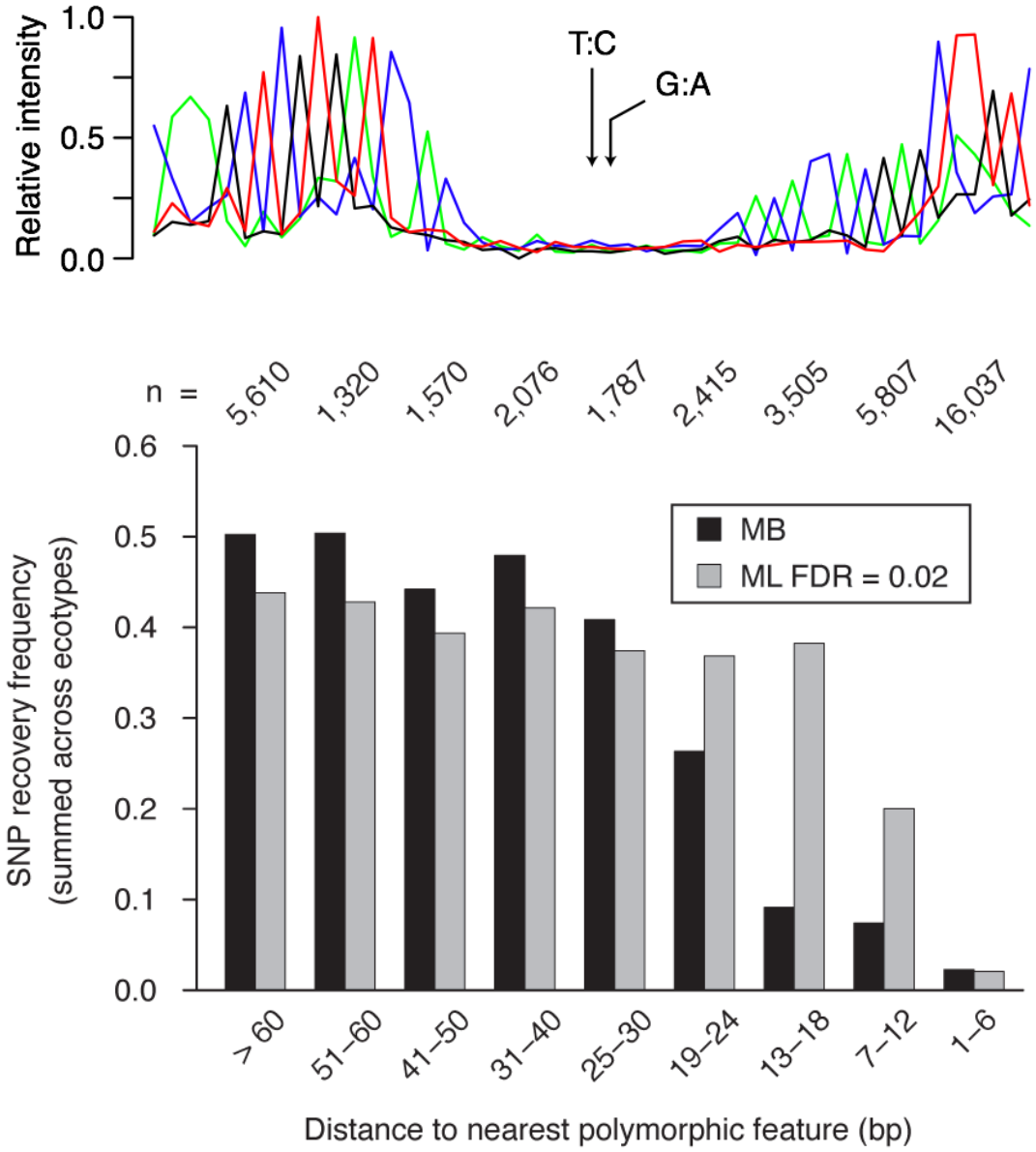
Results

- Performance drops, when other SNPs are in vicinity
- Least predicted SNPs in highly polymorphic regions!
- ML more sensitive

New Approach

- **P**olymorphic **R**egion **P**rediction (PRP)
- Novel ML techniques for predicting genomic segments

Limitations of the Technique



Results

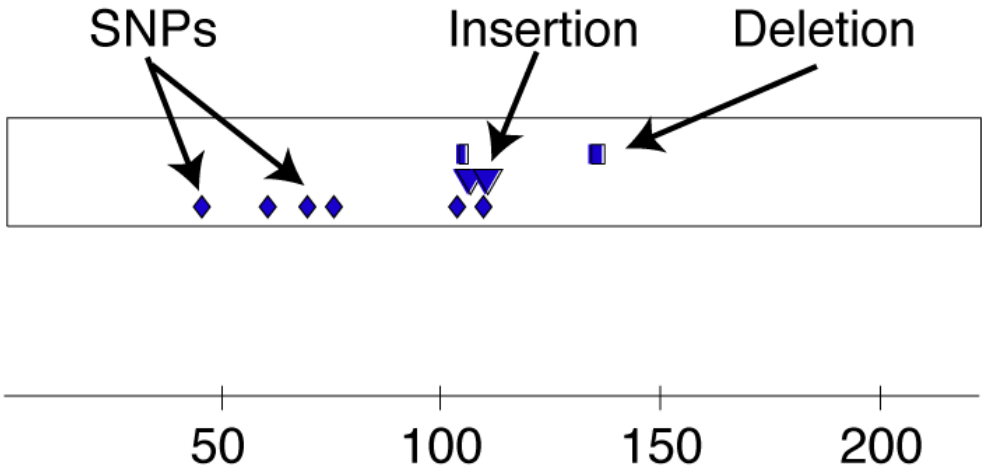
- Performance drops, when other SNPs are in vicinity
- Least predicted SNPs in highly polymorphic regions!
- ML more sensitive

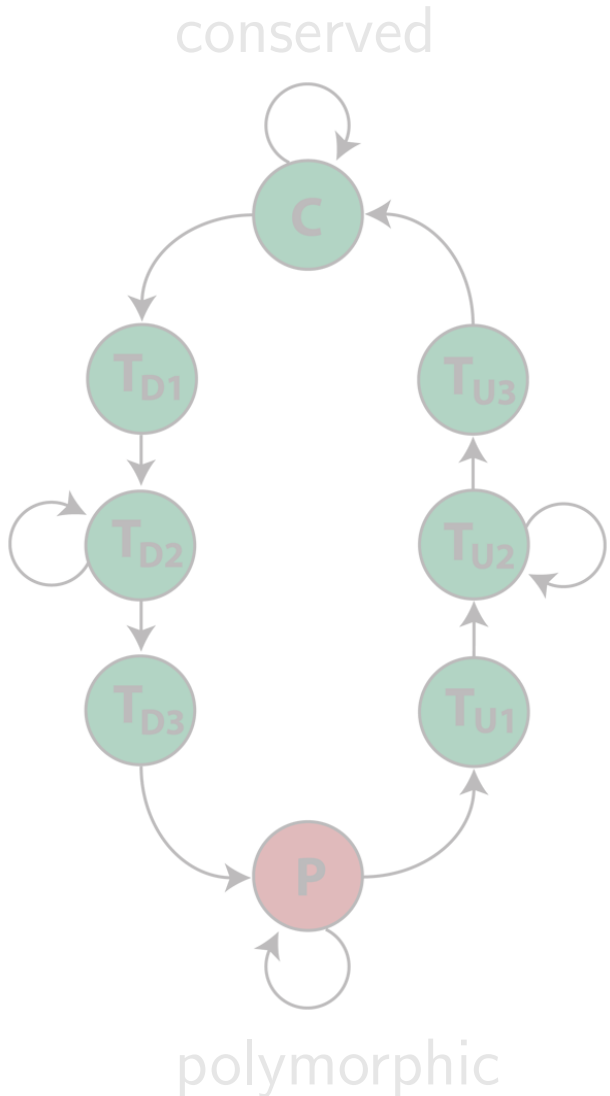
New Approach

- **P**olymorphic **R**egion **P**rediction (PRP)
- Novel ML techniques for predicting genomic segments

Modeling Polymorphic Regions

Example sequence from Bor-4





Segmentation of measurement data into

- Conserved and
- Polymorphic regions (distance <18nt)

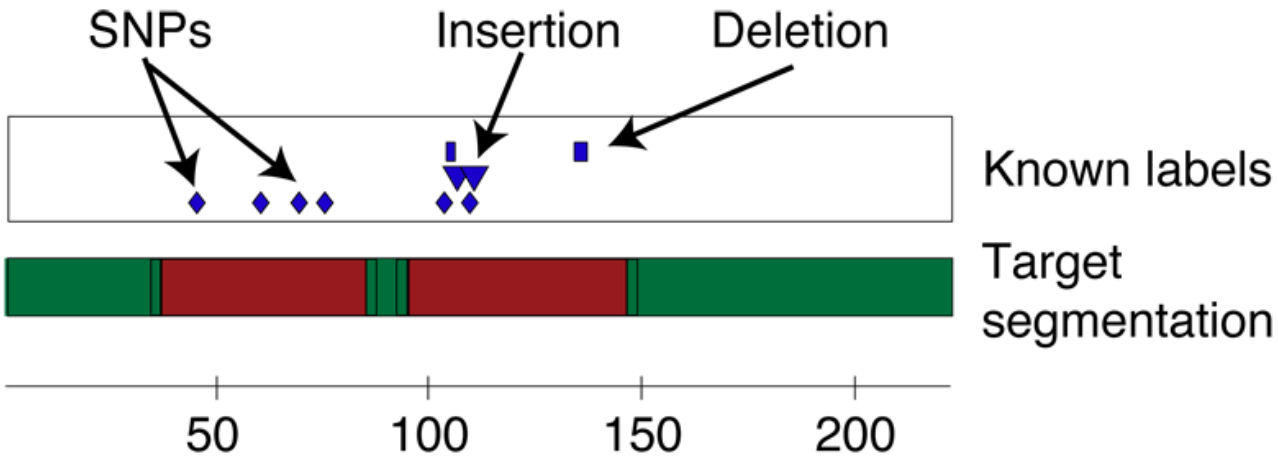
Predict segmentation using state-model

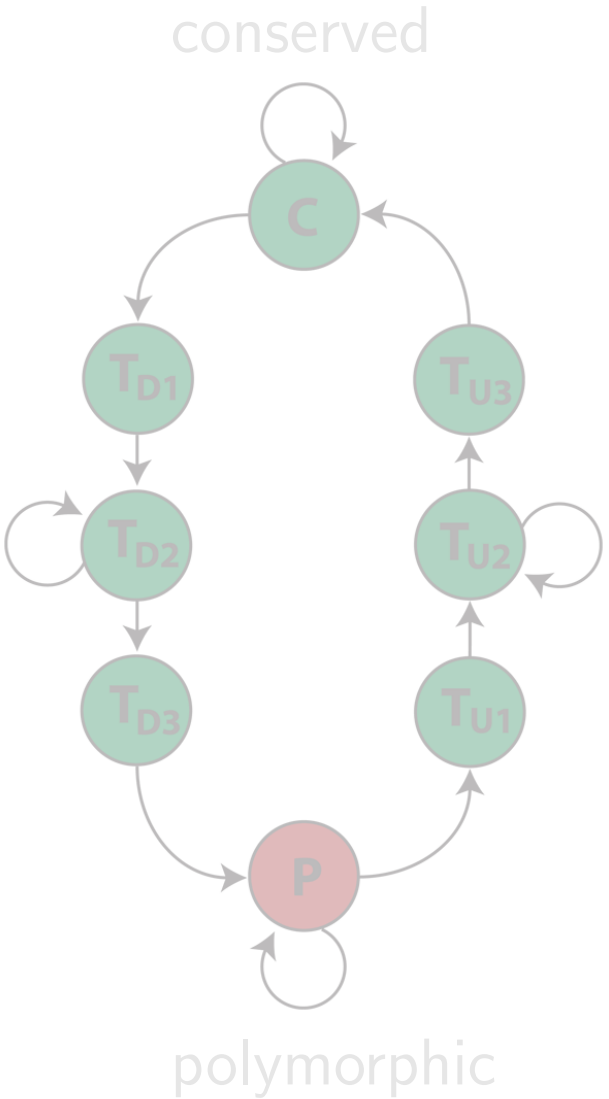
- Similar to Hidden Markov Models
- Discriminative training with HM-SVMs

Zeller et al., 2008.

Modeling Polymorphic Regions

Example sequence from Bor-4





Segmentation of measurement data into

- Conserved and
- Polymorphic regions (distance <18nt)

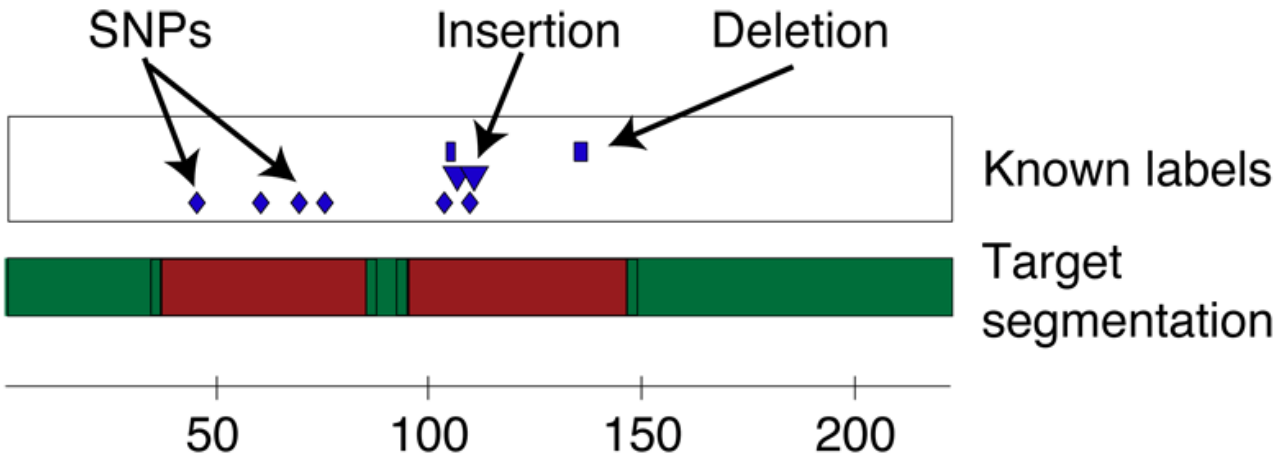
Predict segmentation using state-model

- Similar to Hidden Markov Models
- Discriminative training with HM-SVMs

Zeller et al., 2008.

Modeling Polymorphic Regions

Example sequence from Bor-4



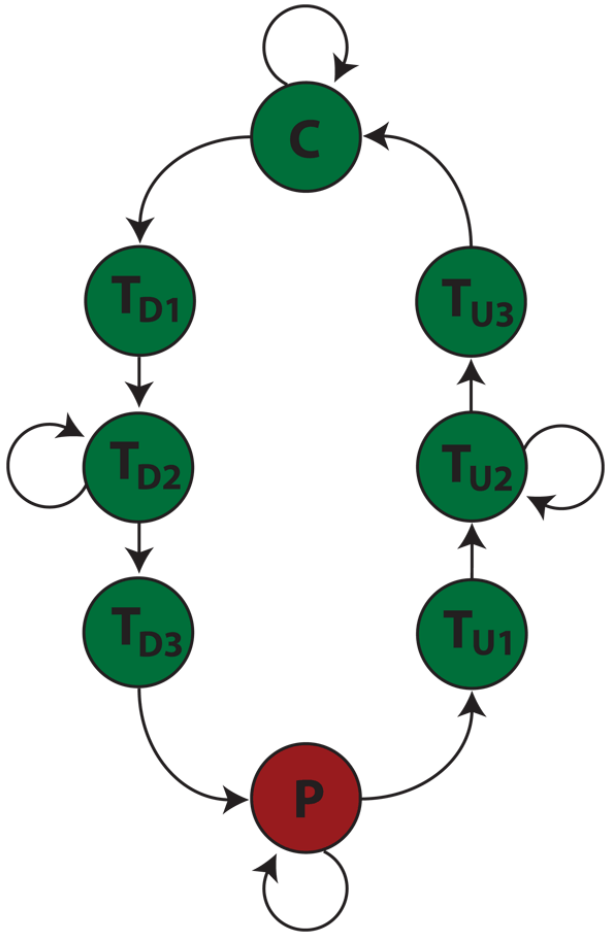
Segmentation of measurement data into

- Conserved and
- Polymorphic regions (distance <18nt)

Predict segmentation using state-model

- Similar to Hidden Markov Models
- Discriminative training with HM-SVMs

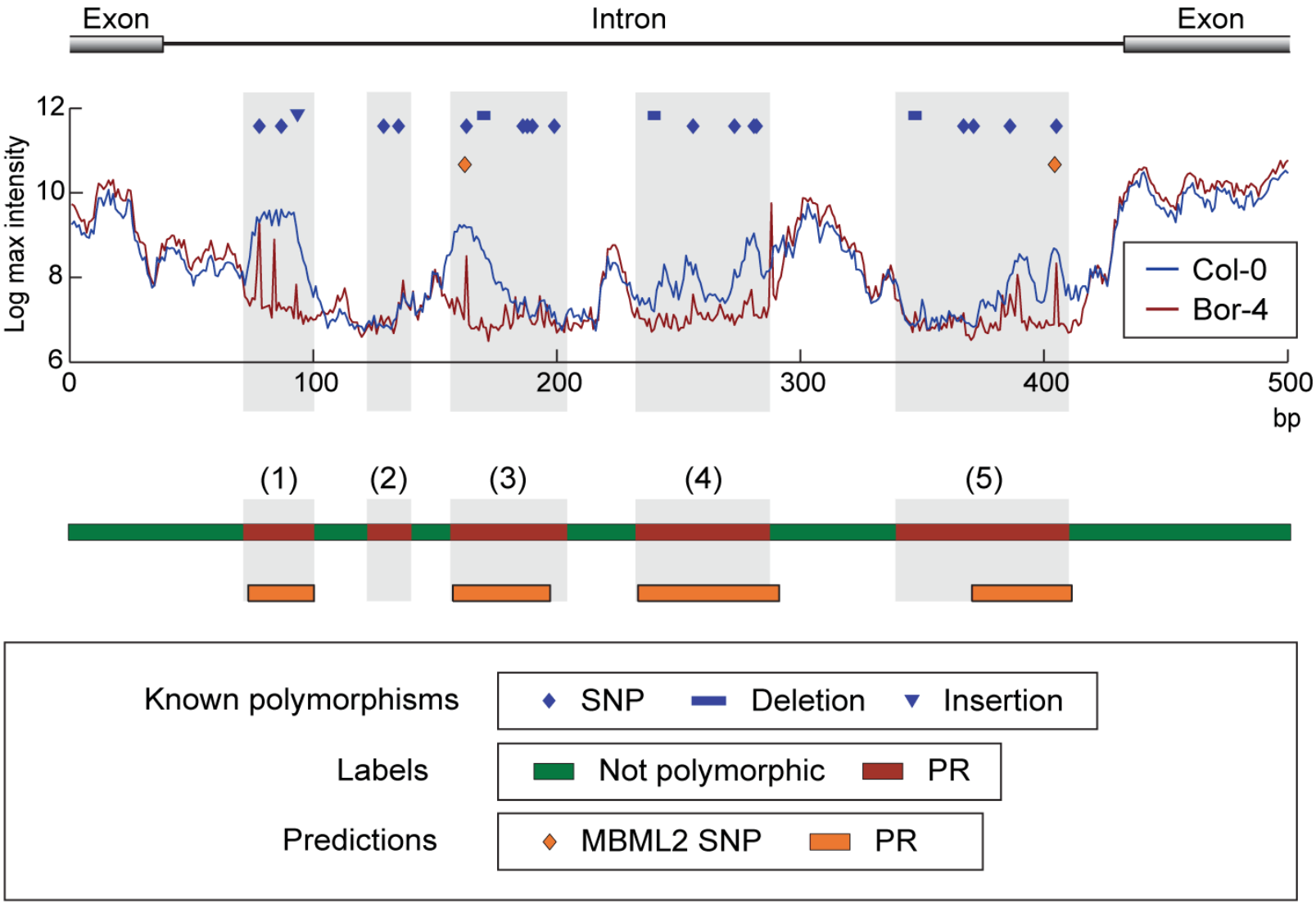
conserved



polymorphic

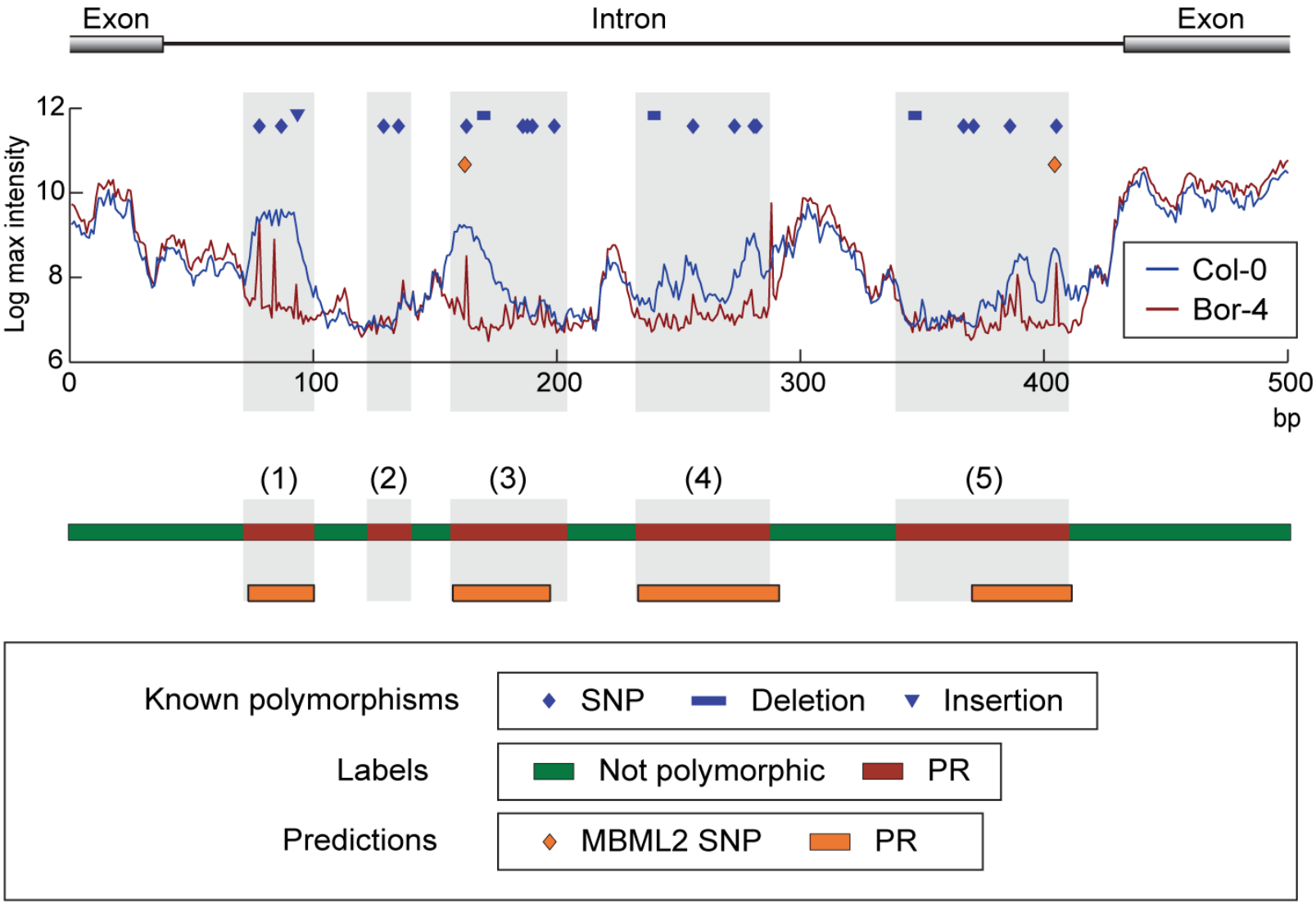
Zeller et al., 2008.

Example



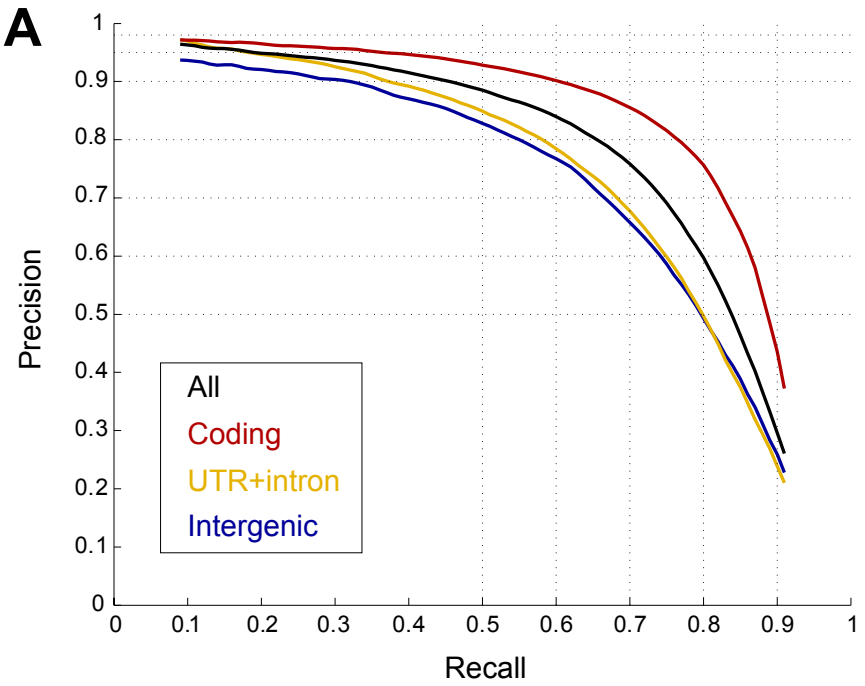
Count prediction as correct if it overlaps by 75%

Example

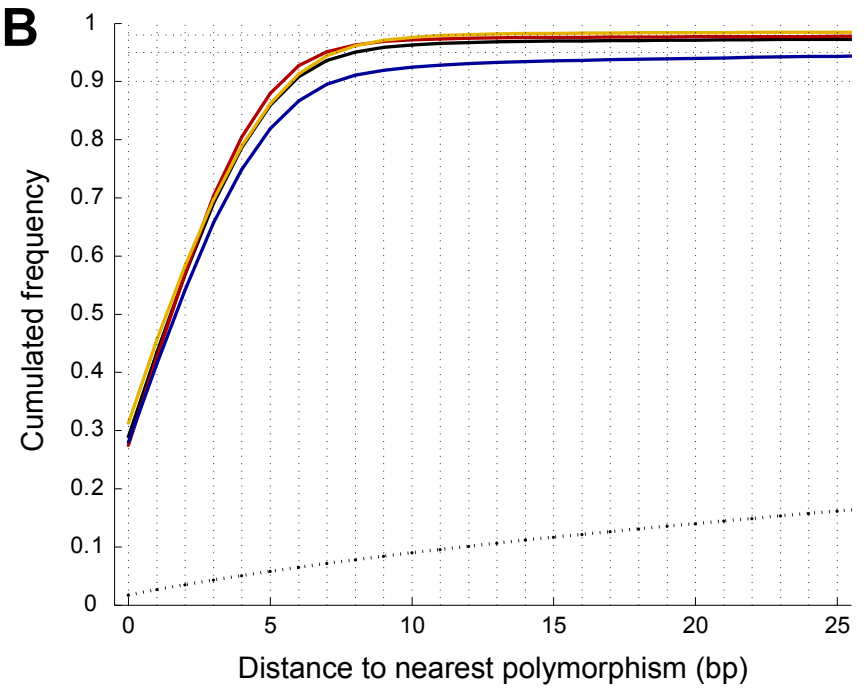


Count prediction as correct if it overlaps by 75%

Detection Performance

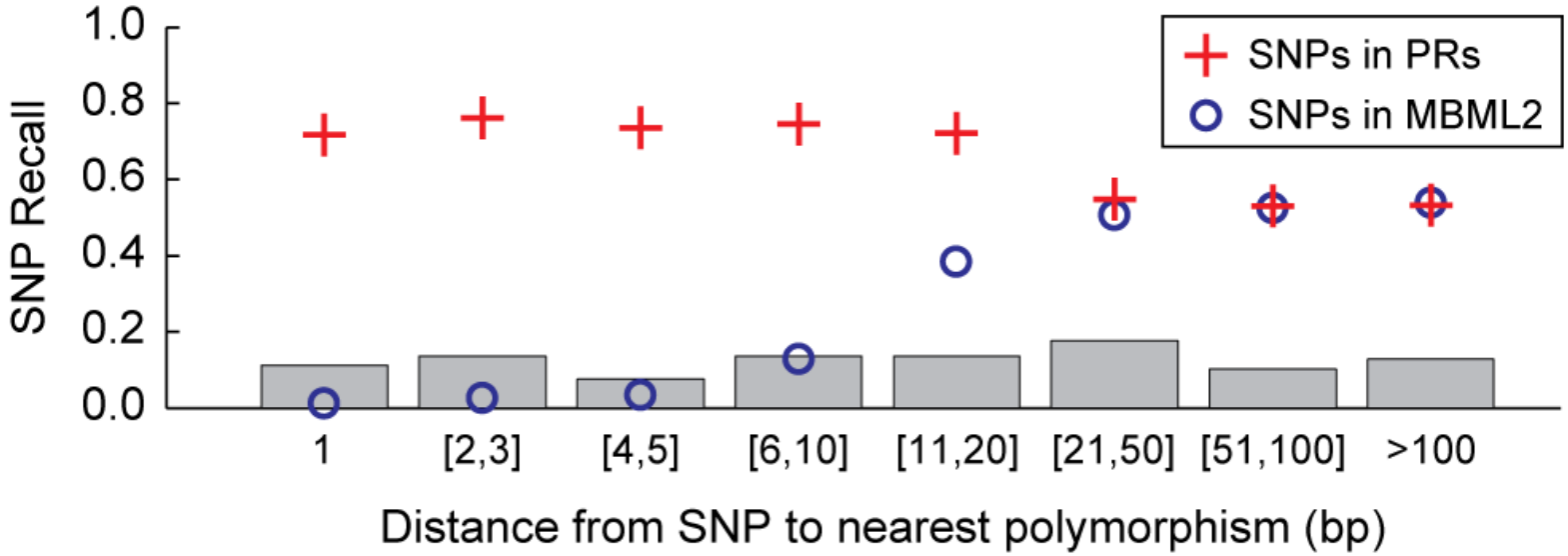


A. Precision-recall curves averaged over cross-validation test subsets for different sequence types



B. Precision at the nucleotide level as calculated for each position within a prediction.

Complementing SNP Calls



Fraction of called/covered polymorphisms (test set):

	SNP calling (MB+ML)	Region predictor
clustered SNPs	∼12%	∼74%
all SNPs (exclusive)	∼32% (8%)	∼64% (42%)
Deletions (per base)		∼53%
Insertions		∼39%

Predicted Effects on Gene Products

SNPs may interfere with signal recognition

- Transcription start & stop
- Splice sites
- Translation start & stop
- Regulatory elements

Consequences

- Modification of consensus sequence (e.g. splice sites)
  - Permanent change

  $\Rightarrow$ Easy to check
- Weaker or stronger recognition of signals
  - Different expression levels
  - (Alternative) splicing pattern may change

  $\Rightarrow$ Needs ab initio gene finding

Predicted Effects on Gene Products

SNPs may interfere with signal recognition

- Transcription start & stop
- Splice sites
- Translation start & stop
- Regulatory elements

Consequences

- Modification of consensus sequence (e.g. splice sites)
  - Permanent change

  $\Rightarrow$ Easy to check
- Weaker or stronger recognition of signals
  - Different expression levels
  - (Alternative) splicing pattern may change

  $\Rightarrow$ Needs ab initio gene finding

Effects on Genes

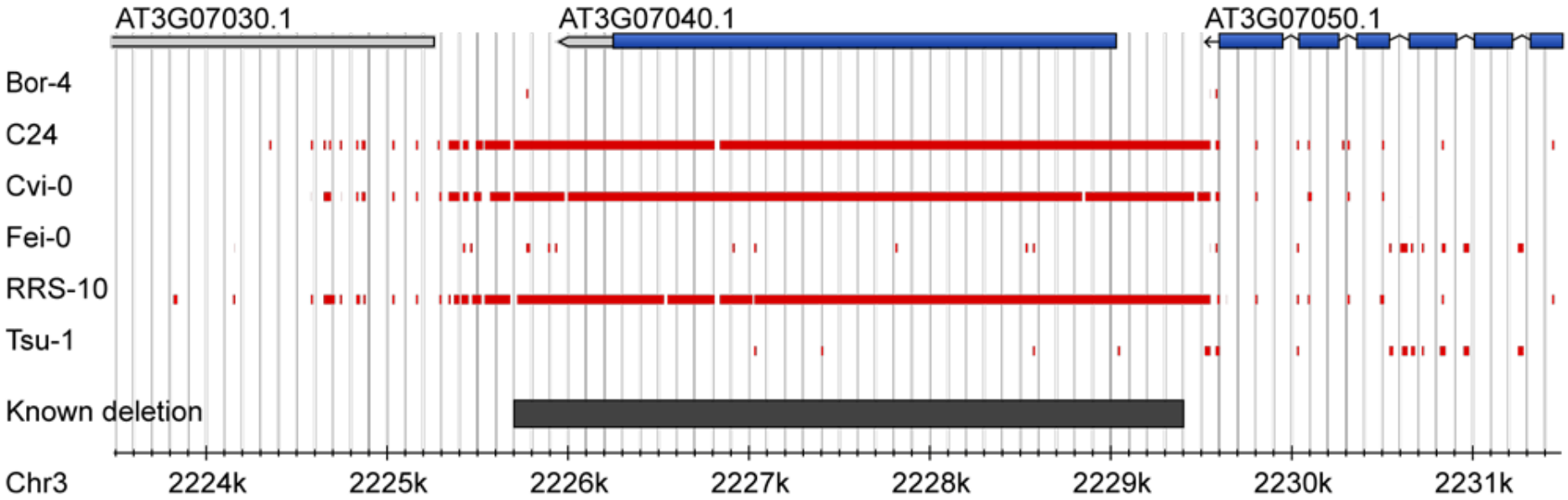
SNPs (MB+ML) leading to ...

- 109,976 amino acid changes
- 1,227 premature stops, 198 removed stops
- 156 initiation methionine changes
- 435 splice site changes
- Major changes in 573 genes validated by dideoxy sequencing

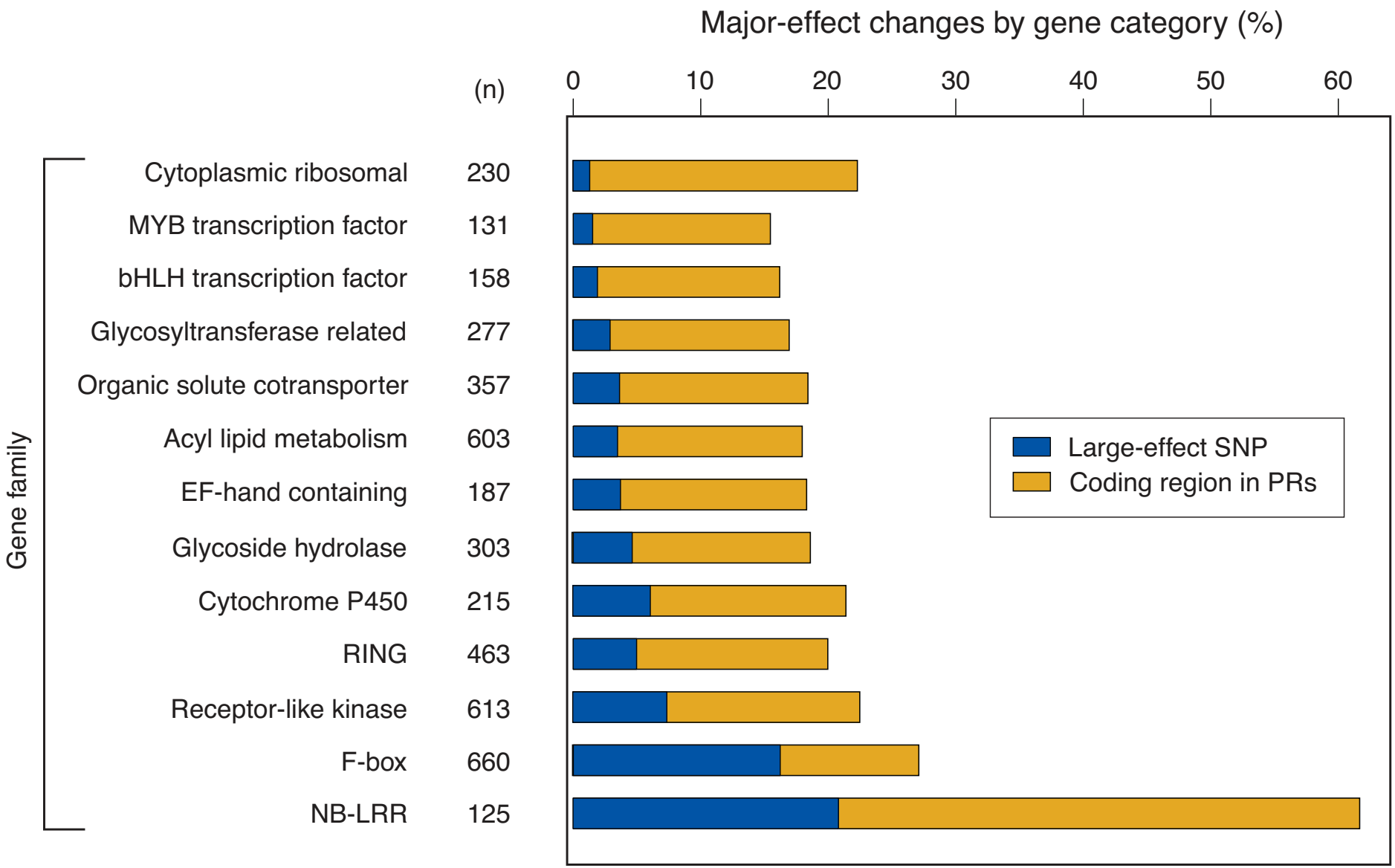
Effects on Genes

SNPs (MB+ML) leading to ...

- 109,976 amino acid changes
- 1,227 premature stops, 198 removed stops
- 156 initiation methionine changes
- 435 splice site changes
- Major changes in 573 genes validated by dideoxy sequencing



Effects on Genes



Ab initio Gene Finding

Given the DNA, predict resulting mRNA and protein

- Requires very accurate identification of
  - Splice sites, translation & transcription starts & stops
  - Sites of regulation (transcription, splicing, etc.)

  $\Rightarrow$ Accurate large scale ML techniques
- Develop methods to integrate single site predictions

  $\Rightarrow$ Novel learning methods for *structured outputs*

Newly developed system "mSplicer" and "mGene"

- Performance illustrated on *C. elegans*
- Highly accurate gene structure prediction (drastic improvements)
- Prediction of alternative splicing

Rätsch et al., 2004-2007

$\Rightarrow$ **Use system to predict effects of SNPs on splicing**

Ab initio Gene Finding

Given the DNA, predict resulting mRNA and protein

- Requires very accurate identification of
  - Splice sites, translation & transcription starts & stops
  - Sites of regulation (transcription, splicing, etc.)

  $\Rightarrow$ Accurate large scale ML techniques
- Develop methods to integrate single site predictions

  $\Rightarrow$ Novel learning methods for *structured outputs*

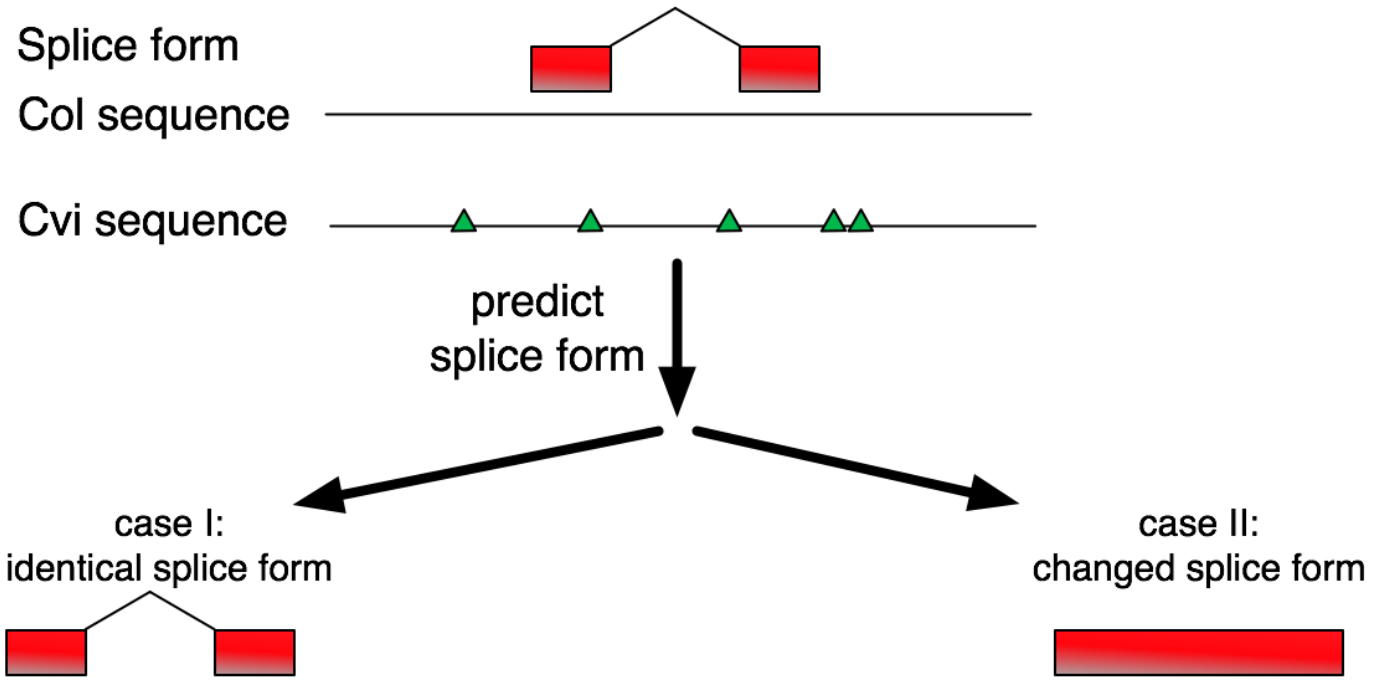
Newly developed system "mSplicer" and "mGene"

- Performance illustrated on *C. elegans*
- Highly accurate gene structure prediction (drastic improvements)
- Prediction of alternative splicing

Rätsch et al., 2004-2007

$\Rightarrow$ **Use system to predict effects of SNPs on splicing**

Predicted Effects by Gene Finding (Preliminary)



16664 genes

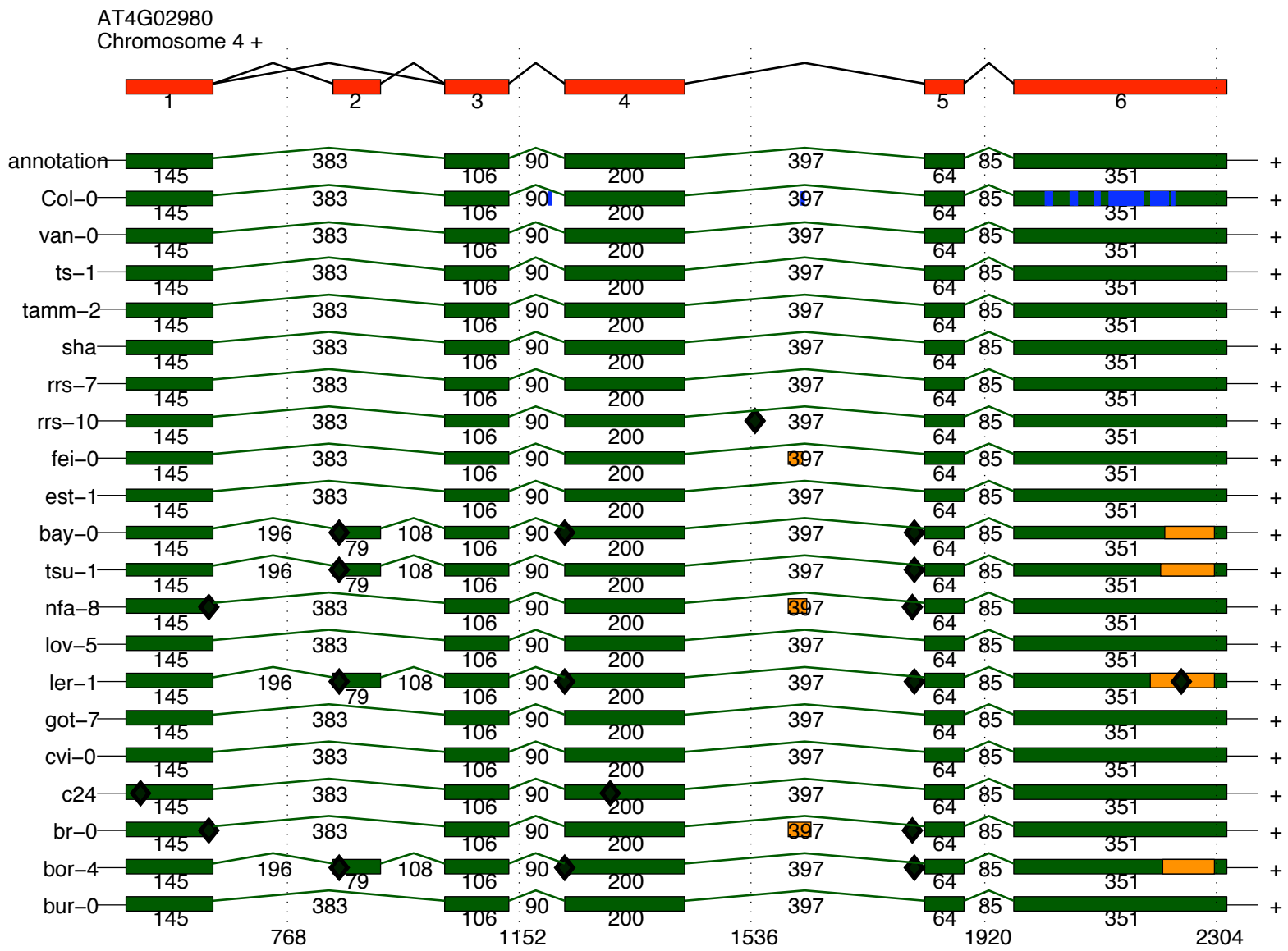
~92 %

1366 genes

~8 %

intron retentions:	**771**
single exon skips:	**380**
alt. 5 prime sites:	**325**
alt. 3 prime sites:	**477**
multiple exon skips:	**23**

Example of Predicted Splice Form Change



Conclusions

- Created inventory for polymorphisms in *Arabidopsis thaliana*
  - Highly polymorphic: ≈0.5% in SNPs, ≈25% in PRPs
- New method for SNP calling; more accurate than Perlegen's
- Accurate polymorphic region predictions
  - Important for further analysis (e.g. dideoxy sequencing)
- Large number of major effect changes
  - Overrepresented in R genes, F-box genes, Receptor-like kinases
- More predicted changes by *ab initio* gene finding
- Application to other genomes (rice, mouse, human?)
- Study variations using mRNA tiling arrays
  - Expression levels, splicing, ...

Conclusions

- Created inventory for polymorphisms in *Arabidopsis thaliana*
  - Highly polymorphic: ≈0.5% in SNPs, ≈25% in PRPs
- New method for SNP calling; more accurate than Perlegen's
- Accurate polymorphic region predictions
  - Important for further analysis (e.g. dideoxy sequencing)
- Large number of major effect changes
  - Overrepresented in R genes, F-box genes, Receptor-like kinases
- More predicted changes by *ab initio* gene finding
- Application to other genomes (rice, mouse, human?)
- Study variations using mRNA tiling arrays
  - Expression levels, splicing, ...

Conclusions

- Created inventory for polymorphisms in *Arabidopsis thaliana*
  - Highly polymorphic: ≈0.5% in SNPs, ≈25% in PRPs
- New method for SNP calling; more accurate than Perlegen's
- Accurate polymorphic region predictions
  - Important for further analysis (e.g. dideoxy sequencing)
- Large number of major effect changes
  - Overrepresented in R genes, F-box genes, Receptor-like kinases
- More predicted changes by *ab initio* gene finding
- Application to other genomes (rice, mouse, human?)
- Study variations using mRNA tiling arrays
  - Expression levels, splicing, ...

Conclusions

- Created inventory for polymorphisms in *Arabidopsis thaliana*
  - Highly polymorphic: ≈0.5% in SNPs, ≈25% in PRPs
- New method for SNP calling; more accurate than Perlegen's
- Accurate polymorphic region predictions
  - Important for further analysis (e.g. dideoxy sequencing)
- Large number of major effect changes
  - Overrepresented in R genes, F-box genes, Receptor-like kinases
- More predicted changes by *ab initio* gene finding
- Application to other genomes (rice, mouse, human?)
- Study variations using mRNA tiling arrays
  - Expression levels, splicing, ...

Acknowledgments

Friedrich Miescher Laboratory
of the Max Planck Society

Friedrich Miescher Laboratory
- Gabriele Schweikert
- Georg Zeller

The Salk Institute, CA
- Paul Shinn
- Joseph Ecker

MPI for Biological Cybernetics
- Bernhard Schölkopf

Perlegen Sciences, CA
- Kelly Frazer
- David Hinds

MPI for Developmental Biology
- Richard Clark
- Stephan Ossowski
- Norman Warthmann
- Korbinian Schneeberger
- Detlef Weigel

University of Southern California
- Christopher Toomajian
- Tina Hu
- Magnus Nordborg

ZBIT, University of Tübingen
- Daniel Huson

References I

Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden Markov support vector machines. In *Proc. 20th Int. Conf. Mach. Learn.*, pages 3–10, 2003.

A. Benhur, C.S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch. Support vector machines and kernels for computational biology. *submitted*, 2008.

F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch. Optimal spliced alignments of short sequence reads. In *Proc. ECCB'08*, 2008. in press.

R. Clark, G. Schweikert, C. Toomajian, S. Ossowski, G. Zeller, P. Shinn, N. Warthmann, T. Hu, G. Fu, D. Hinds, H. Chen, K. Frazer, D. Huson, B. Schölkopf, M. Nordborg, G. Rätsch, Joseph Ecker, and Detlef Weigel. Common sequence polymorphisms shaping genetic diversity in arabidopsis thaliana. *Science*, 317(5836):338–342, 2007. URL `http://www.sciencemag.org/cgi/content/abstract/317/5836/338`.

References II

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, and Cox DR. Free full text whole-genome patterns of common dna variation in three human populations. *Science*, 307(5712):1072–9, February 2005.

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall JD, Wang J, Zhao K, Kalbfleisch T, Schulz V, Kreitman M, and Bergelson J. The pattern of polymorphism in arabidopsis thaliana. *PLoS Biology*, 3(7):e196, 2005.

G. Rätsch and S. Sonnenburg. Large scale semi hidden markov svms. In *Advances in Neural Information Processing Systems 19*, Cambridge, MA, 2006. MIT Press.

References III

G. Rätsch, S. Sonnenburg, J. Srinivasan, H. Witte, K.R. Müller, R. Sommer, and B. Schölkopf. Improving the caenorhabditis elegans genome annotation using machine learning. *PLoS Comput Biol*, 3(2): e20, 2007.

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

Uta Schulze, Bettina Hepp, Cheng Soon Ong, and Gunnar Rätsch. Palma: mrna to genome alignments using large margin algorithms. *Bioinformatics*, 2007.

S. Sonnenburg and et al. The need for open source software in machine learning. *Journal of Machine Learning Research*, 2007. submitted.

S. Sonnenburg, G. Rätsch, and K. Rieck. *Large Scale Kernel Machines*, chapter Large Scale Learning with String Kernels. MIT Press, 2007a.

References IV

S. Sonnenburg, G. Schweikert, P. Philips, J. Behr, and G. Rätsch. Accurate splice recognition with svms. *BMC Bioinformatics*, August 2007b. accepted.

S. Sonnenburg, A. Zien, P. Philips, and G. Rätsch. Poims: Positional oligomer importance matrices - understanding support vector machine based signal detectors. *Bioinformatics*, 2008. in press.

Sören Sonnenburg, Alexander Zien, and Gunnar Rätsch. ARTS: Accurate Recognition of Transcription Starts in Human. *Bioinformatics*, 22(14): e472–480, 2006.

I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6:1453–1484, 2005. ISSN 1533-7928. URL `http://jmlr.csail.mit.edu/papers/v6/tsochantaridis05a.html`.

References V

G. Zeller, R. Clark, K. Schneeberger, A. Bohlen, D. Weigel, and G. Rätsch. Detecting polymorphic regions in arabidopsis thaliana with resequencing microarrays. *Genome Research*, 2008. in press.