Overview



1. Introduction to Bioinformatics
   - Basic Biology and Central Dogma
   - Typical Data Types
   - Common Analysis Tasks
2. Sequence Analysis (with SVMs)
   - String Kernels
   - Large Scale Data Structures
   - Heterogeneous Data
3. Structured Output Learning
   - Hidden Markov Models & Dynamic Programming
   - Discriminative Approaches (CRFs & HMSVMs)
   - Large Scale Approaches
4. Some Applications
   - Spliced Alignments (PALMA)
   - Gene Finding (mGene)
   - Analysis of Resequencing Arrays (SNPs and Polymorphic regions)

Part IV: Some Applications



Part IV: Some Applications

- Optimal Spliced Alignments (PALMA)
- Discriminative Gene Finding (mGene)
- Analysis of Resequencing Arrays (SNPs & PRs)

NB: Help from Gabriele Schweikert, Georg Zeller & Richard Clark

For more applications see e.g. Noble [2004].

PALMA: Perfect Alignments using Large Margins

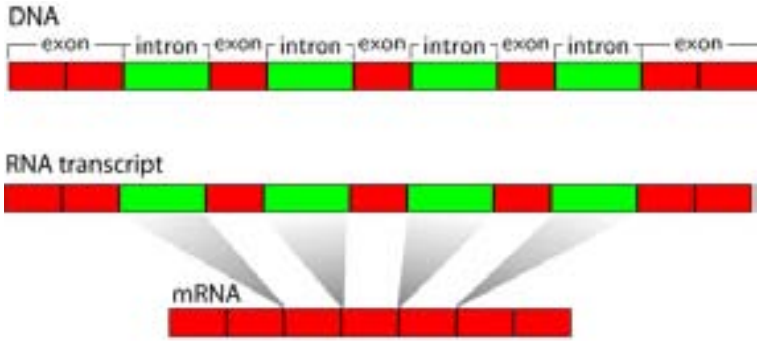
Gunnar Rätsch,[1] Bettina Hepp,[2] Uta Schulze,[1,3] and Cheng Soon Ong[1,4]

[1] Friedrich Miescher Laboratory, Tübingen
[2] Fraunhofer FIRST, Berlin
[3] University of Leipzig
[4] Max Planck Institute for Biological Cybernetics, Tübingen

http://www.fml.mpg.de/raetsch/projects/palma

Motivation & Background

- Abundant experimental data:
  - Expressed Sequence Tags (EST)
  - Full length mRNAs
- Alignment to genomic sequences helps
  - Discovery of new genes,
  - Delineation of exon/intron boundaries,
  - Identification of alternative splice forms,
  - Finding SNPs,
- Problems
  - Repetitive elements, paralogs, pseudo-genes
  - Sequencing errors, polymorphisms
  - Non-canonical splice sites
  - Microexons



Previous Work



More than 10 years of research on spliced alignments

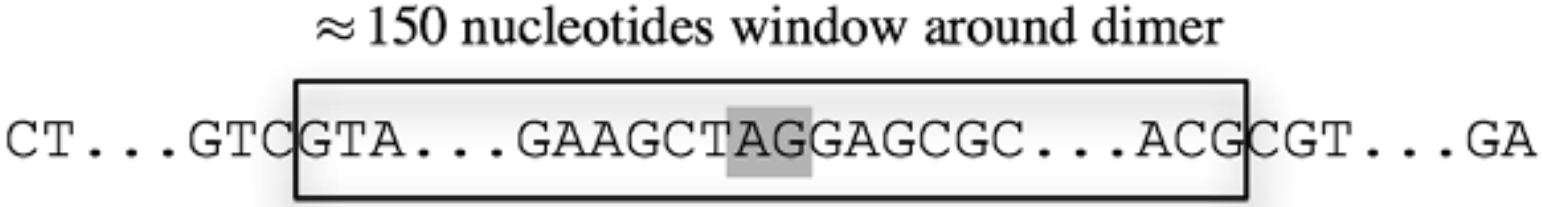
- Greedy algorithms (extend seed words or `Blast` based)
  - `Sim4` [Florea et al., 1998], `Spidey` [Wheelan et al., 2001]
  - `Blat` (prefers `AG`/`GT`) [Kent, 2002]
- `EST_Genome` (DP based, prefers `AG`/`GS`) [Mott, 1997]
- `Exalin` (DP based, `AG`/`GS` only) [Zhang and Gish, 2006]
  - Fixed substitution and gap costs
  - Splice site model (PWMs)

  } Maximum likelihood combination

Why another tool?

- More accurate splice site models (SVM based)
- Intron length model
- More principled combination (based on large margins)

2-Class Splice Site Detection



- True sites: Fixed window around a true splice site
- Decoys sites: Generated by shifting the window



⇒ Very unbalanced problem (1:200)
⇒ Millions of points from EST databases
⇒ Large scale methods necessary
(here: Support Vector Machines)

Alignment Algorithms

- **Input**
  - Two sequences over the alphabet $\{A,\ C,\ G,\ T/U,\ N\}$
    - EST sequence S_E of length m
    - DNA sequence S_D of length n
  - Substitution matrix $M : \Sigma \times \Sigma \rightarrow \mathbf{R}$,
    where $\Sigma := \{A,\ C,\ G,\ T,\ N, -\}$
- **Output**
  - Sequence alignment $\mathcal{A}$
    - Sequence of pairs, i.e. $\mathcal{A} = (a_r, b_r)_{r=1,\ldots,R},\ a_r, b_r \in \Sigma$
    - $R \leq m + n$ depends on the alignment
  - Alignment that maximizes the alignment score

$$s(\mathcal{A}) = \sum_{r=1}^{R} M(a_r, b_r)$$

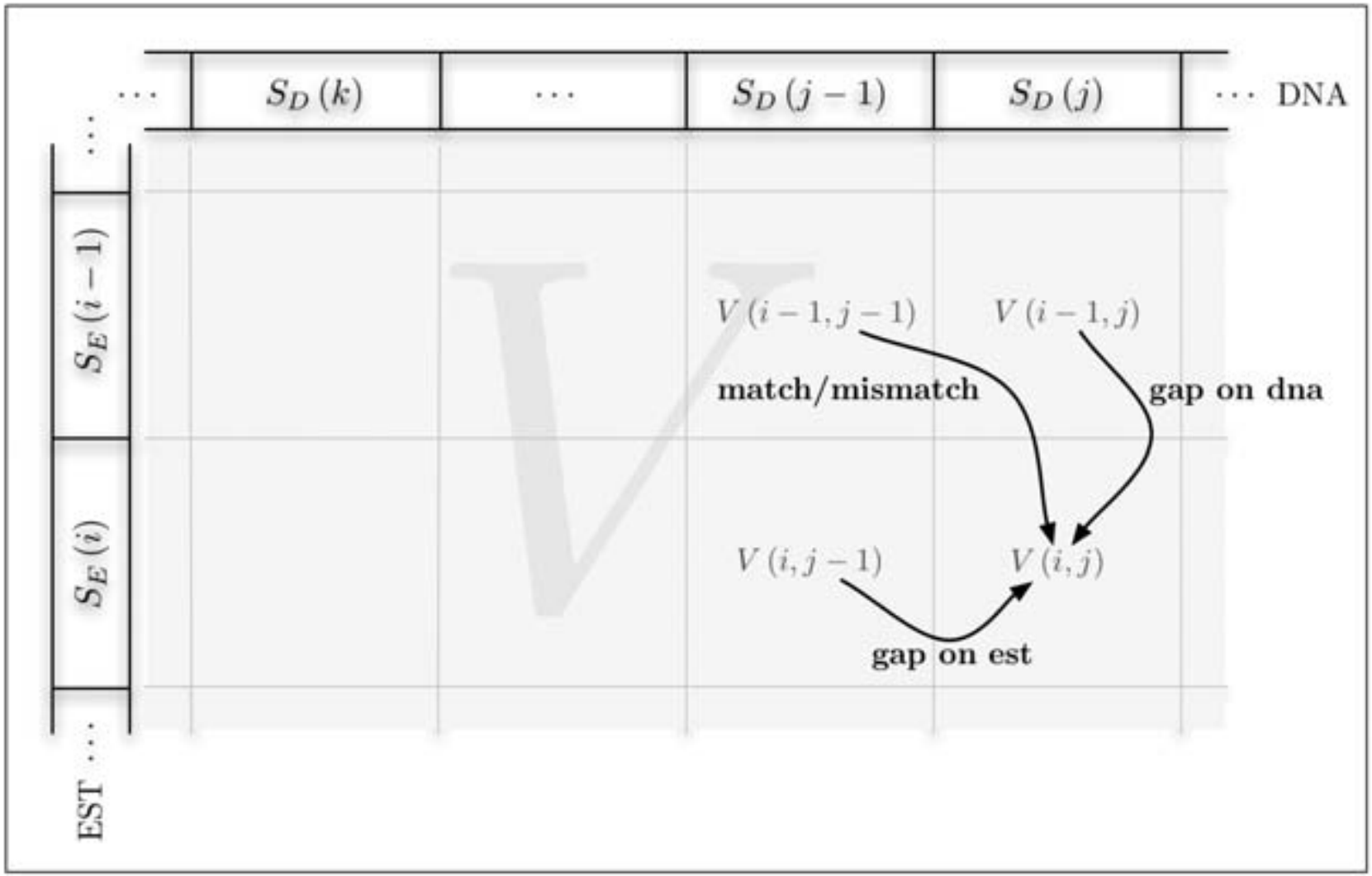
Maximizing the Alignment Score

- **Needleman-Wunsch algorithm**
  - Maximizes alignment score by dynamic programming
  - Fills $m \cdot n$ alignment matrix V:
    - $V(i,0) := 0$ and $V(0,j) := 0$ for all i,j
    - Recursion

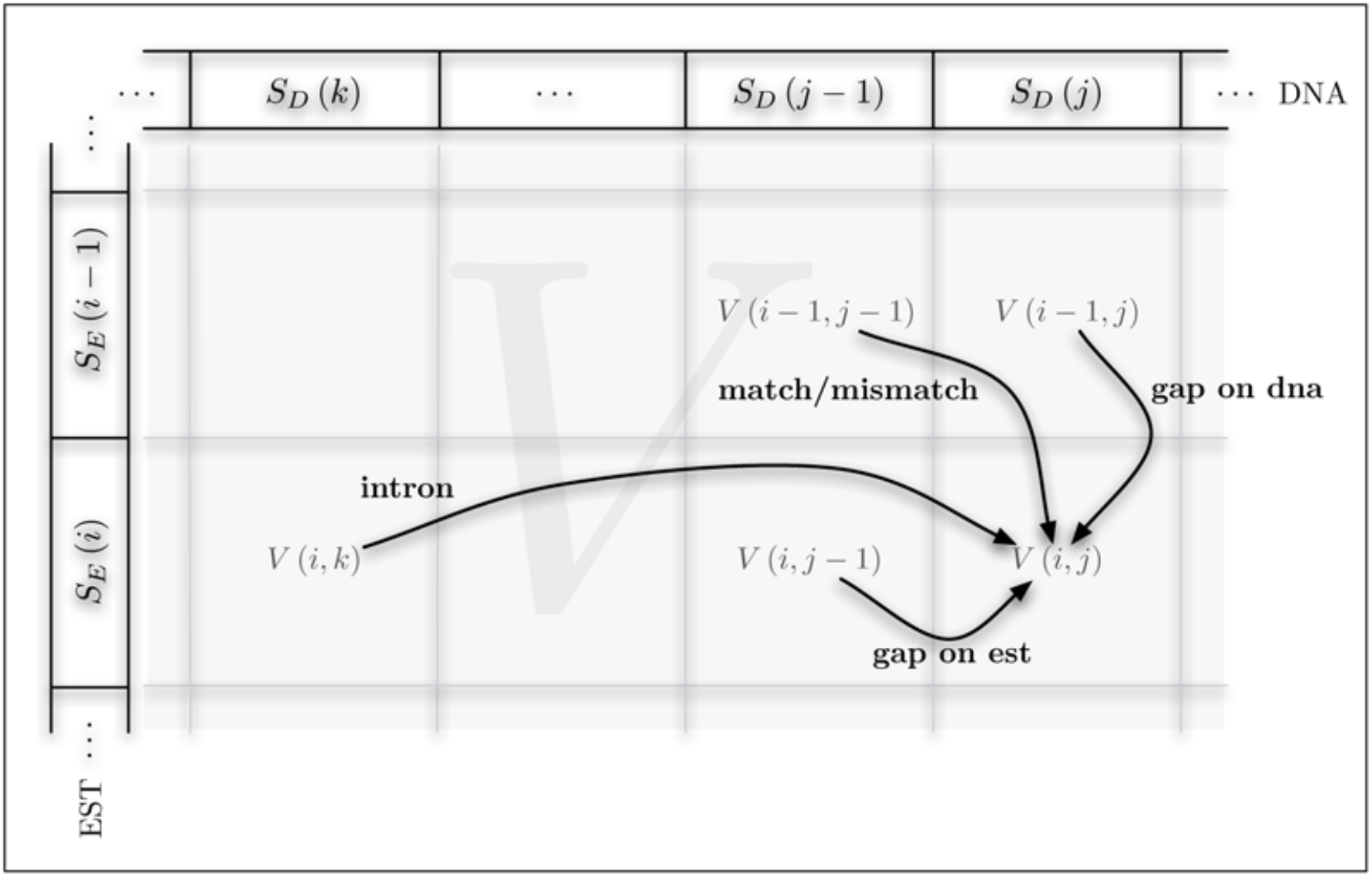
$$V(i,j) = \max \begin{cases} V(i-1,j-1) + M(S_E(i), S_D(j)) \\ V(i-1,j) + M(S_E(i),'-') \\ V(i,j-1) + M('-', S_D(j)) \end{cases}$$

  - Runtime time and space complexity: $\mathcal{O}(m \cdot n)$
- **Problems:**
  - Does not distinguish between gaps and intron
  - How to choose M? No splice site model!
  - Too expensive for alignments against whole genomes

Needleman-Wunsch Algorithm



Needleman-Wunsch Algorithm With Introns



Recursion with Intron Model

- Extended recursion formula ($\forall i = 1 \dots m, j = 1 \dots n$)

$$V(i,j) = \max \begin{cases} V(i-1,j-1) + M(S_E(i), S_D(j)) \\ V(i-1,j) + M(S_E(i), '-') \\ V(i,j-1) + M('-', S_D(j)) \\ \max_{1 \le k \le j-1} (V(i,k) + f_I(k,j)) \end{cases}$$

- For intron score $f_I(k,j)$ consider
  - Splice sites scores s_k^{Don} and s_k^{Acc} (SVM predictions)
    $\Rightarrow$ Contribute $f^{\mathrm{Don}}(s_k^{\mathrm{Don}}) + f^{\mathrm{Acc}}(s_k^{\mathrm{Acc}})$
  - Length of intron
    $\Rightarrow$ Contributes $f^{\mathrm{Len}}(j-k)$
- Unspecified functions $f^{\mathrm{Don}}, f^{\mathrm{Acc}}, f^{\mathrm{Len}}$ as well as M!

Idea: Learn functions on training set with known alignments

Parameterization

- Substitution matrix $M : \Sigma \times \Sigma \rightarrow \mathbb{R}$
- Functions f^{Len}, f^{Acc} and f^{Don}
  - Piecewise linear functions (support points $x_1, \ldots, x_s$):

$$f(x) = \begin{cases} \theta_1 & x \leq x_1 \\ \frac{\theta_i(x_{i+1}-x)+\theta_{i+1}(x-x_i)}{x_{i+1}-x_i} & x_i \leq x \leq x_{i+1} \\ \theta_s & x \geq x_s \end{cases} .$$

  - $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_s)$ parametrizes function
- Let $\boldsymbol{\theta} := (\boldsymbol{\theta}^{\mathrm{Acc}}, \boldsymbol{\theta}^{\mathrm{Don}}, \boldsymbol{\theta}^{\mathrm{Len}}, \boldsymbol{\theta}^{M})$
- Given $\boldsymbol{\theta}$, alignment score $s_{\boldsymbol{\theta}}(\mathcal{A})$ is fully defined

Parameters: Optimization

Idea

Find θ such that for a known alignment $\mathcal{A}^+$

$$s_\theta(\mathcal{A}^+) \gg s_\theta(\mathcal{A}^-)$$

where $\mathcal{A}^- \neq \mathcal{A}^+$ is any wrong alignment

Given N known alignments $\mathcal{A}_i^+, i = 1, \ldots, N$

Solve quadratic optimization problem (QP)

$$\begin{aligned} \min_{\xi \geq \mathbf{0}, \theta} \quad & \frac{1}{N} \sum_{i=1}^{N} \xi_i + \mathbf{P}(\theta) \\ \text{s.t.} \quad & s_\theta(\mathcal{A}_i^+) - s_\theta(\mathcal{A}_i^-) \geq 1 - \xi_i \quad \forall \mathcal{A}_i^- \neq \mathcal{A}_i^+, i = 1, \ldots, N \end{aligned}$$

- ξ_i: Slack-variables to implement a soft-margin
- $\mathbf{P}(\theta)$: Regularizer leading to smooth functions

Problem: Exponentially many constraints

Iterative Algorithm

- Set $\boldsymbol{\theta} := (\boldsymbol{\theta}^{\mathrm{Acc}}, \boldsymbol{\theta}^{\mathrm{Don}}, \boldsymbol{\theta}^{\mathrm{Len}}, \boldsymbol{\theta}^{M})$ randomly, $A_i^- = \emptyset$
- For $t = 1, \ldots$
  - For $i = 1, \ldots, N$
    - Compute (wrong) alignments $\mathcal{A}^-$ based on $\boldsymbol{\theta}$
    - If $\mathcal{A}^- \neq \mathcal{A}_i^+$, then $A_i^- := A_i^- \cup \{\mathcal{A}^-\}$
  - Obtain new parameters $\boldsymbol{\theta}$ by solving the restricted QP

$$\begin{aligned} \min_{\boldsymbol{\xi} \geq \mathbf{0}, \boldsymbol{\theta}} \quad & \frac{1}{N} \sum_{i=1}^{N} \xi_i + \mathbf{P}(\boldsymbol{\theta}) \\ \text{s.t.} \quad & s_{\boldsymbol{\theta}}(\mathcal{A}_i^+) - s_{\boldsymbol{\theta}}(\mathcal{A}^-) \geq 1 - \xi_i \quad \forall \mathcal{A}^- \in A_i^-, i = 1, \ldots, N \end{aligned}$$

- Only need to solve small optimization problems!
- Guaranteed convergence!

Microexon Simulation Study

Artificial Data

- Consider EST confirmed exon triples (*C. elegans*)
- Shorten middle exon in central region (EST & DNA)



(Found by `blat` & strict filtering)

Microexon generated
$\Rightarrow$ Splice sites still intact

- Generate insertions/deletions/mutations in artificial EST ($\sigma = 0\%, 1\%, 2\%, 10\%, 20\%, 50\%$)
- Train `PALMA` on 4608 exon triples ($\approx 2h$)
- Test `blat`, `sim4`, `exalin` and `PALMA` on 4358 tripples
- Correct only if all boundaries are correctly predicted

Microexon Simulation Study

Artificial Data

- Consider EST confirmed exon triples (*C. elegans*)
- Shorten middle exon in central region (EST & DNA)



(Found by `blat` & strict filtering)

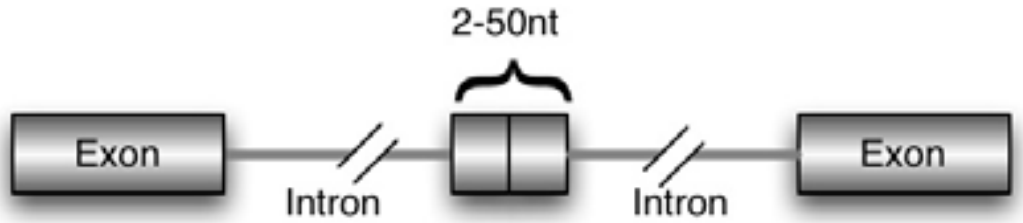
Microexon generated
$\Rightarrow$ Splice sites still intact

- Generate insertions/deletions/mutations in artificial EST ($\sigma = 0\%, 1\%, 2\%, 10\%, 20\%, 50\%$)
- Train `PALMA` on 4608 exon triples ($\approx 2h$)
- Test `blat`, `sim4`, `exalin` and `PALMA` on 4358 tripples
- Correct only if all boundaries are correctly predicted

Microexon Simulation Study

Artificial Data

- Consider EST confirmed exon triples (*C. elegans*)
- Shorten middle exon in central region (EST & DNA)



(Found by `blat` & strict filtering)

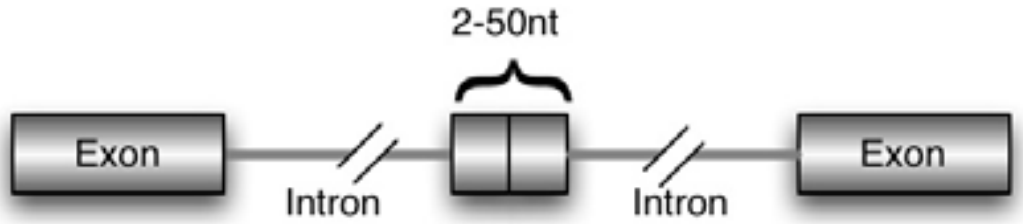
Microexon generated
$\Rightarrow$ Splice sites still intact

- Generate insertions/deletions/mutations in artificial EST ($\sigma = 0\%, 1\%, 2\%, 10\%, 20\%, 50\%$)
- Train `PALMA` on 4608 exon triples ($\approx 2h$)
- Test `blat`, `sim4`, `exalin` and `PALMA` on 4358 tripples
- Correct only if all boundaries are correctly predicted

Microexon Simulation Study

Artificial Data

- Consider EST confirmed exon triples (*C. elegans*)
- Shorten middle exon in central region (EST & DNA)

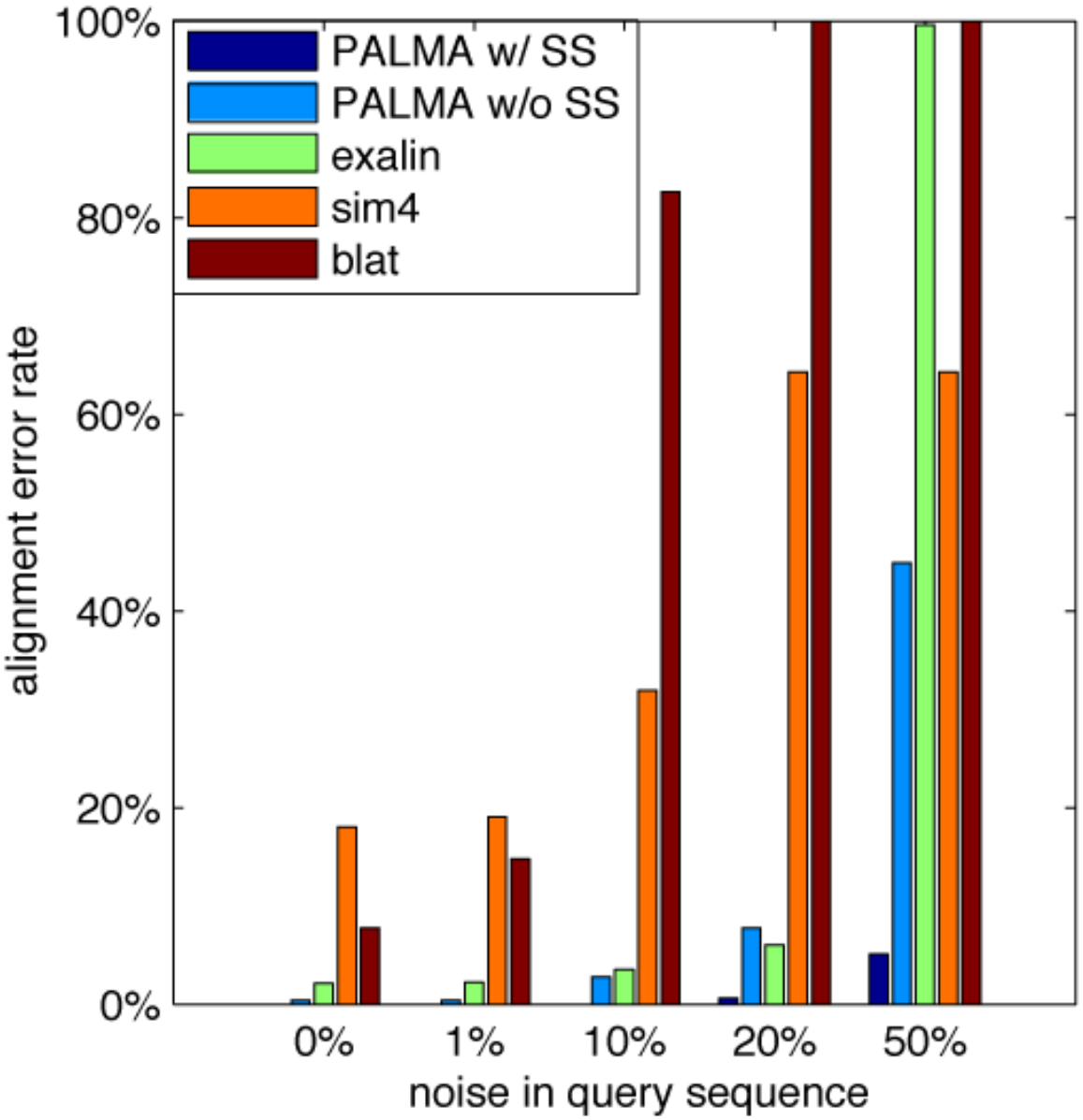


(Found by `blat` & strict filtering)

Microexon generated
$\Rightarrow$ Splice sites still intact

- Generate insertions/deletions/mutations in artificial EST
  ($\sigma = 0\%, 1\%, 2\%, 10\%, 20\%, 50\%$)
- Train `PALMA` on 4608 exon triples ($\approx 2h$)
- Test `blat`, `sim4`, `exalin` and `PALMA` on 4358 tripples
- Correct only if all boundaries are correctly predicted

Results (*C. elegans*)



- Error rate increases drastically for `blat` and `sim4`
- `Exalin` performs quite well, but always $> 3\%$ wrong
- `PALMA` is always correct for $\sigma \leq 10\%$
- `PALMA w/o SS` only slightly worse

Conclusion

- New alignment algorithm for *perfect* alignments of mRNA and DNA
- Exploits very accurate SVM-based splice site predictions
- New idea of combining different sources of information:
  - Similarity, splice site scores and intron lengths
  - Large margin based iterative algorithm
  - Guaranteed convergence
- Significantly reduced error rates on test data (short exons/much noise)
- Better detection of microexons & altern. spliced exons
- Current work: Reduce computational complexity
- Source code (Python/C++, GPL) and data available at `http://www.fml.mpg.de/raetsch/projects/palma`

Discriminative Gene Finding

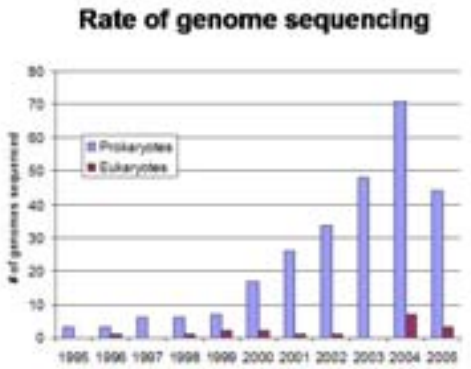
Gabriele Schweikert,[1,2,3] Georg Zeller,[1,3], Alexander Zien,[1,4] Cheng Soon Ong,[1,2] Fabio de Bona,[1] Petra Philips,[1] Sören Sonnenburg,[4] Gunnar Rätsch[1]

[1] Friedrich Miescher Laboratory, Tübingen
[2] Max Planck Institute for Biological Cybernetics, Tübingen
[3] Max Planck Institute for Developmental Biology, Tübingen
[4] Fraunhofer FIRST, Berlin,

http://www.fml.mpg.de/raetsch/projects/mgene

Introduction

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

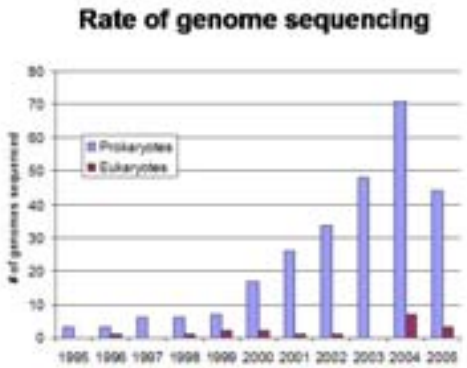


Example: Nematodes:

- *C. elegans* sequenced in 1998, *C. briggsae* in 2002
- Three more nematode genomes are recently being sequenced (*C. remanei*, *C. japonica*, *C.* sp. CB5161)
- 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

Introduction



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

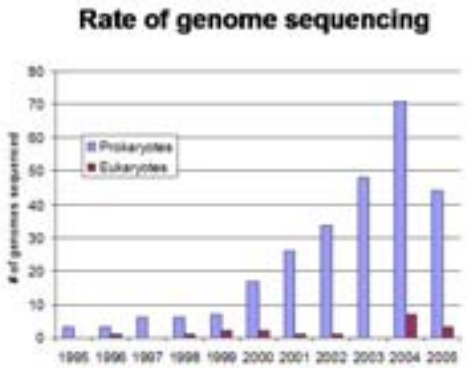
Example: Nematodes:



- *C. elegans* sequenced in 1998, *C. briggsae* in 2002
- Three more nematode genomes are recently being sequenced (*C. remanei*, *C. japonica*, *C.* sp. CB5161)
- 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

Introduction



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

Example: Nematodes:

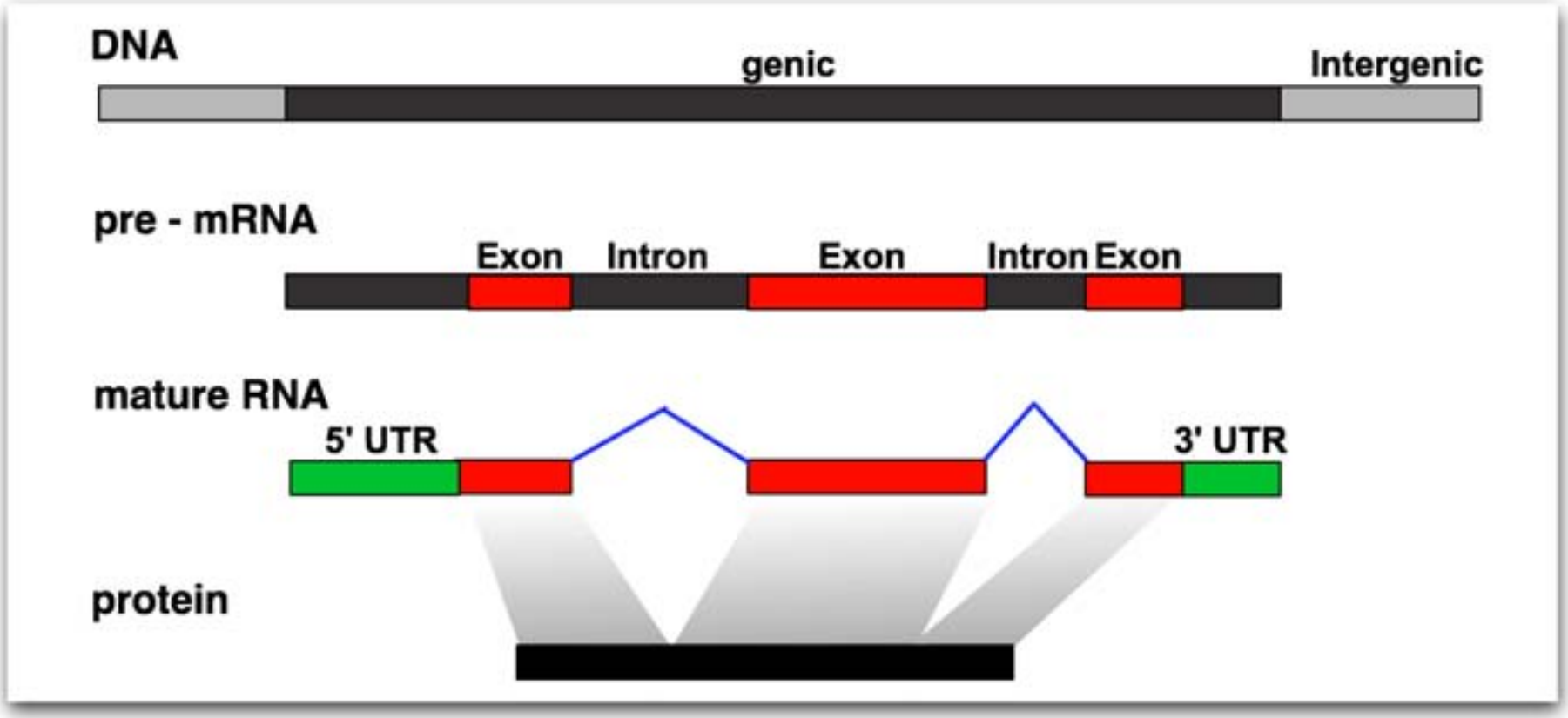


- *C. elegans* sequenced in 1998, *C. briggsae* in 2002
- Three more nematode genomes are recently being sequenced (*C. remanei*, *C. japonica*, *C.* sp. CB5161)
- Genomes need to be annotated! Which gene finder?
- nGASP competition:
  - Controlled competition conditions
  - Identify the most accurate method

Our contribution: A novel accurate gene finding system

Ab initio Gene Prediction

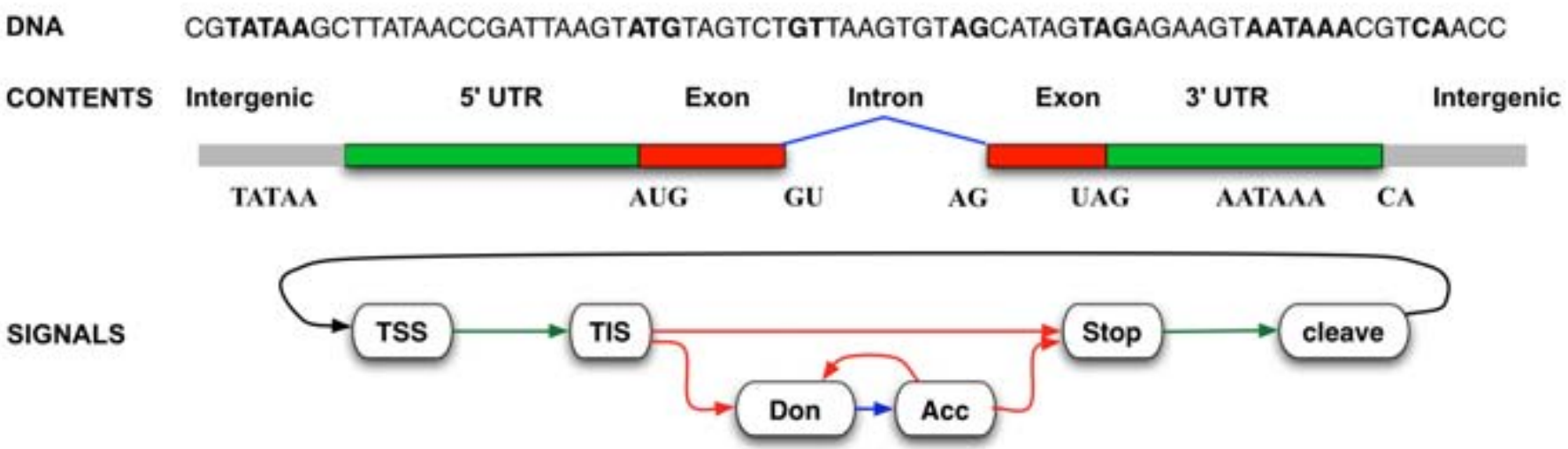
A label sequence learning problem



- Given a DNA sequence $\mathbf{x} \in \{\text{'A','C','G','T'}\}^L$
- Find the correct sequence of labels $\mathbf{y} = \upsilon_1 \upsilon_2 \ldots \upsilon_L$
  ($\upsilon_i \in \mathcal{Y} = \{\text{'intergenic', 'exon', 'intron', } \ldots\}$)

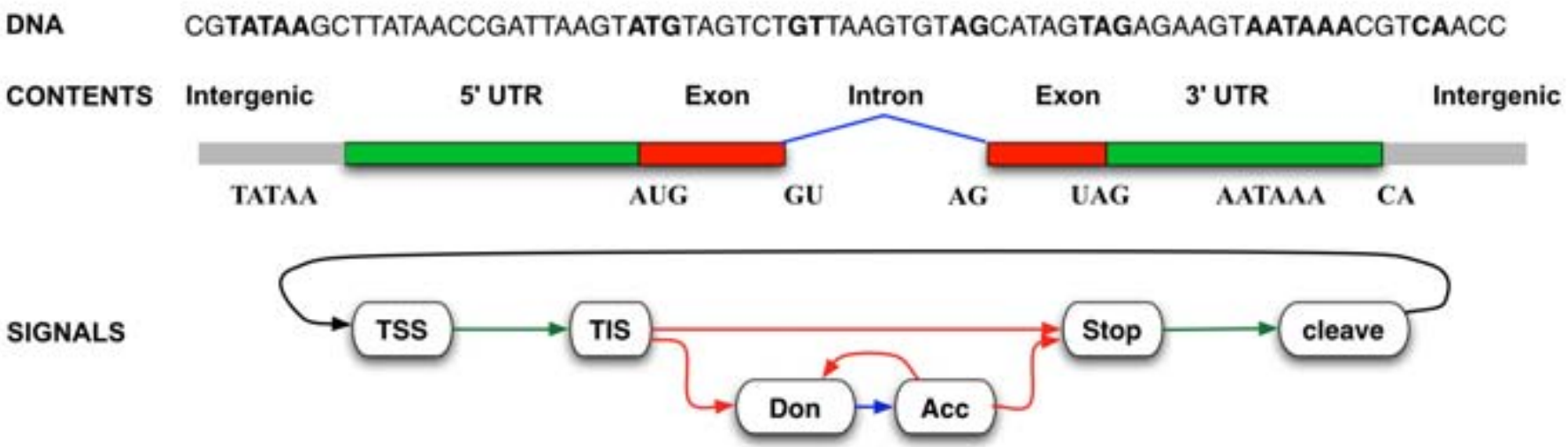
Gene Finding as a Segmentation Task

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



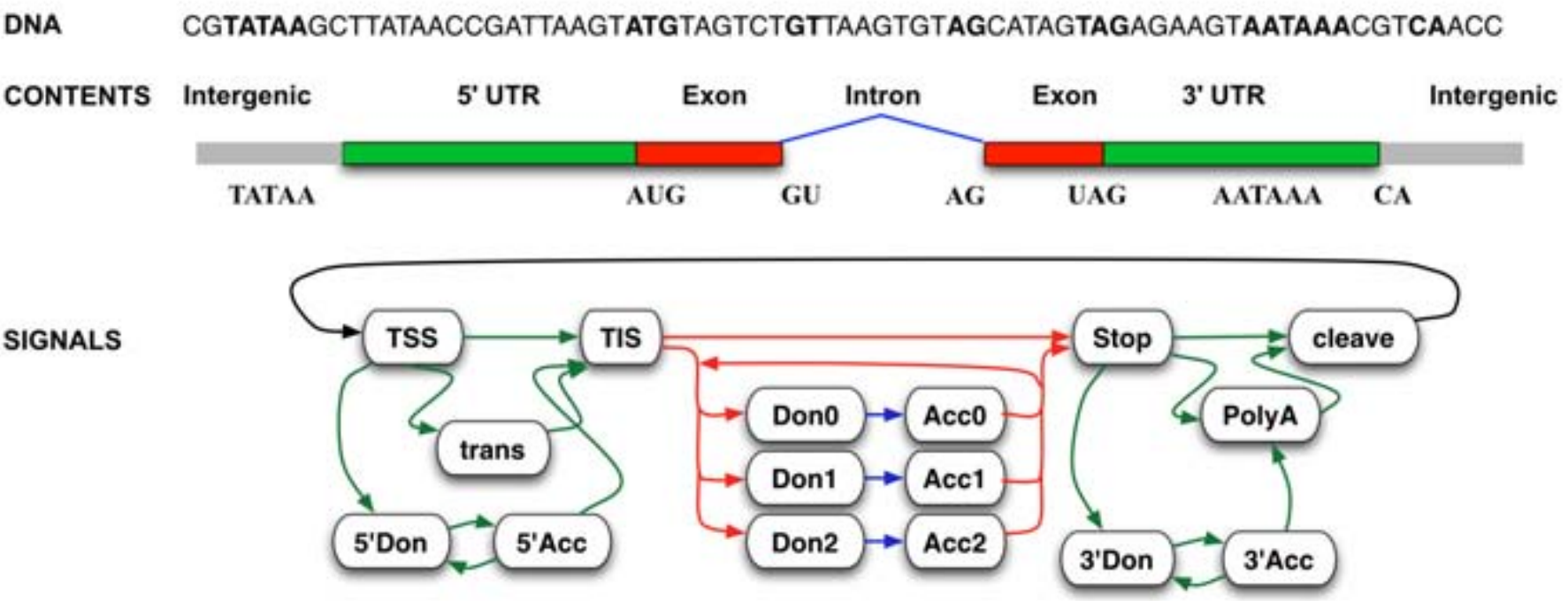
Gene Finding as a Segmentation Task

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



Gene Finding as a Segmentation Task

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



Signal and Content Sensors

Sensors to recognize signals:

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

Every non-signal position is a negative $\Rightarrow$ very unbalanced problem

Sensors to recognize contents:

- Exons & UTR
- Introns
- Intergenic

Recognize one content among others

Typical approach: PSSMs or higher order Markov chains

Here: Support Vector Machines

Signal and Content Sensors



Sensors to recognize signals:

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

Every non-signal position is a negative $\Rightarrow$ very unbalanced problem

Sensors to recognize contents:

- Exons & UTR
- Introns
- Intergenic

Recognize one content among others

Typical approach: PSSMs or higher order Markov chains

Here: Support Vector Machines

Signal and Content Sensors



Sensors to recognize signals:

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

Every non-signal position is a negative $\Rightarrow$ very unbalanced problem

Sensors to recognize contents:

- Exons & UTR
- Introns
- Intergenic

Recognize one content among others

Typical approach: PSSMs or higher order Markov chains

Here: Support Vector Machines

Signal and Content Sensors



Sensors to recognize signals:

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

Every non-signal position is a negative $\Rightarrow$ very unbalanced problem

Sensors to recognize contents:

- Exons & UTR
- Introns
- Intergenic

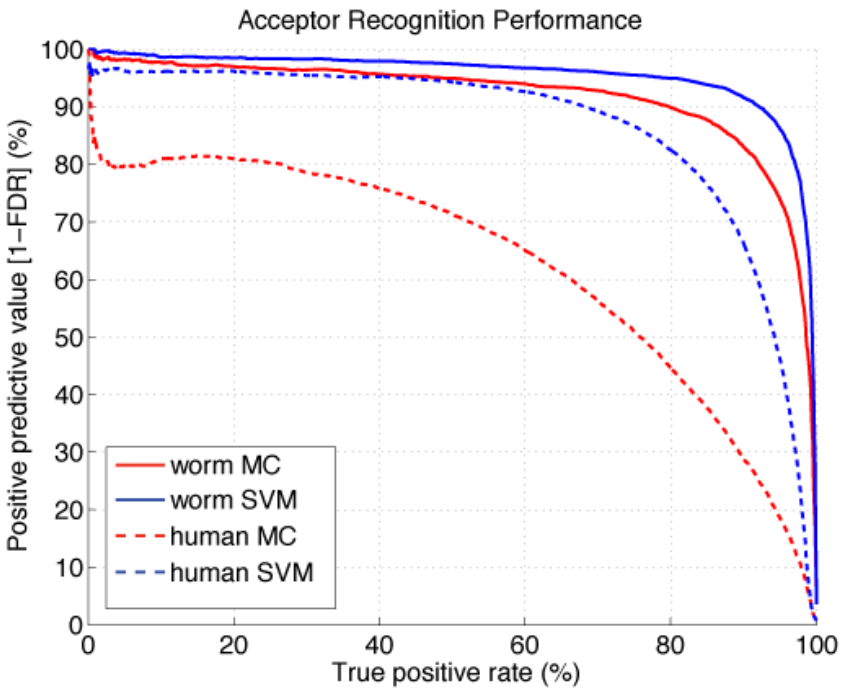
Recognize one content among others

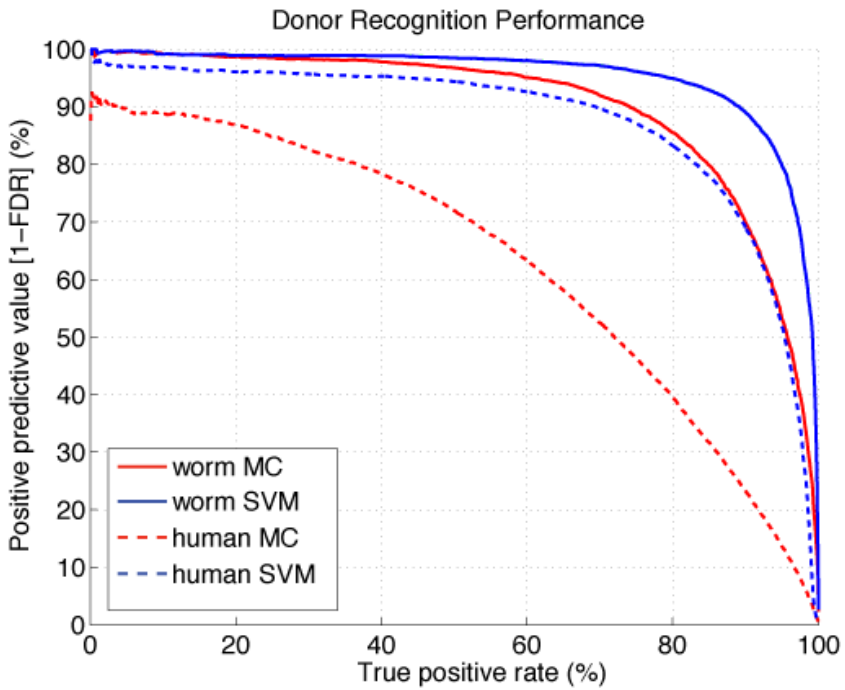
Typical approach: PSSMs or higher order Markov chains

Here: Support Vector Machines

Example: SVMs for Splice Site Recognition

	Worm		Fly		Cress		Fish		Human	
	Acc	**Don**	**Acc**	**Don**	**Acc**	**Don**	**Acc**	**Don**	**Acc**	**Don**
Markov Chains										
auROC(%)	99.62	99.55	98.78	99.12	99.12	99.44	98.98	99.19	96.03	97.78
auPRC(%)	92.09	89.98	80.27	78.47	87.43	88.23	63.59	62.91	16.20	24.98
SVM										
auROC(%)	99.80	99.82	99.12	99.51	99.43	99.68	99.38	99.61	97.86	98.57
auPRC(%)	95.89	95.34	86.67	87.47	92.16	92.88	86.58	86.94	54.42	56.86





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

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
    (implemented in Shogun toolbox: `http://www.shogun-toolbox.org`)

(Sonnenburg, Rieck, Rätsch, 2007)

- No probabilistic outputs
  - Prerequisite for use in generative model 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
    (implemented in Shogun toolbox: `http://www.shogun-toolbox.org`)

(Sonnenburg, Rieck, Rätsch, 2007)

- No probabilistic outputs
  - Prerequisite for use in generative model 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:

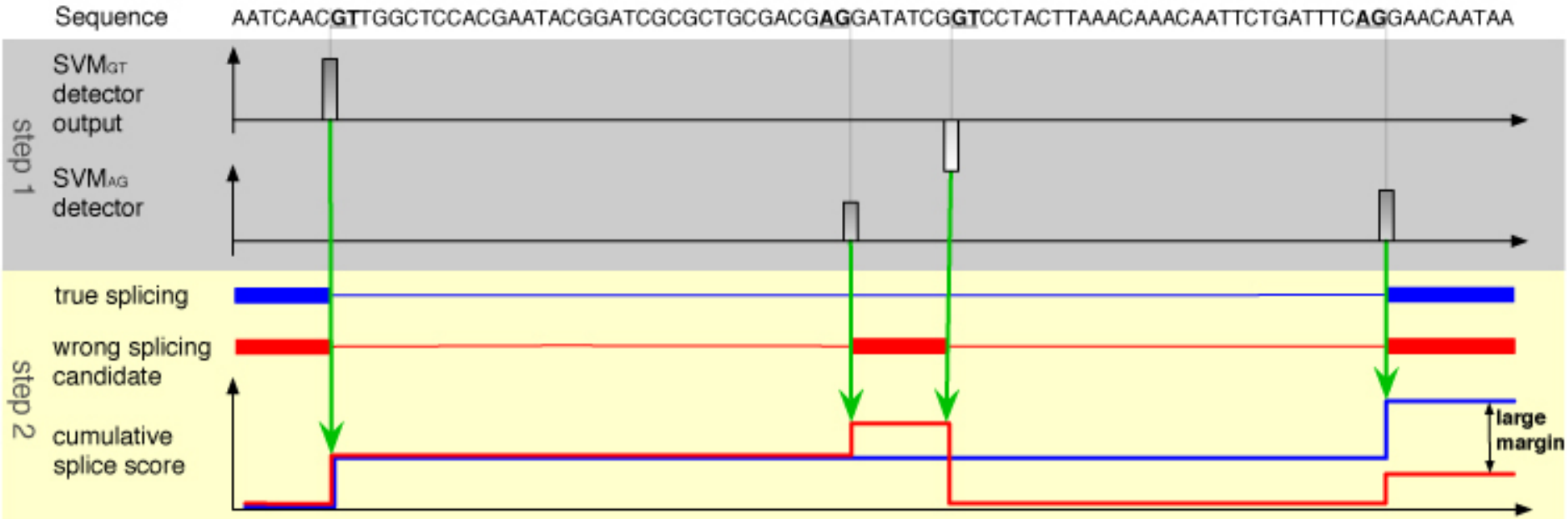
- (Originally) not large scale
  - New developments: Can deal with 10 million examples
    (implemented in Shogun toolbox: `http://www.shogun-toolbox.org`)

(Sonnenburg, Rieck, Rätsch, 2007)

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

(Tsochantaridis et al., 2005, Rätsch & Sonnenburg 2006)

Large Margin Learning (simplified)

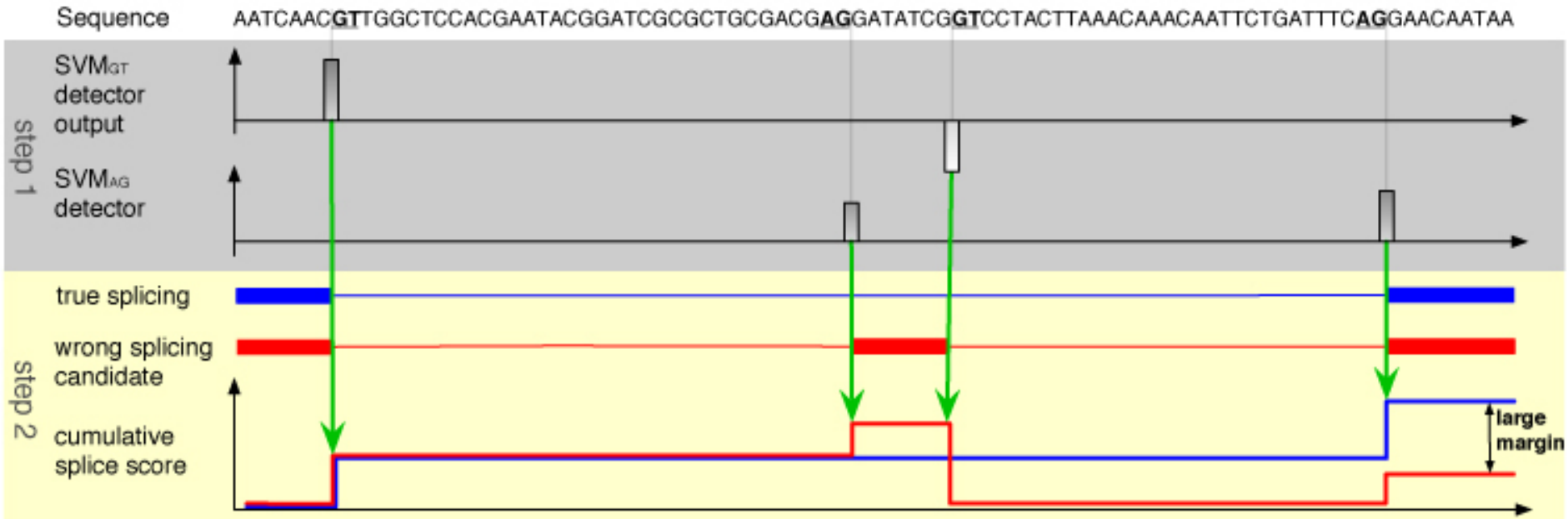


- Simplified Model: Score for splice form $\mathbf{y} = \{(p_j, q_j)\}_{j=1}^{J}$:

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

- Tune free parameters (in functions $S_{GT}, S_{AG}, S_{L_E}, S_{L_I}$) by solving **linear program** using training set with known splice forms

Large Margin Learning (simplified)

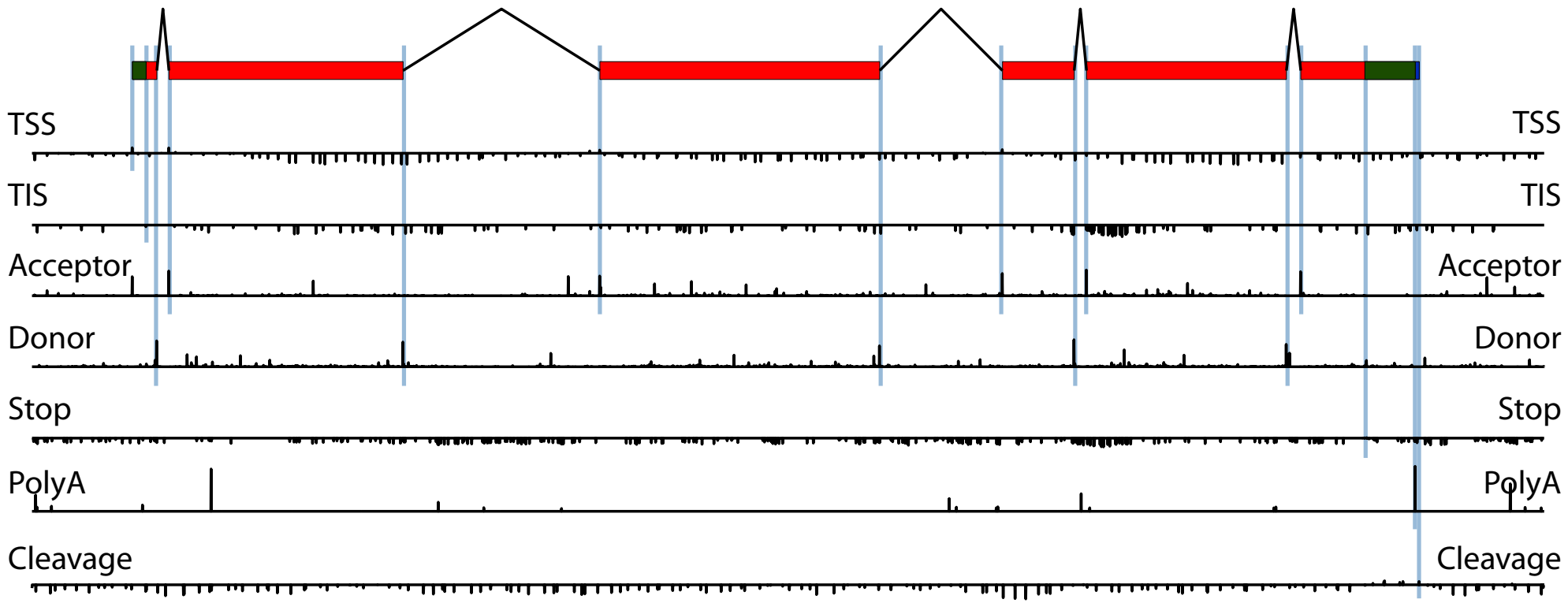


- Simplified Model: Score for splice form $\mathbf{y} = \{(p_j, q_j)\}_{j=1}^{J}$:

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

- Tune free parameters (in functions $S_{GT}, S_{AG}, S_{L_E}, S_{L_I}$) by solving **linear program** using training set with known splice forms

Example


TSS
TSS
TIS
TIS
Acceptor
Acceptor
Donor
Donor
Stop
Stop
PolyA
PolyA
Cleavage
Cleavage

Learning to Predict Segmentations

- Learn function $f(\mathbf{y}|\mathbf{x})$ scoring segmentations $\mathbf{y}$ for sequences $\mathbf{x}$
- f considers signal, content and length information
- Maximize $f(\mathbf{y}|\mathbf{x})$ w.r.t. $\mathbf{y}$ to perform prediction: $\operatorname{argmax}_{\mathbf{y}} f(\mathbf{y}|\mathbf{x})$
- Determine f such that there is a large margin between true and wrong segmentations

$$\begin{aligned} \min_{f} \quad & \sum_{n=1}^{N} \xi_n + \mathbf{P}[f] \\ \text{w.r.t.} \quad & f(\mathbf{y}_n|\mathbf{x}_n) - f(\mathbf{y}|\mathbf{x}_n) \geq 1 - \xi_n \\ & \text{for all } \mathbf{y} \neq \mathbf{y}_n, n = 1, \ldots, N \end{aligned}$$

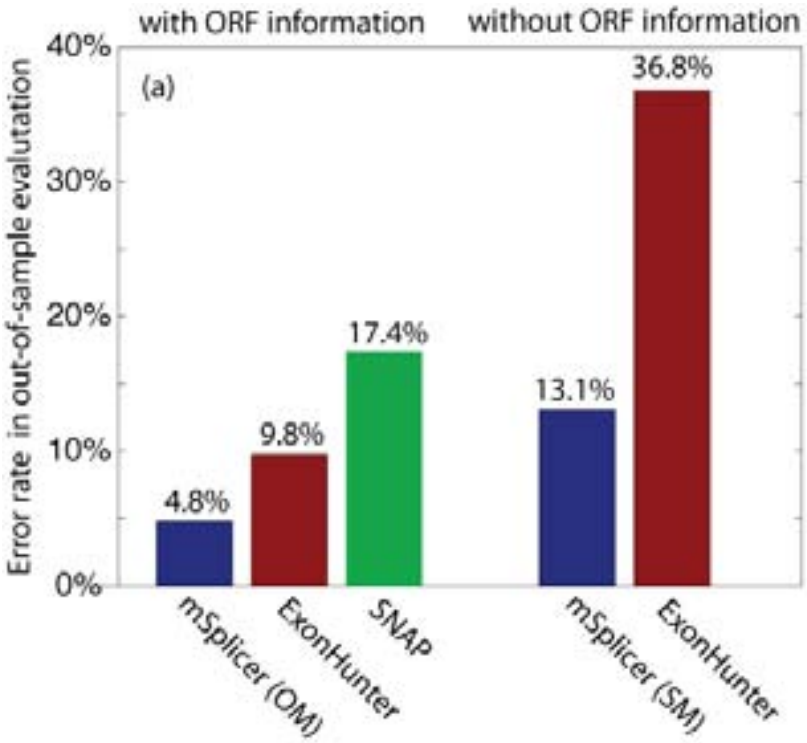
- Use approximation (Rätsch & Sonnenburg, NIPS'06)
  - Train signal and content detectors separately
  - Combine in large margin fashion

Results Overview

- Splice forms only (Rätsch et al., PLoS Comp. Biol., 2007)
  - Comparison with other methods
  - Analysis of a few disagreeing cases
  - Results available on `http://www.wormbase.org`
- Full gene predictions
  - Participation at nGASP competition
  - Preliminary evaluation (officially due 31st of May)

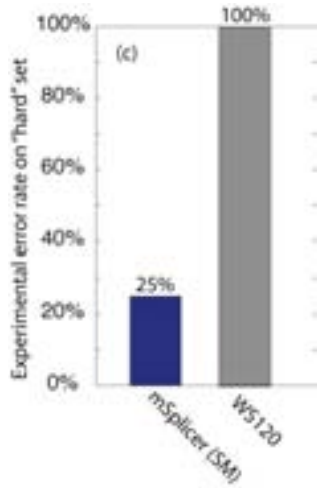
Results I (Splice forms only)

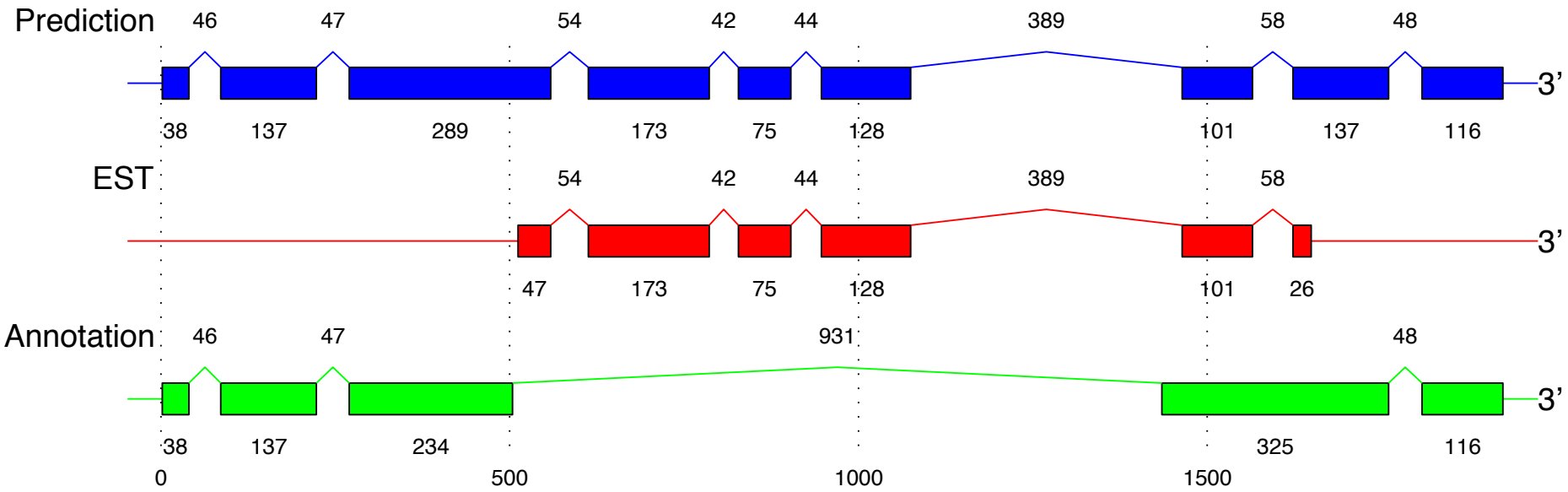
- ≈3,800 gene models derived from cDNAs and ESTs
  - 60% for training and validation
  - 40% for testing (exclude alt. spliced genes)
- Out-of-sample accuracy (≈1100 gene models):
  - Splice form error rate
    - 4.8% (coding)
    - 13.1% (mixed)
  - Much lower error rates than state-of-the-art
    -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



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



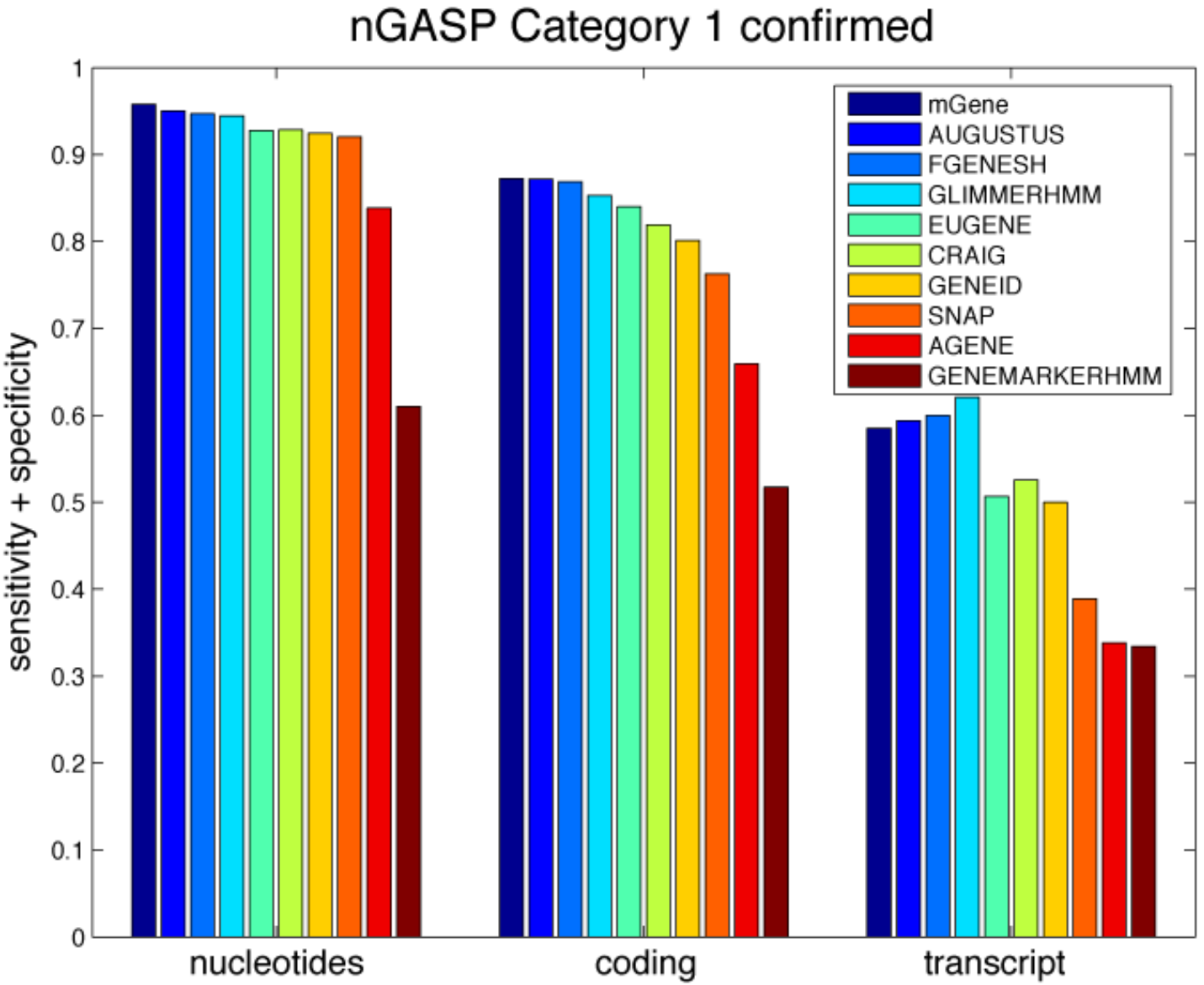


Results III: nGASP Competition

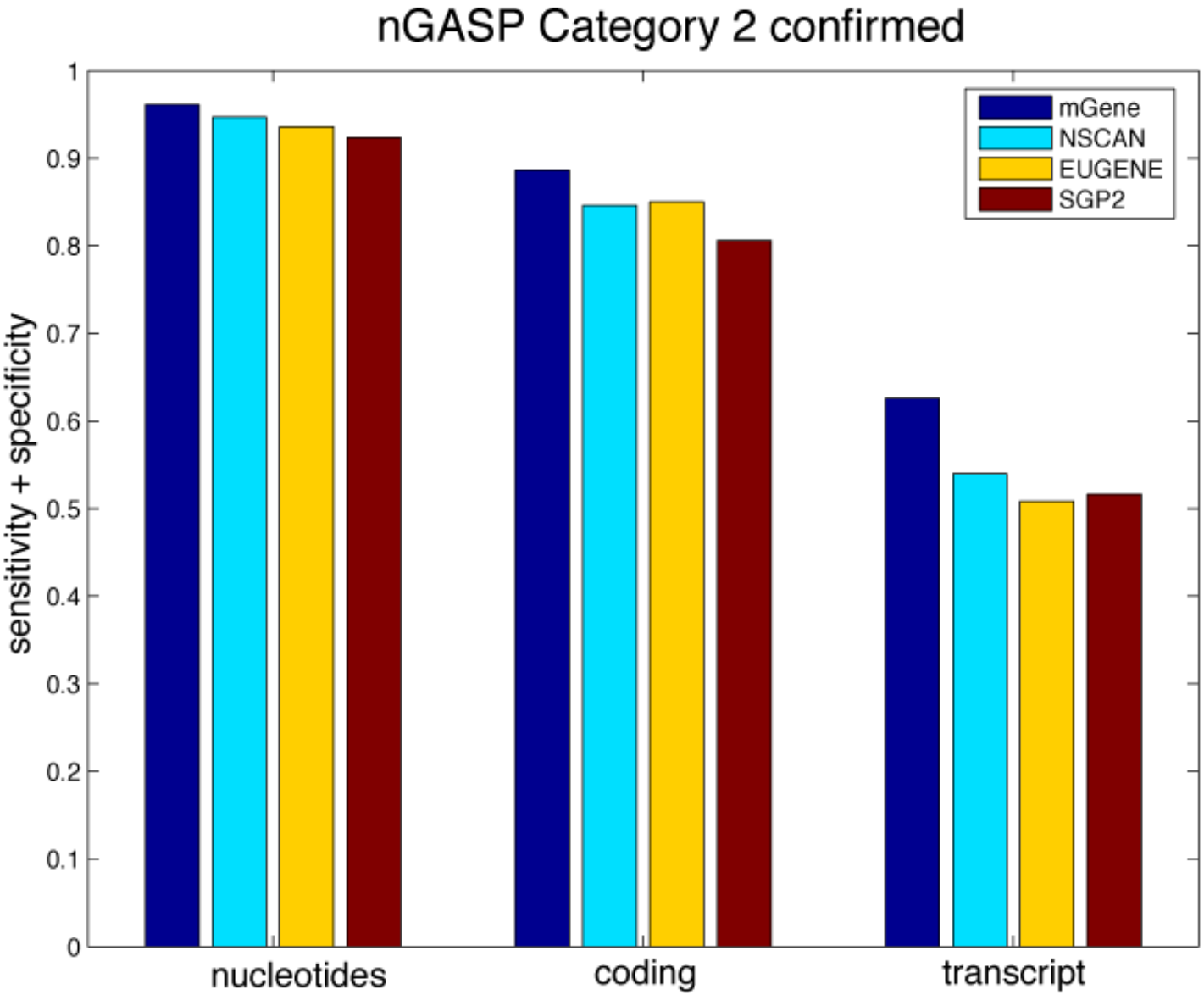


- Controlled competition conditions:
  - 10% for training methods
  - 10% for evaluation
- Phase I: Single predictors
- Phase II: Combining algorithms
- Four categories:
  - *Ab initio* gene finders
  - Dual/Multi-genome gene finders
  - Gene finders that use alignments of sequences
  - Gene finders that use any of above information
- Official evaluation due on 31st of March (not public yet)
- Preliminary evaluation on WS160 fully confirmed genes

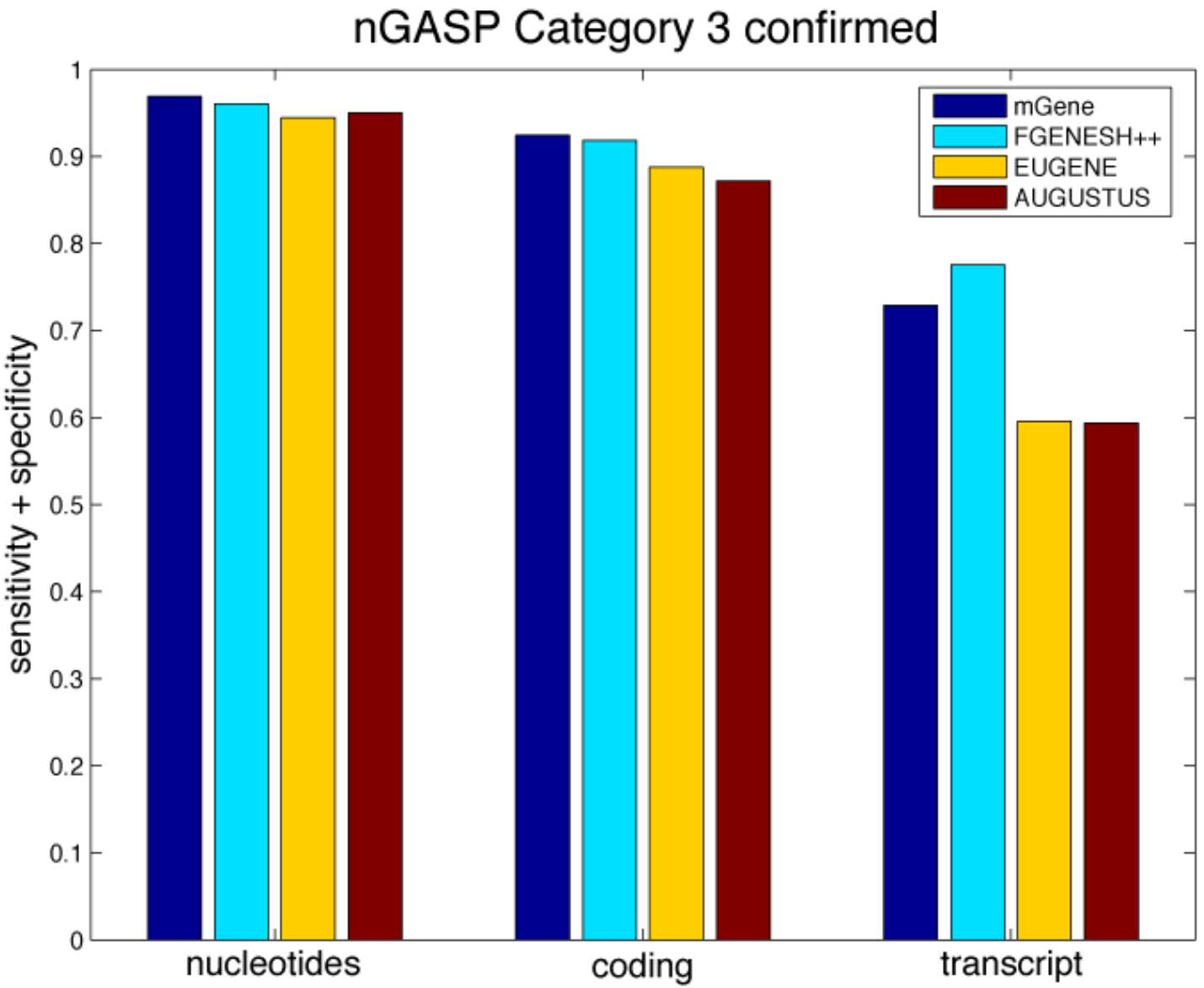
nGASP Cat. 1 evaluation (prelim.)



nGASP Cat. 2 evaluation (prelim.)



nGASP Cat. 3 evaluation (prelim.)

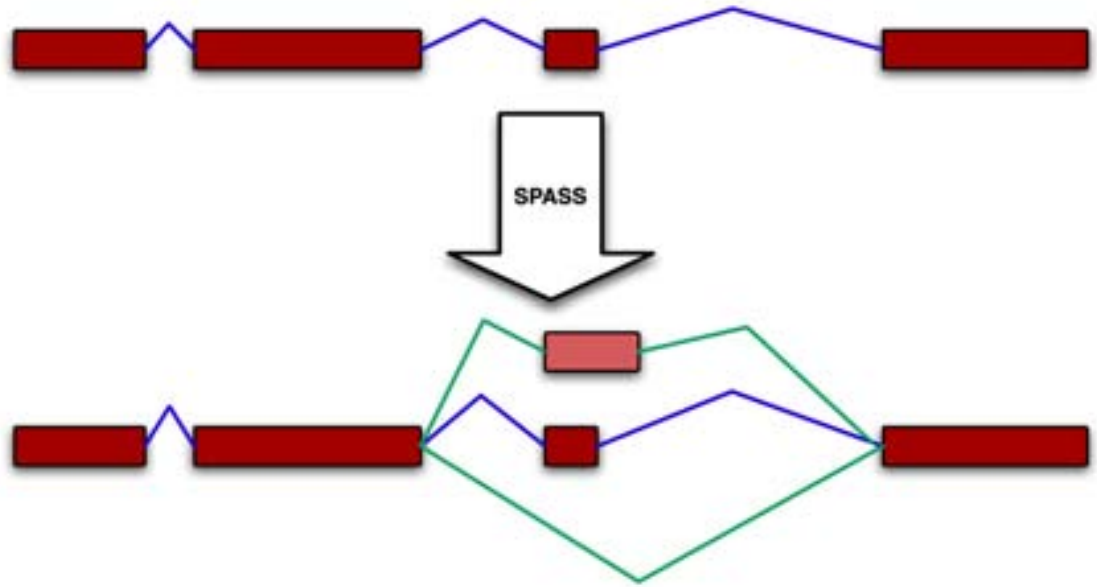


Discussion

- Good on nucleotide and exon level
- Suboptimal on transcript level:
  - Genes are often merged
- Surprise on Category 2: Average performance worse than Category 1
- Simple approach for Category 3 worked pretty well
- Possible improvements:
  - Use RepeatMasker features
  - Revise the loss function (penalize wrong transcripts)
  - Use Protein alignments in Category 3

Alternative Splicing: First Steps

Predictions of Alternative Splicing



- Predict novel alternative splicing as independent events
- Use only information available to splicing machinery (Rätsch et. al, ISMB'05)
- Quite accurate for frequently appearing patterns
- Requires known gene structures

More Steps

- Combine gene finding with prediction of single alternative splicing events
  - Predict the **Splice graph** of a gene
  - Machine learning challenge:
    - Input: DNA sequence
    - Output: Splice-graph
- mSplicer approach can be extended to
  - Include predictions of alternative splicing events
  - Predict **"simple splice graphs"**
- Predicting arbitrary graphs is considerably harder

Discussion and Future Work



- Gene structure prediction is still a big challenge
- High accuracy on constitutively spliced genes for nematodes
- First version of new gene finder system seems very promising
  - Compares favorably compared to CRAIG
  - 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

Acknowledgements & 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, Fabio de Bona, Petra Philips, Sören Sonnenburg

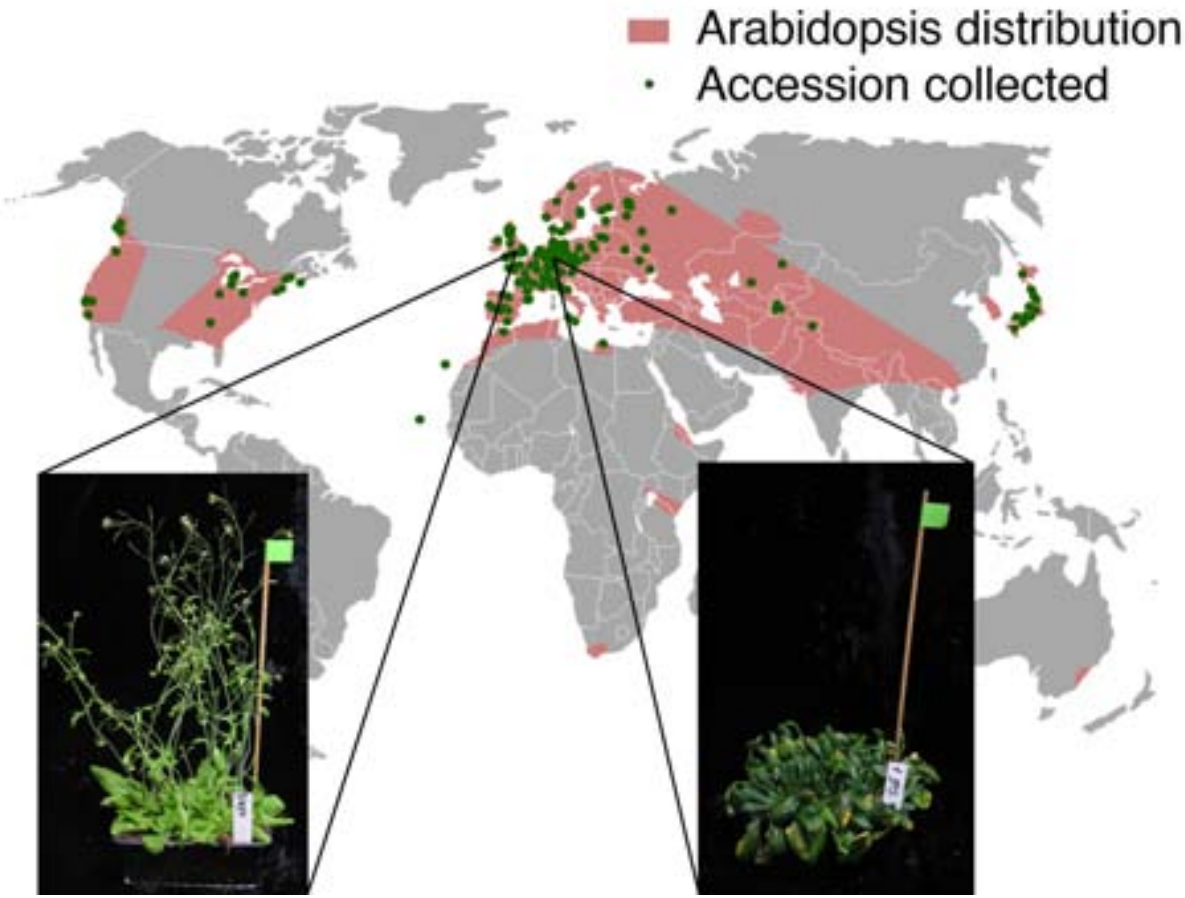
- Homepage:
  `http://www.fml.mpg.de/raetsch`
- Details, Data sets & Predictions:
  `http://www.msplicer.org` and `http://www.mgene.org`
- A large scale open source ML toolbox:
  `http://www.shogun-toolbox.org`
- Other projects:
  `http://www.fml.mpg.de/raetsch/projects`

Discovering Common Sequence Variation in Arabidopsis thaliana

Gabriele Schweikert,[1,2,3] Georg Zeller,[1,3] Richard Clark,[3] Stephan Ossowski,[3] Christopher Toomajian,[4] Norman Warthmann,[3] Kelly Frazer,[5] Joseph Ecker,[6] Magnus Nordborg,[4] Detlef Weigel,[3] Bernhard Schölkopf,[2] Gunnar Rätsch[1]

[1] Friedrich Miescher Laboratory, Tübingen
[2] Max Planck Institute for Biological Cybernetics, Tübingen
[3] Max Planck Institute for Developmental Biology, Tübingen
[4] Molecular and Computational Biology, UCLA, CA
[5] Perlegen Sciences, Mountain View, CA
[6] Genomic Analysis Laboratory, The Salk Institute, La Jolla, CA

Introduction



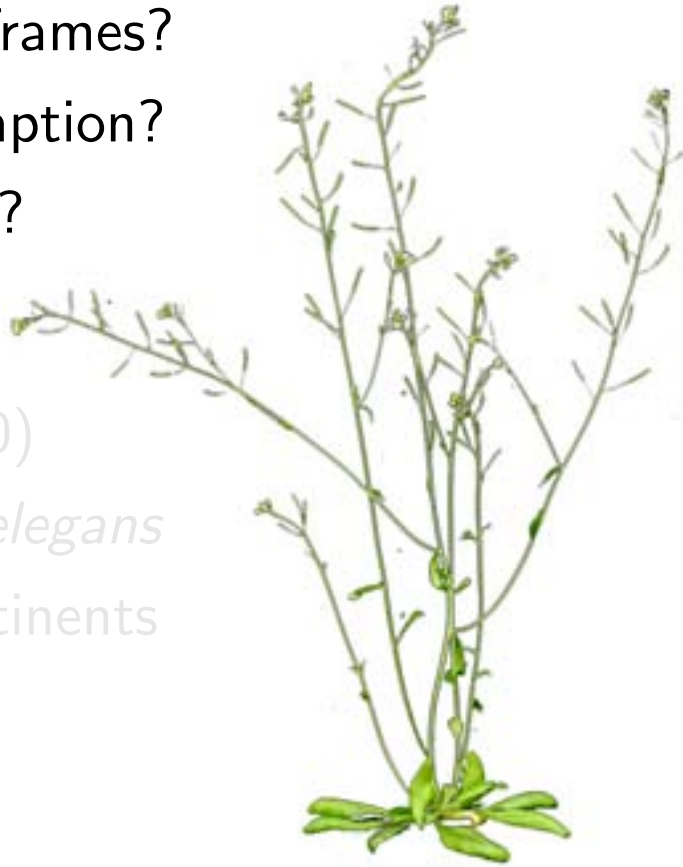
What is the genetic basis of variation?

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

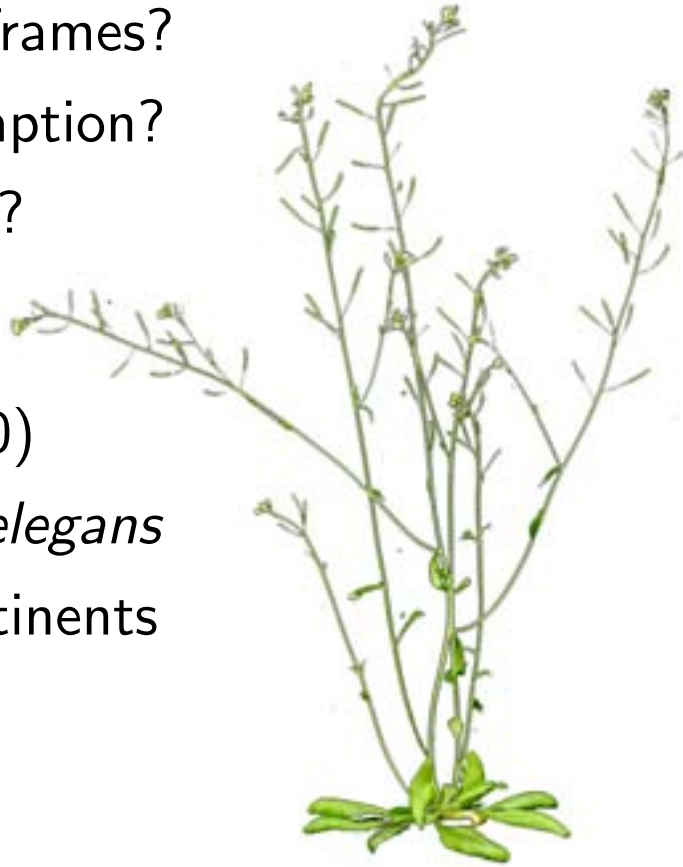
Resequencing of 20 wild strains

- 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

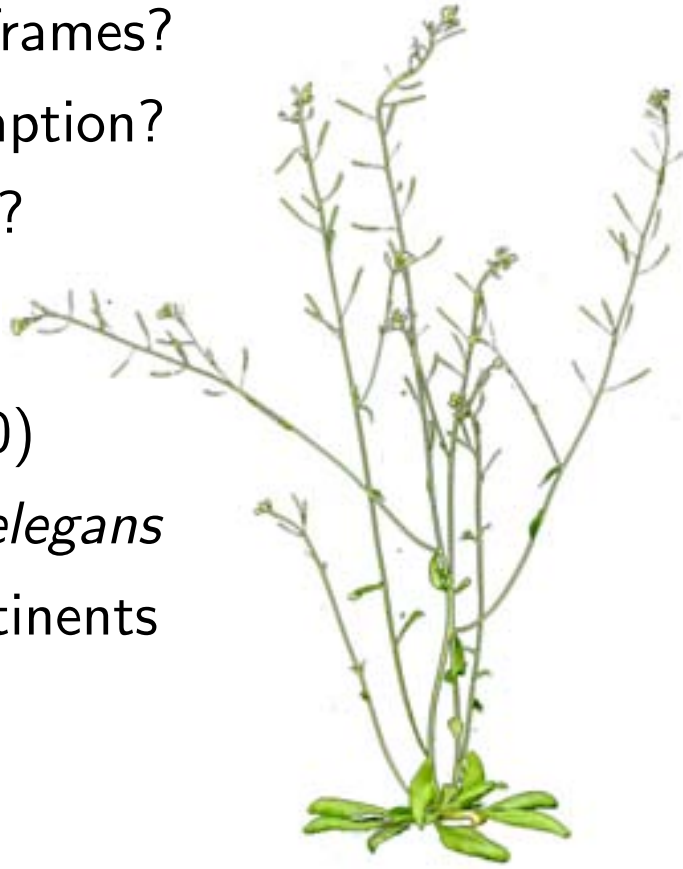
Resequencing of 20 wild strains

- 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?



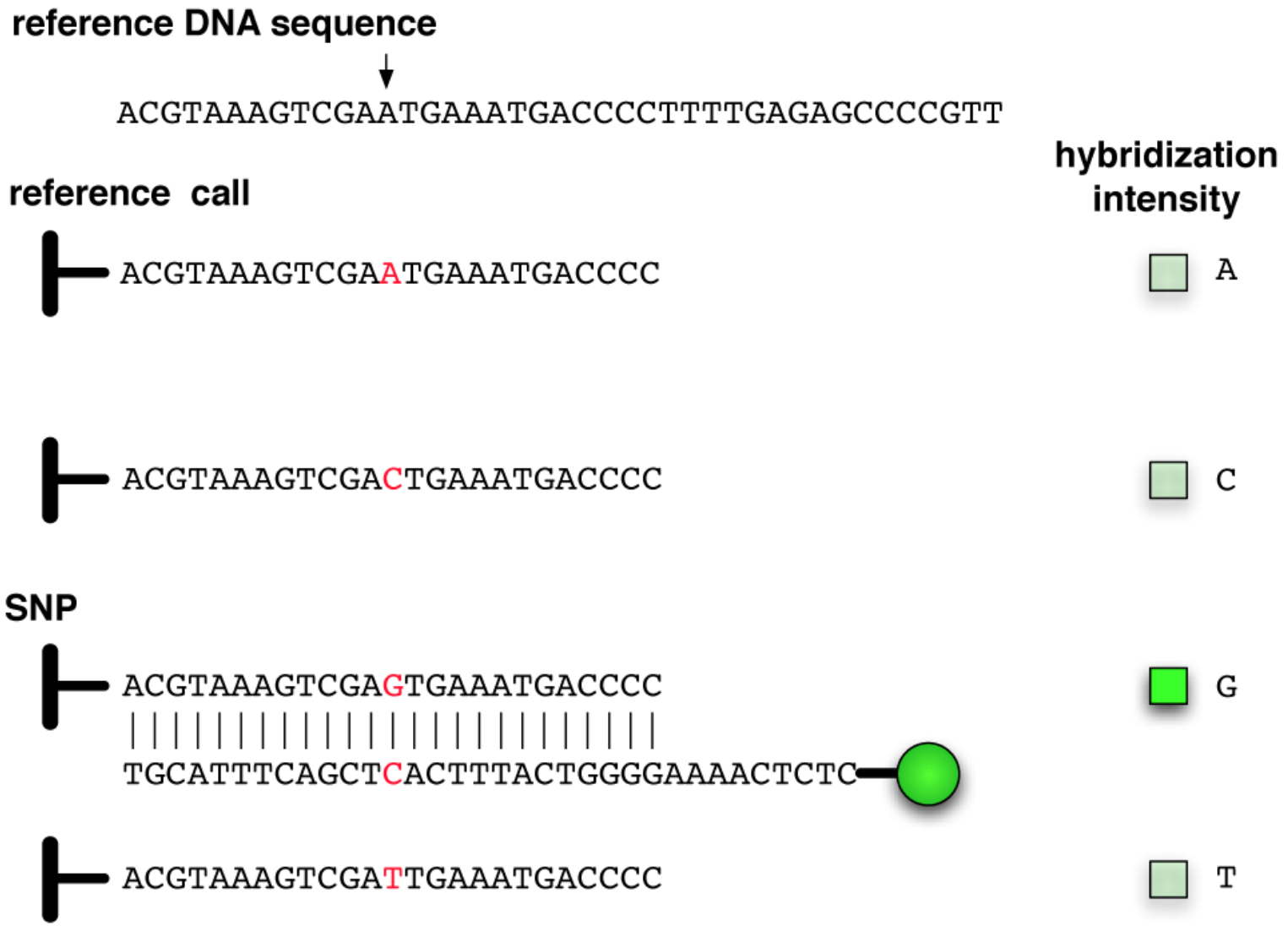
Arabidopsis thaliana:

- 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

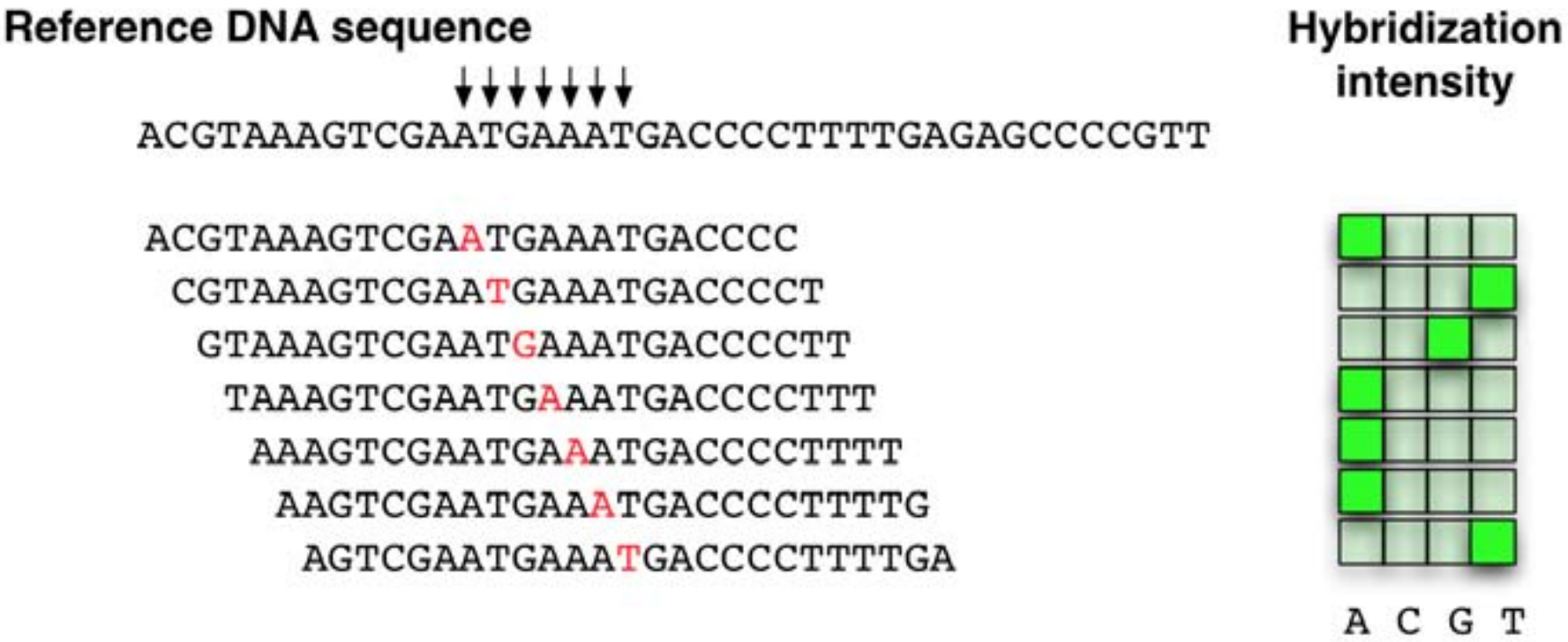
Resequencing of 20 wild strains

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

Resequencing Array Basics I

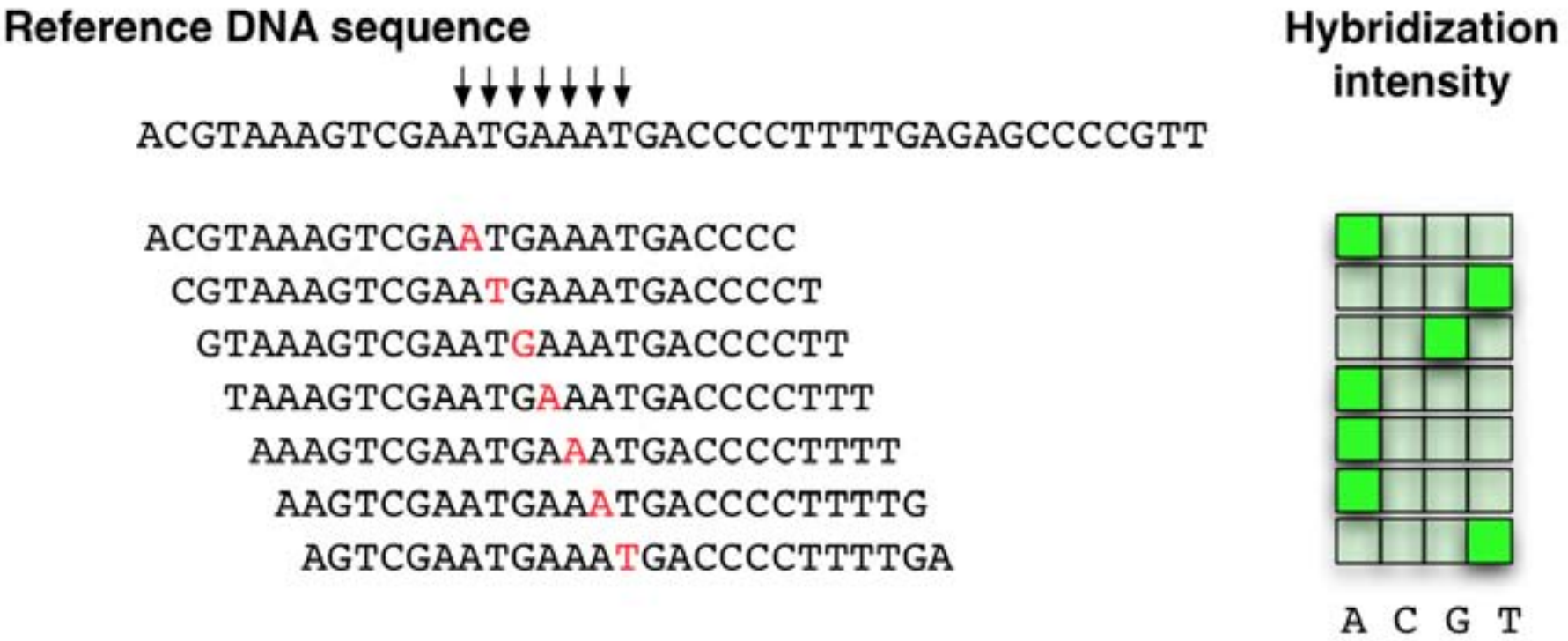


Resequencing Array Basics II



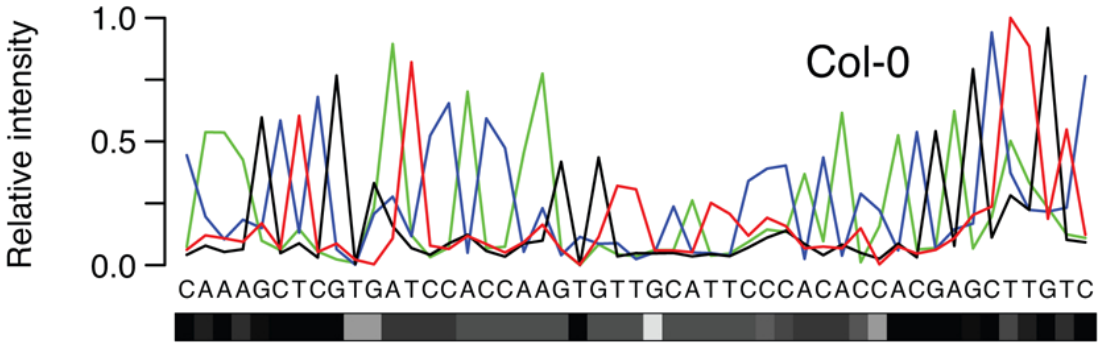
- >99.99% of bases represented
- Each base queried with forward and reverse strand probe quartets
- Nearly 1 billion oligos per ecotype
- 19+1 ecotypes surveyed representing worldwide distribution

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

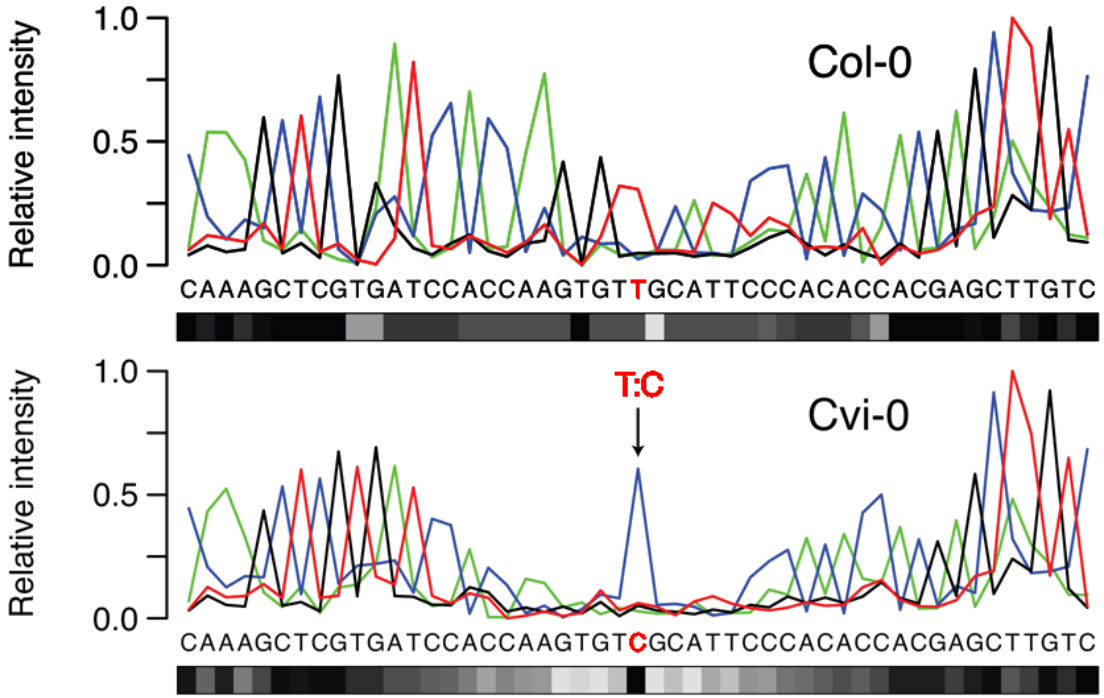
- Hybridization intensities depend on
  - Oligomer
  - Ecotype
  - Repeats
- Measurement noise
- Identify SNPs

Problematic cases

- Highly polymorphic regions
- Deletions/insertions



Resequencing Data





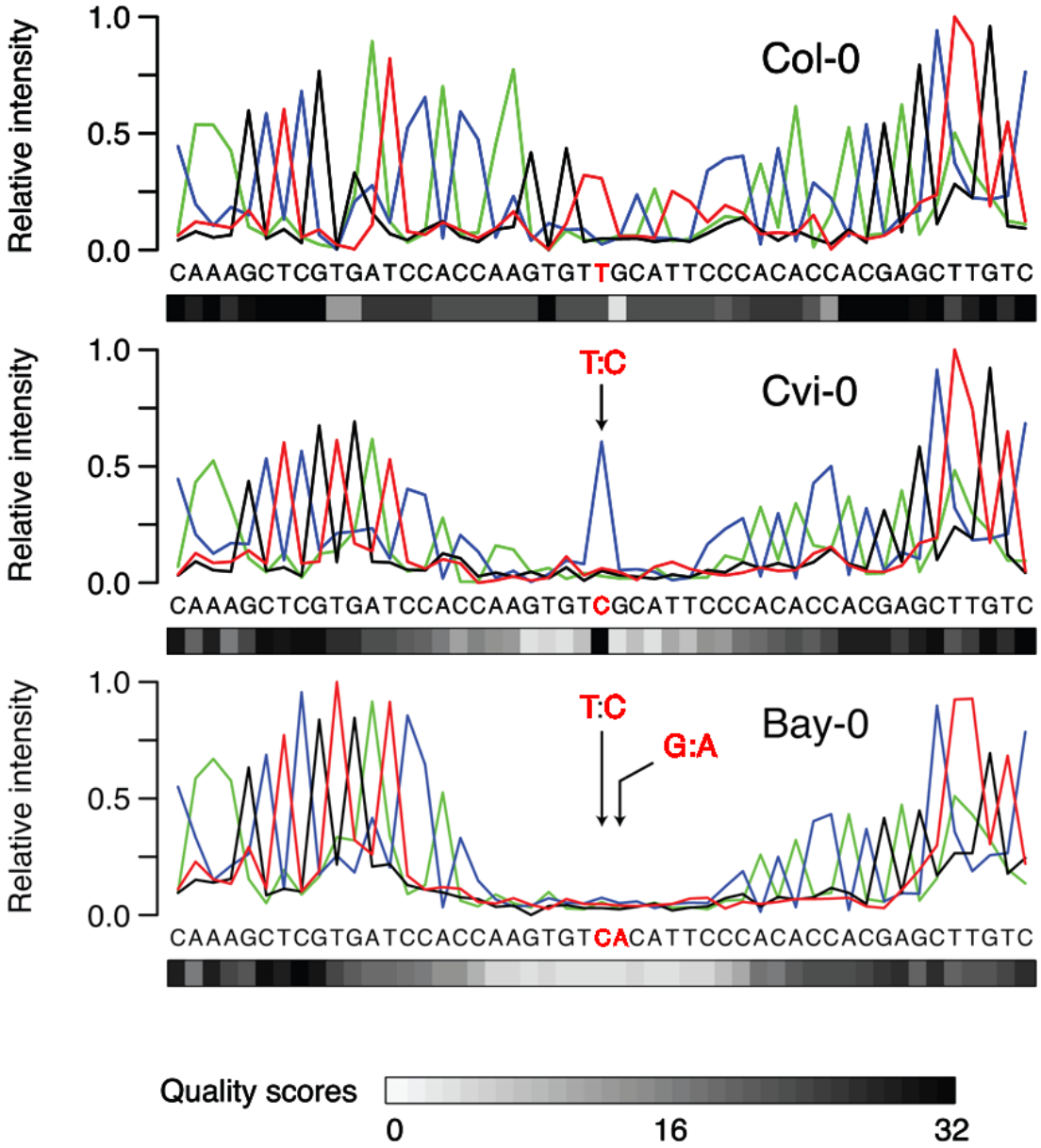
Data analysis challenge

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

Problematic cases

- Highly polymorphic regions
- Deletions/insertions

Resequencing Data



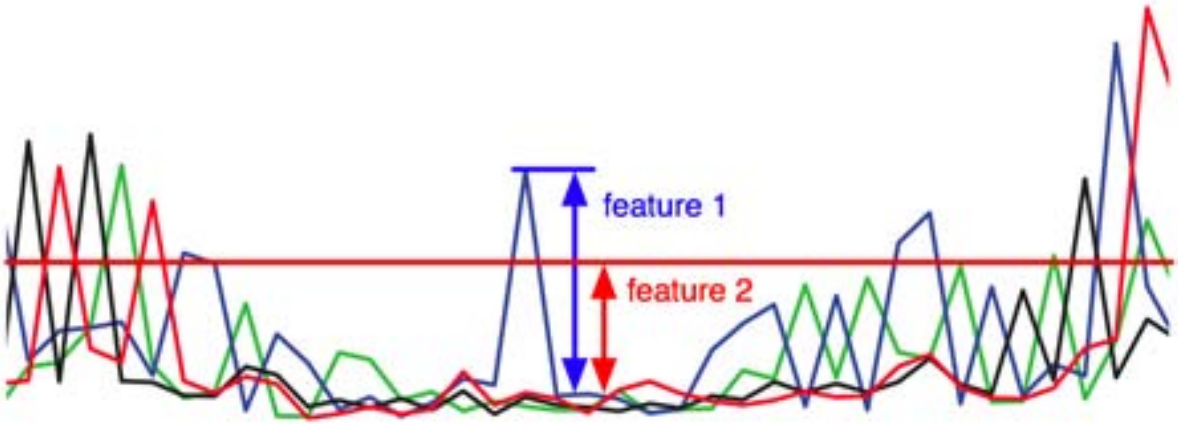
Data analysis challenge

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

Problematic cases

- Highly polymorphic regions
- Deletions/insertions

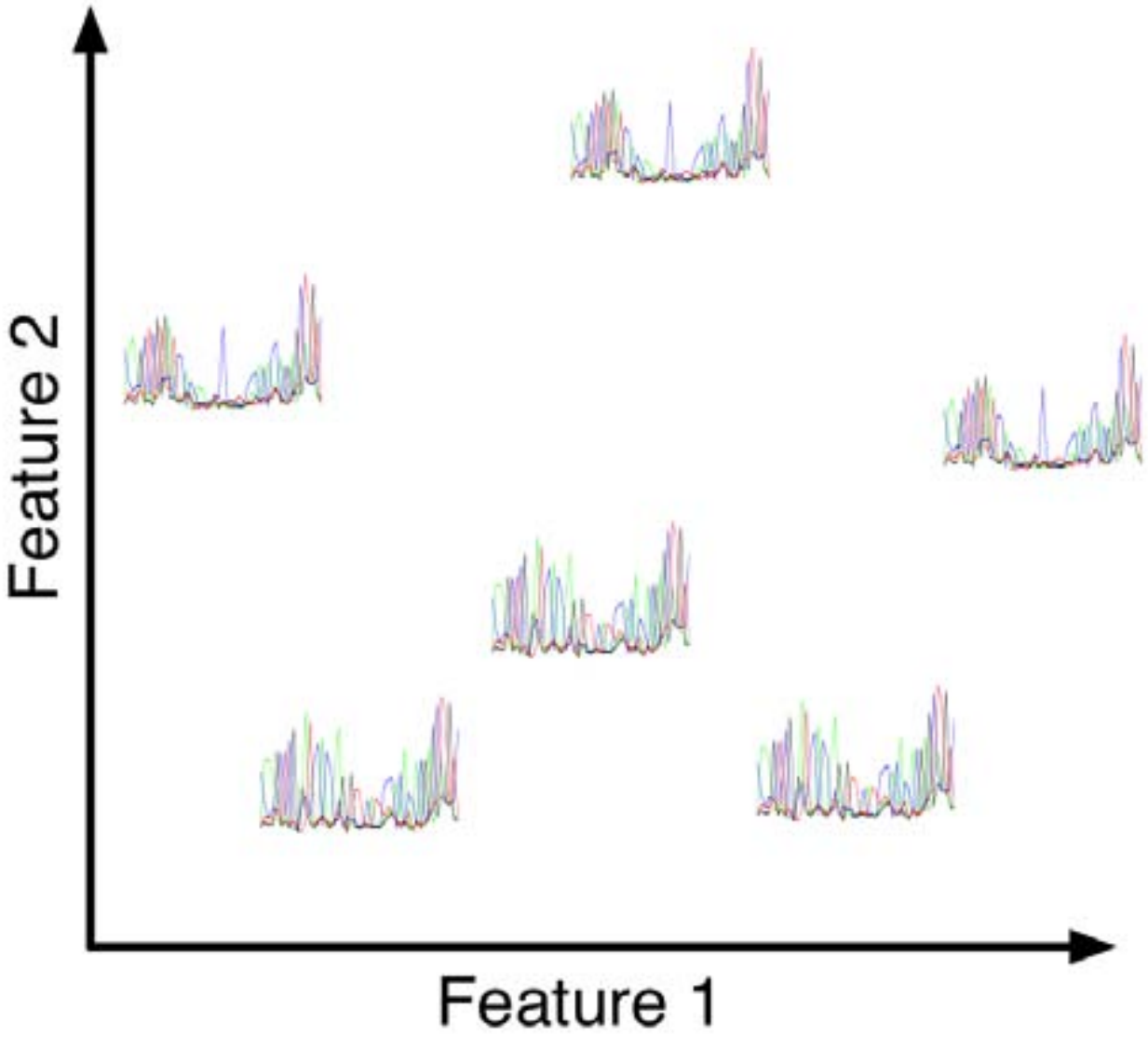
Machine Learning for Resequencing Data



Large Margin Learning

- Extract features
- Find large margin separation
- Classify new points
- Works with

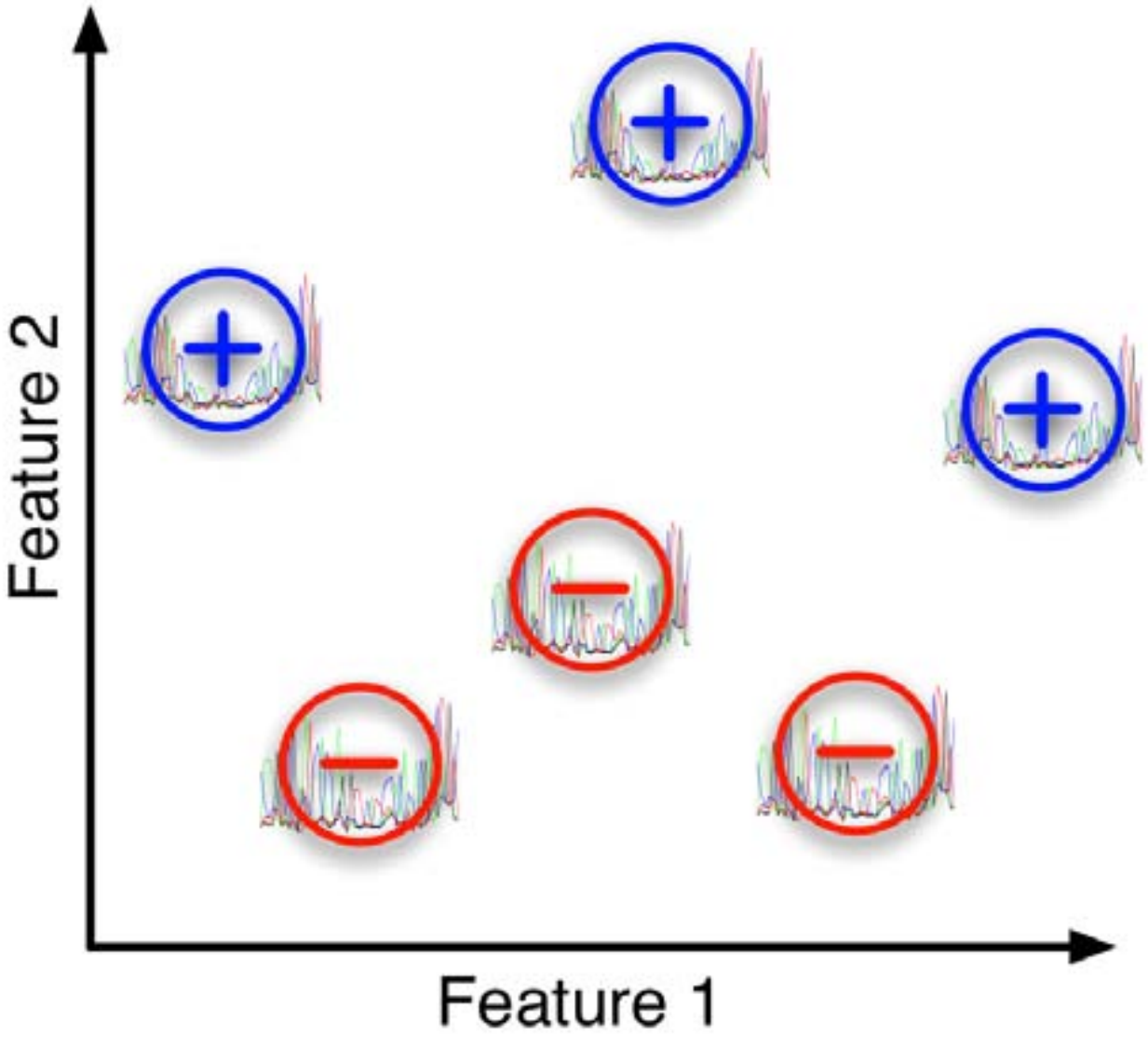
Machine Learning for Resequencing Data



Large Margin Learning

- Extract features
- Find large margin separation
- Classify new points
- Works with
  - many features
  - nonlinearities using RBF kernels

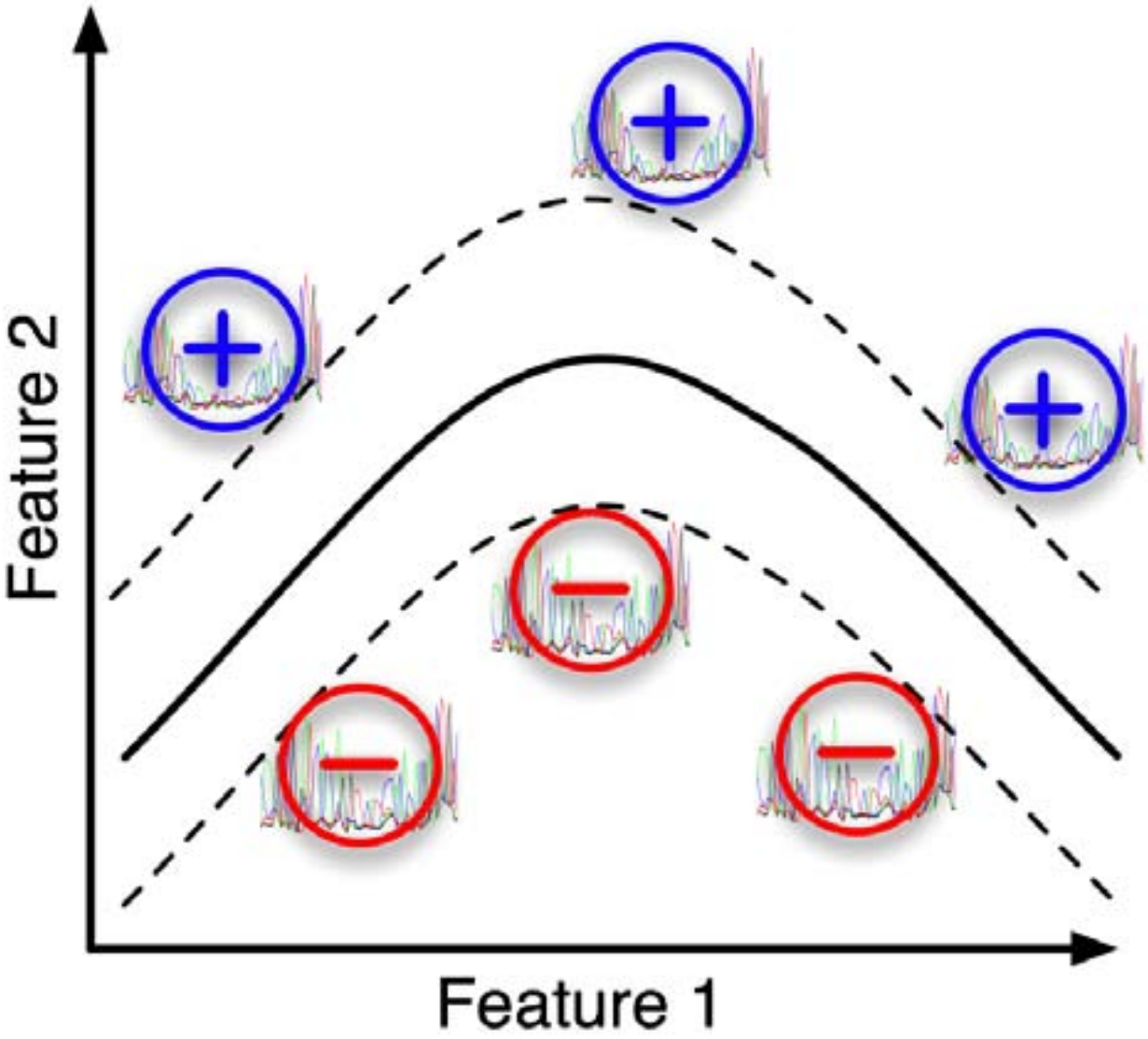
Machine Learning for Resequencing Data



Large Margin Learning

- Extract features
- Find large margin separation
- Classify new points
- Works with
  - many features
  - nonlinearities using RBF kernels

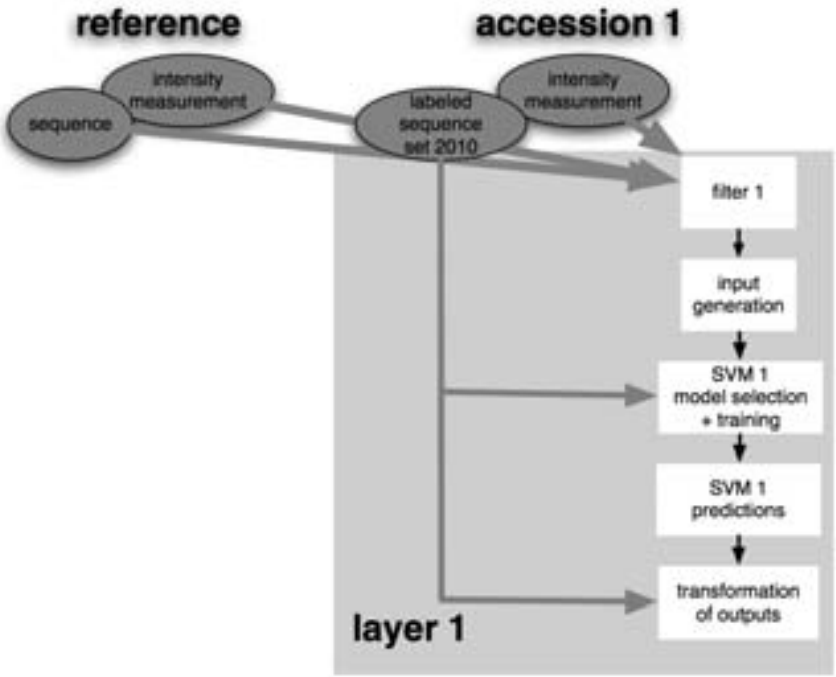
Machine Learning for Resequencing Data



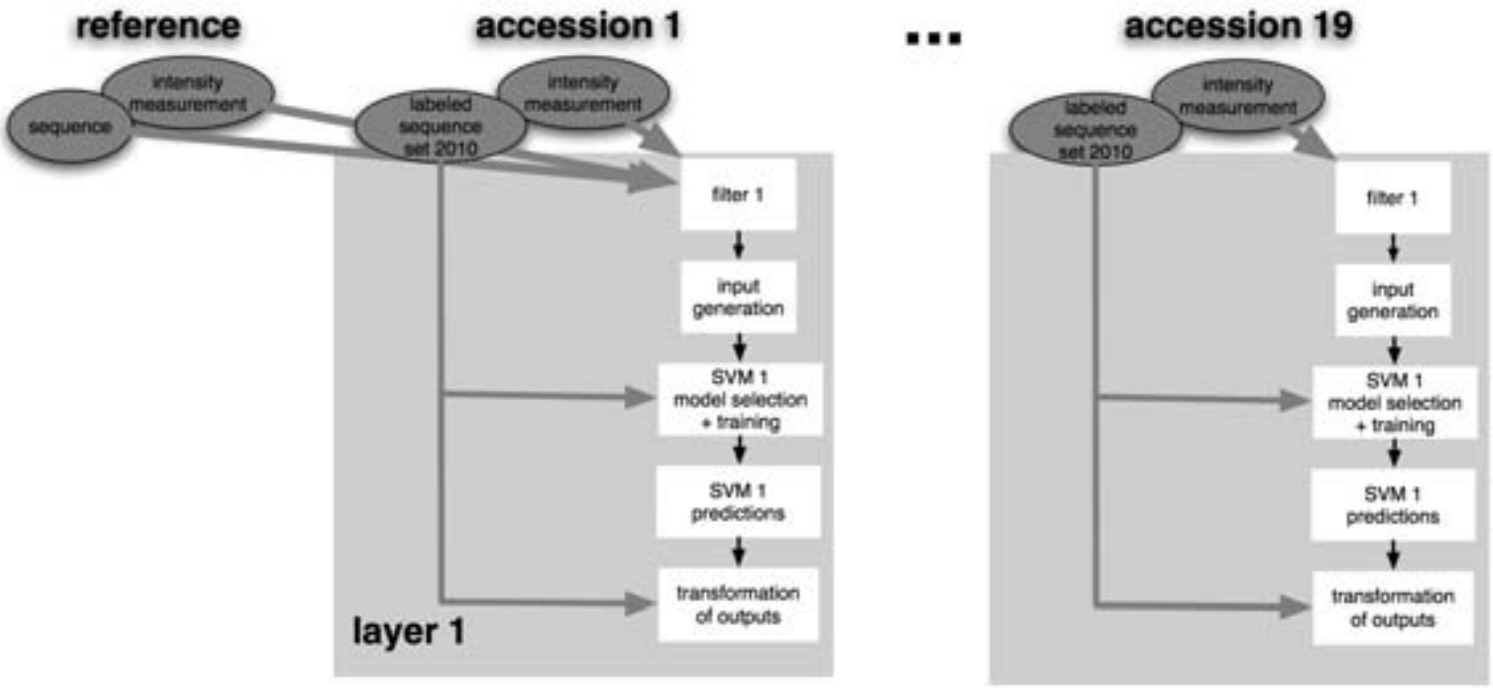
Large Margin Learning

- Extract features
- Find large margin separation
- Classify new points
- Works with
  - many features
  - nonlinearities using RBF kernels

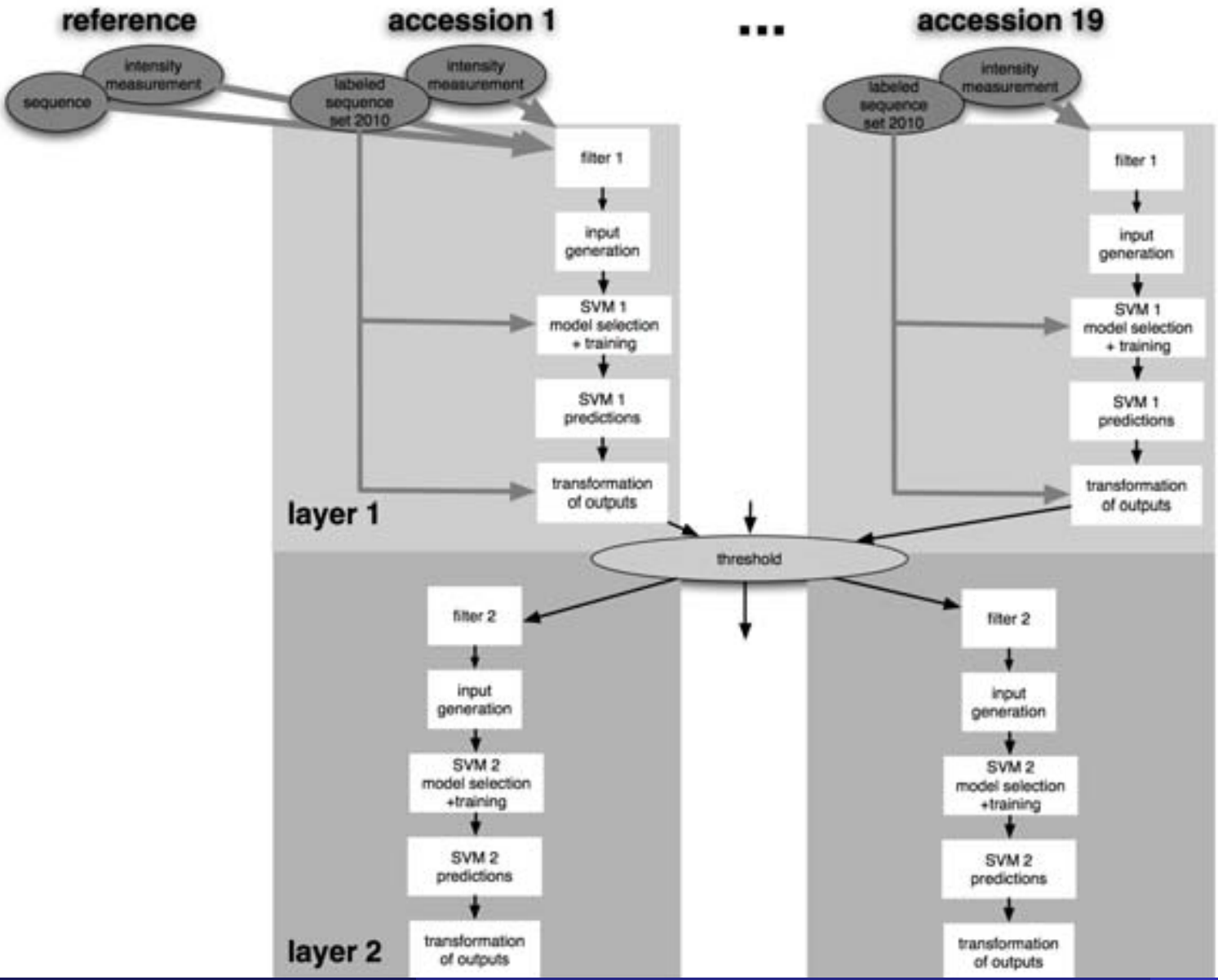
It's Actually A Bit More Complicated...



It's Actually A Bit More Complicated...

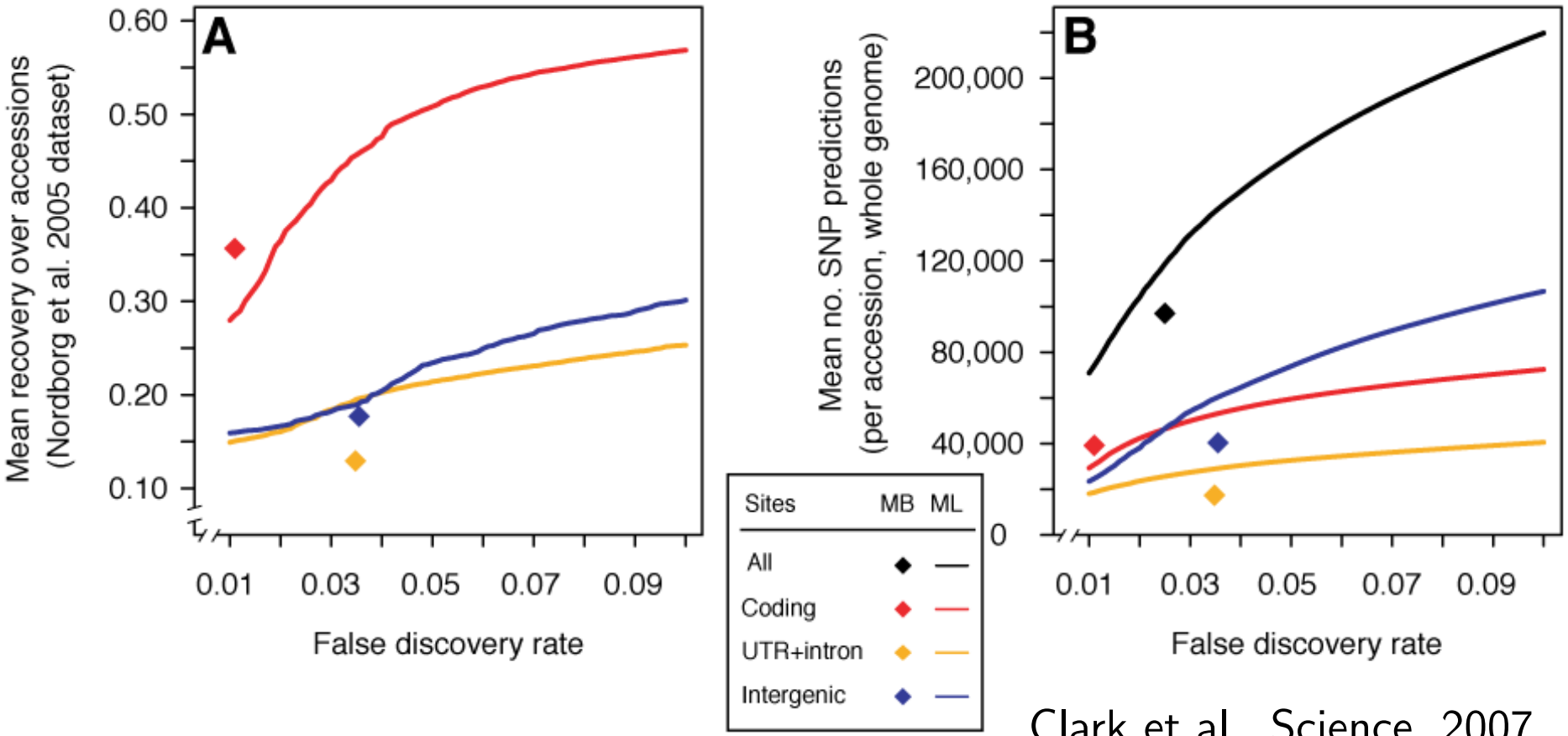


It's Actually A Bit More Complicated...



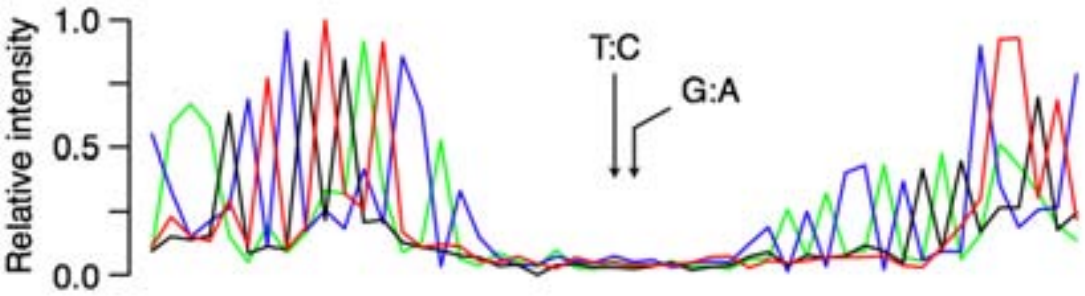
Application to SNP Discovery

- 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.

Identification of Highly Polymorphic Regions



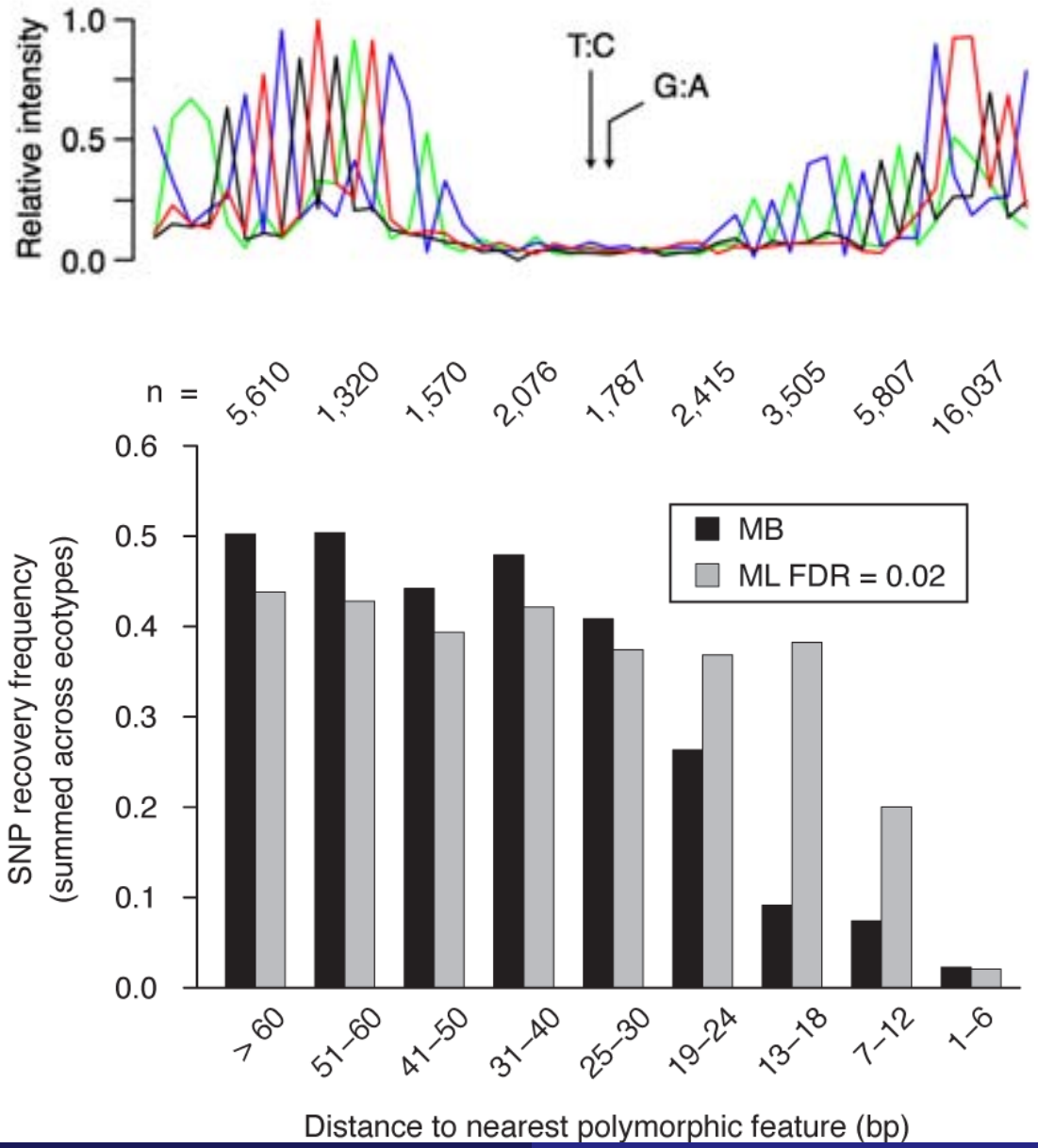
Results

- Performance drops, when other SNPs are in vicinity (1-20nt)
- Least predicted SNPs in highly polymorphic regions!
- ML more sensitive

New Approach

- **P**olymorphic **R**egion **P**rediction (PRP)
- Novel ML techniques for predicting complex properties

Identification of Highly Polymorphic Regions



Results

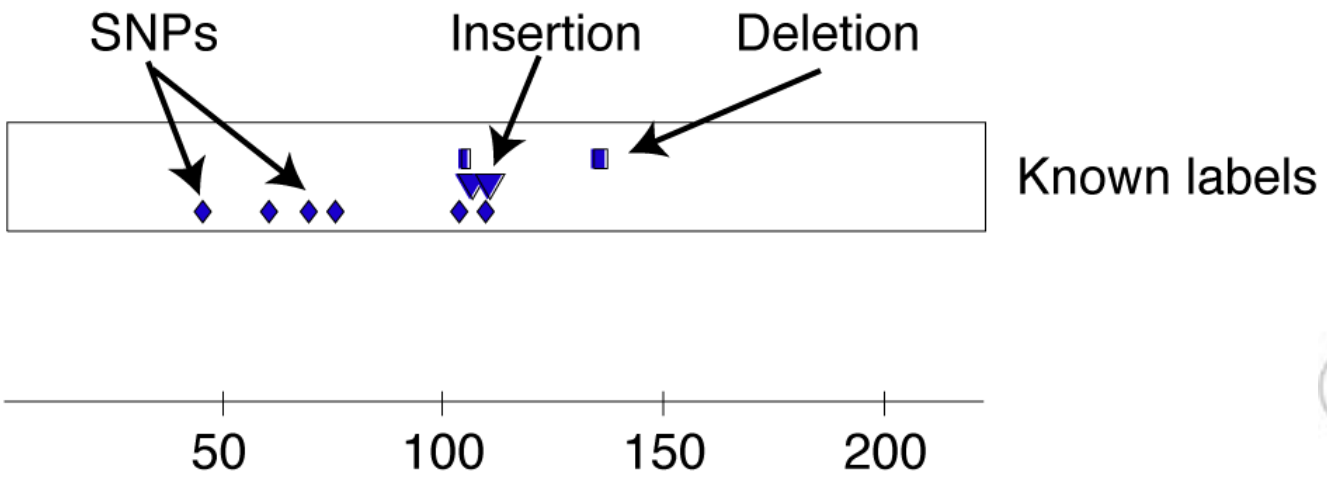
- Performance drops, when other SNPs are in vicinity (1-20nt)
- Least predicted SNPs in highly polymorphic regions!
- ML more sensitive

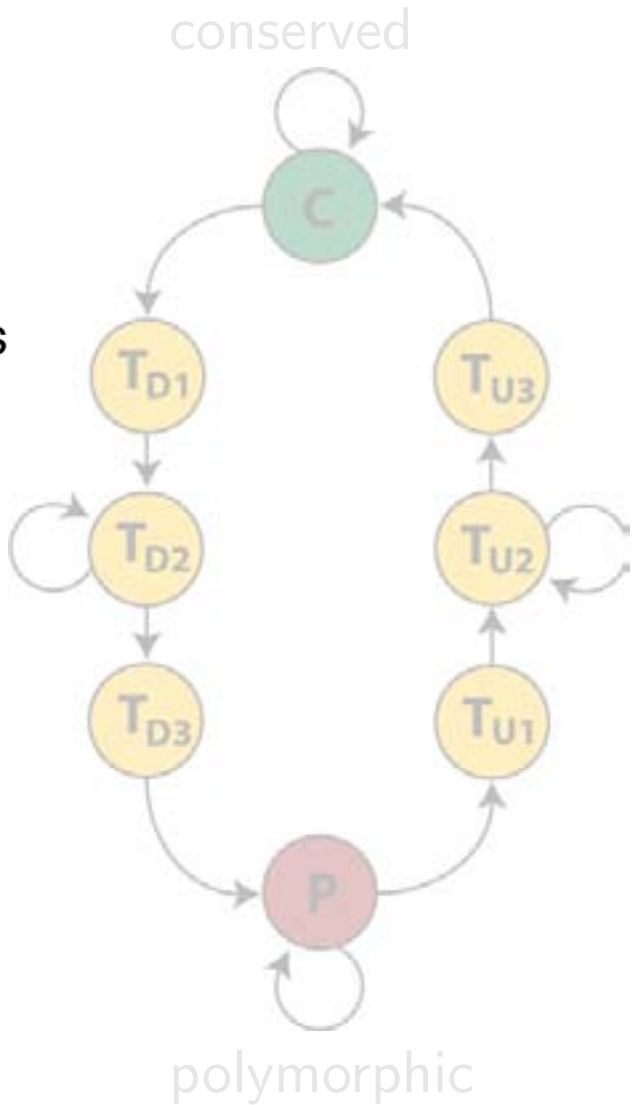
New Approach

- **P**olymorphic **R**egion **P**rediction (PRP)
- Novel ML techniques for predicting complex properties

Modeling Polymorphic Regions

Example sequence from Bor-4





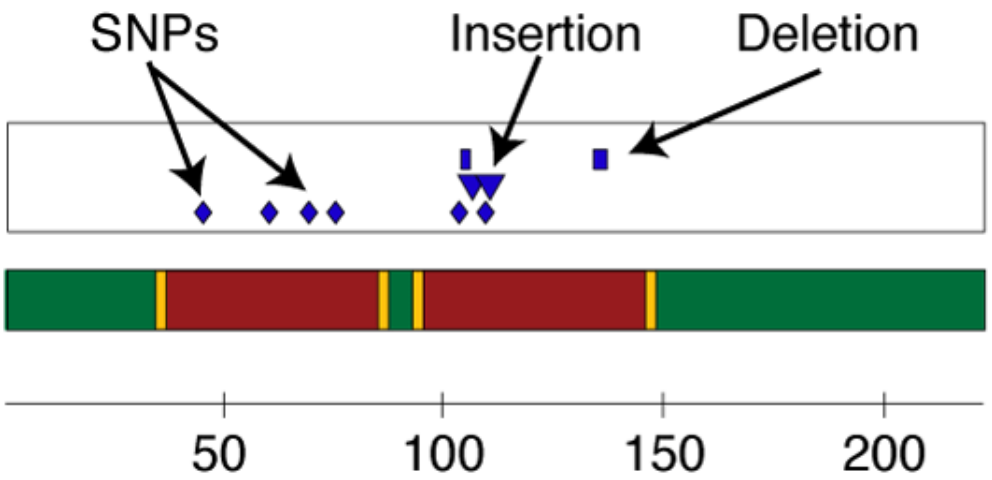
Segmentation of measurement data into

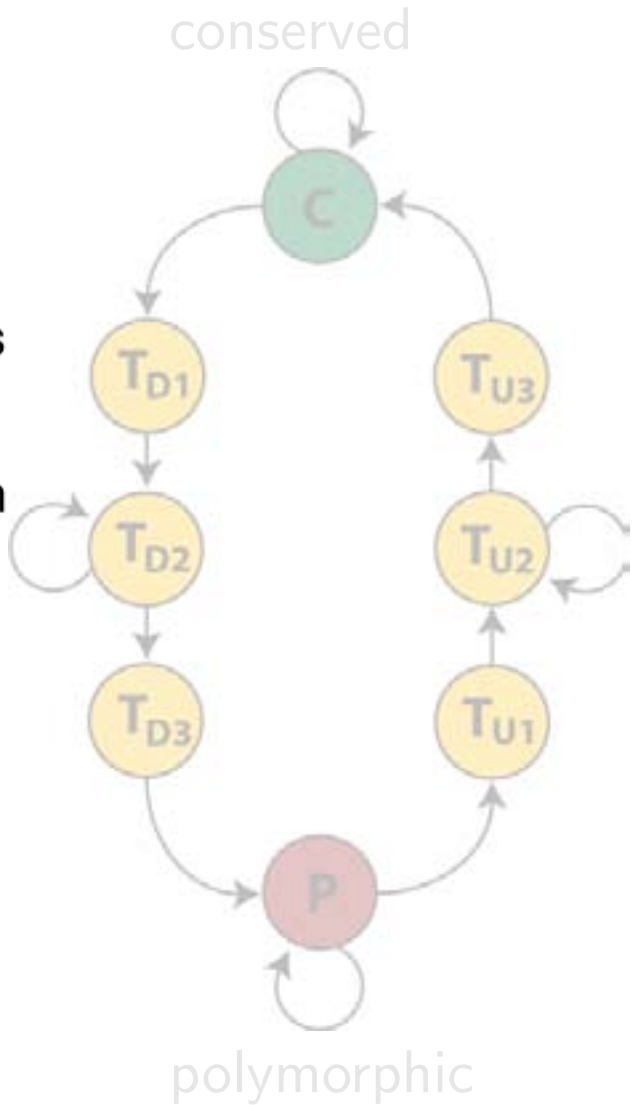
- Conserved and
- Polymorphic regions (distance <25nt)

Predict segmentation using state-model based approach

Modeling Polymorphic Regions

Example sequence from Bor-4





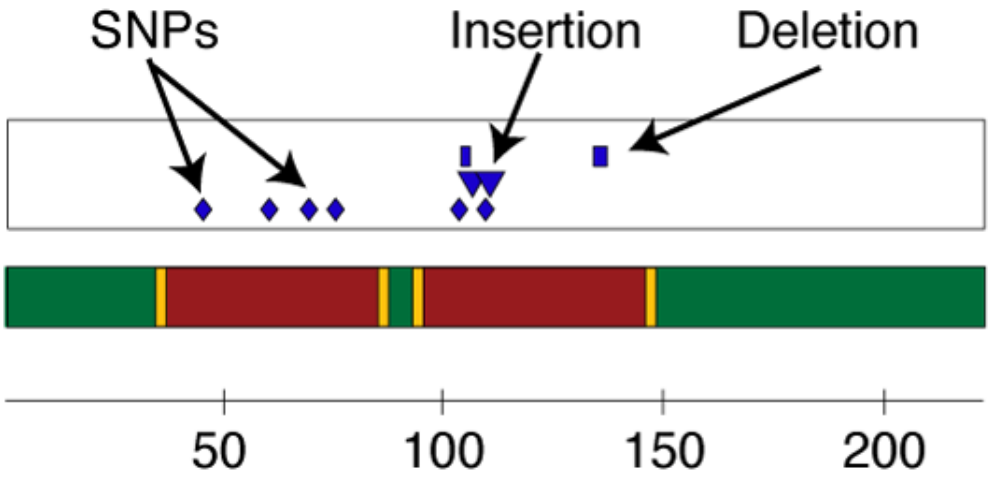
Segmentation of measurement data into

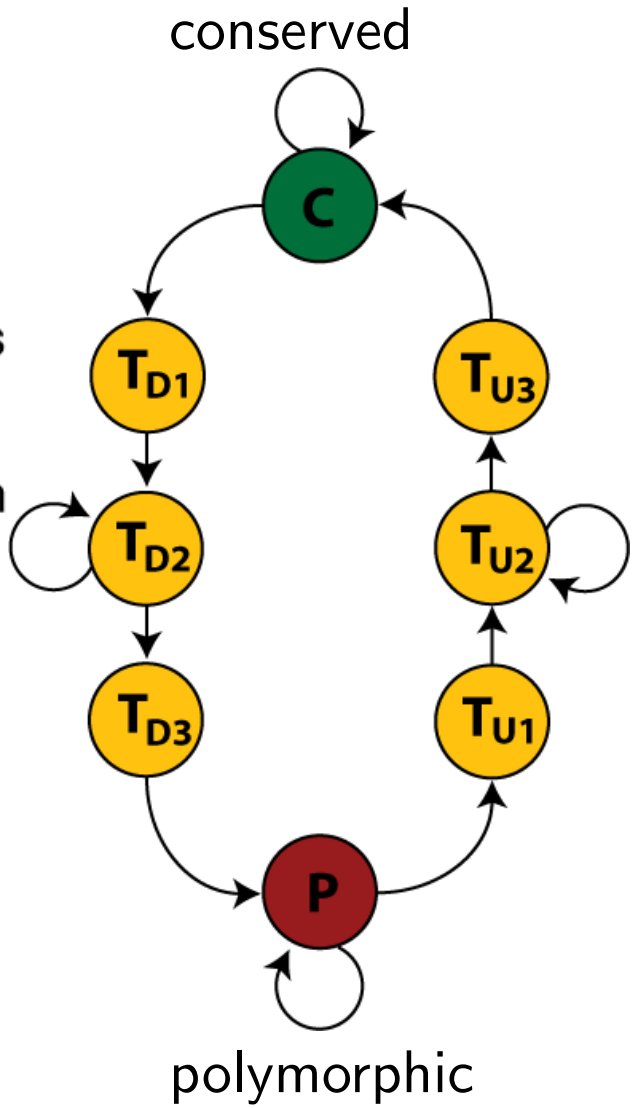
- Conserved and
- Polymorphic regions (distance <25nt)

Predict segmentation using state-model based approach

Modeling Polymorphic Regions

Example sequence from Bor-4





Segmentation of measurement data into

- Conserved and
- Polymorphic regions (distance <25nt)

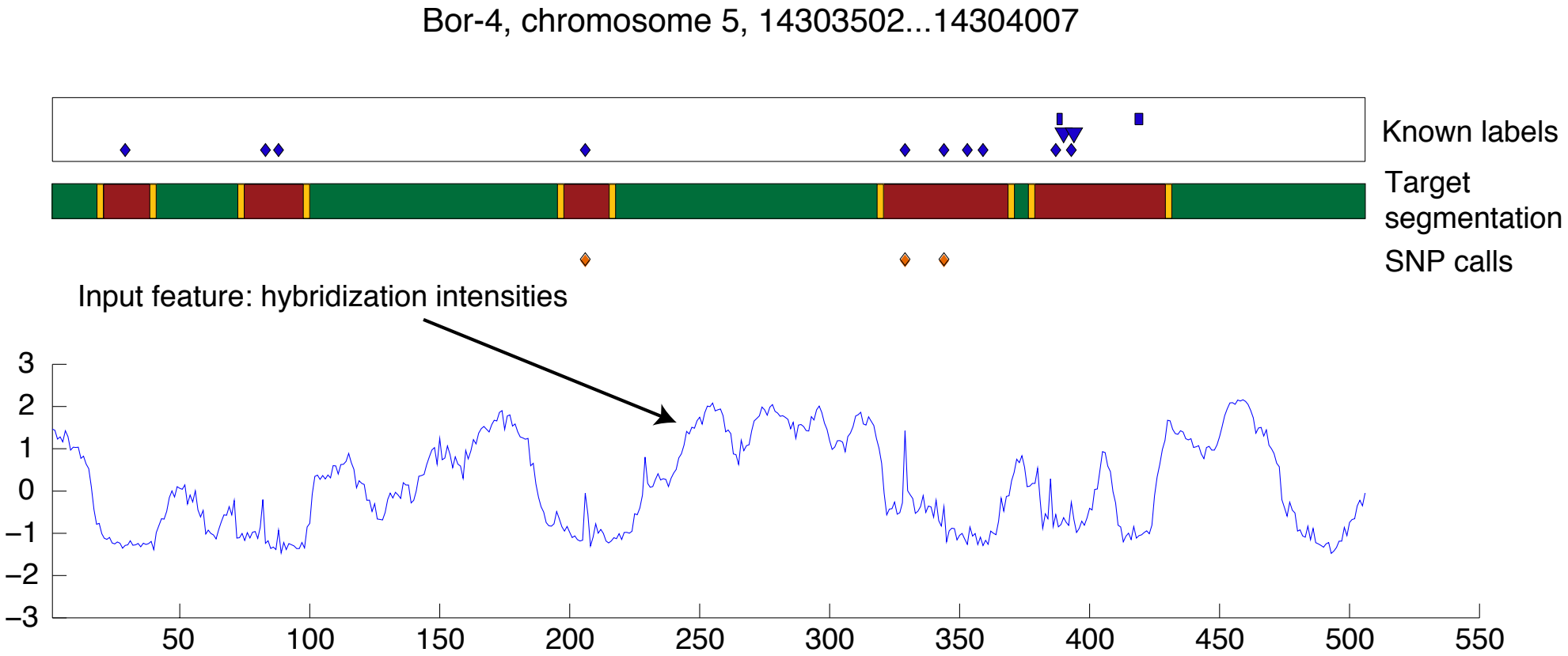
Predict segmentation using state-model based approach

Example



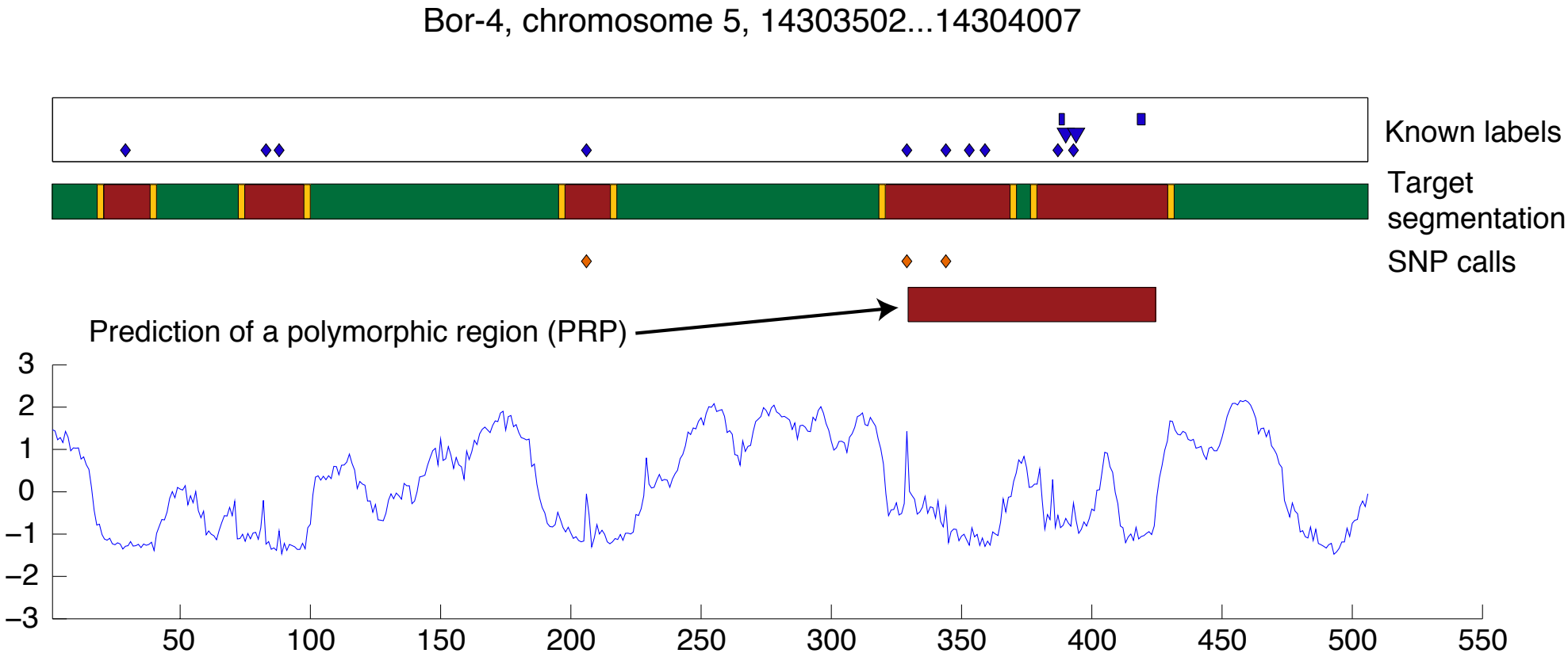
- Count prediction as correct if it overlaps by 75%
- 55% Sensitivity, 90% Specificity (excluding isolated SNPs)
- Previously: Heuristic method with much lower sensitivity

Example



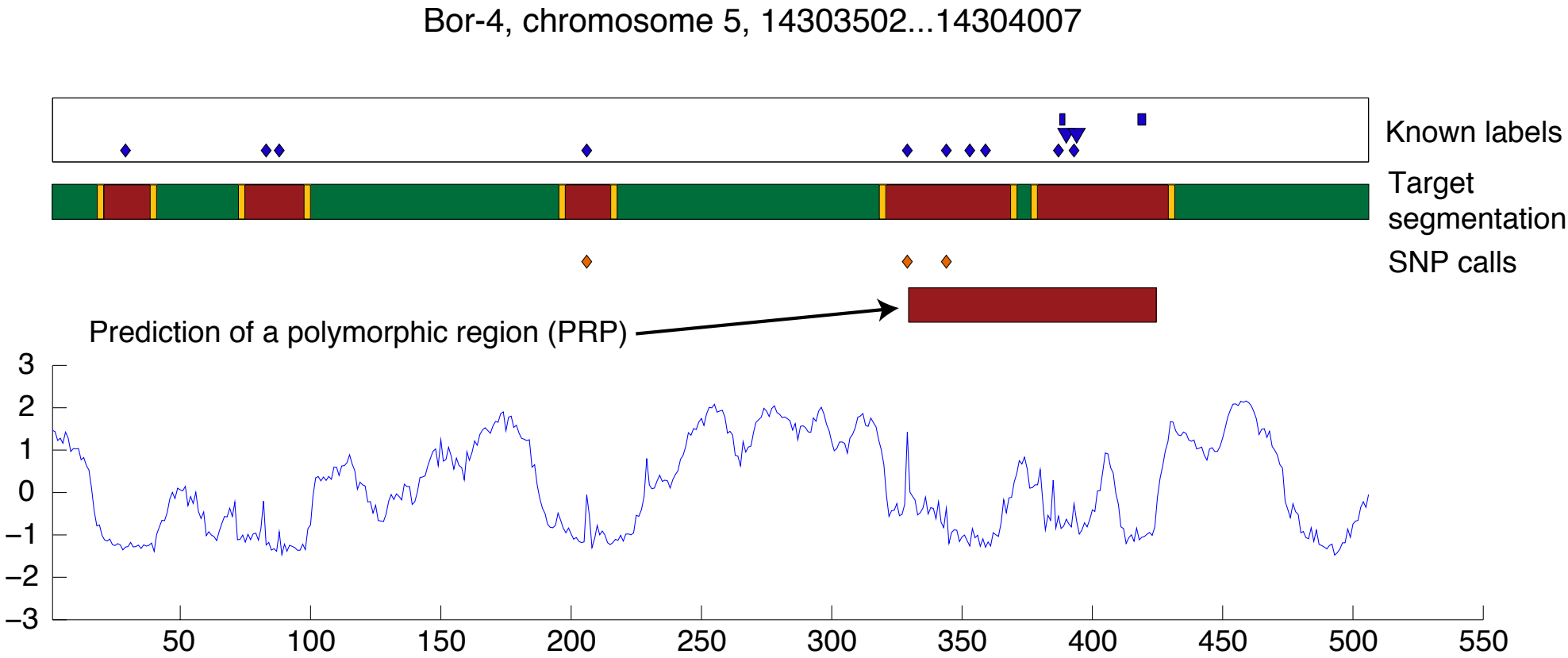
- Count prediction as correct if it overlaps by 75%
- 55% Sensitivity, 90% Specificity (excluding isolated SNPs)
- Previously: Heuristic method with much lower sensitivity

Example



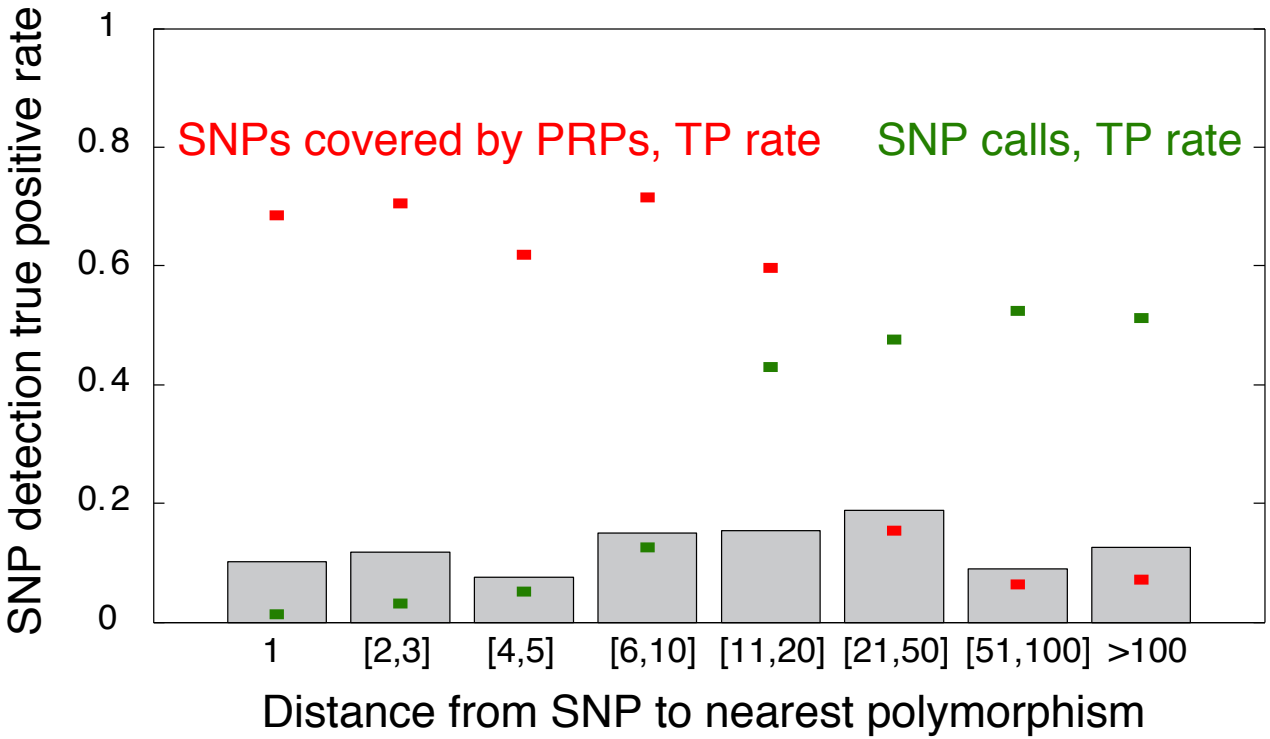
- Count prediction as correct if it overlaps by 75%
- 55% Sensitivity, 90% Specificity (excluding isolated SNPs)
- Previously: Heuristic method with much lower sensitivity

Example



- Count prediction as correct if it overlaps by 75%
- 55% Sensitivity, 90% Specificity (excluding isolated SNPs)
- Previously: Heuristic method with much lower sensitivity

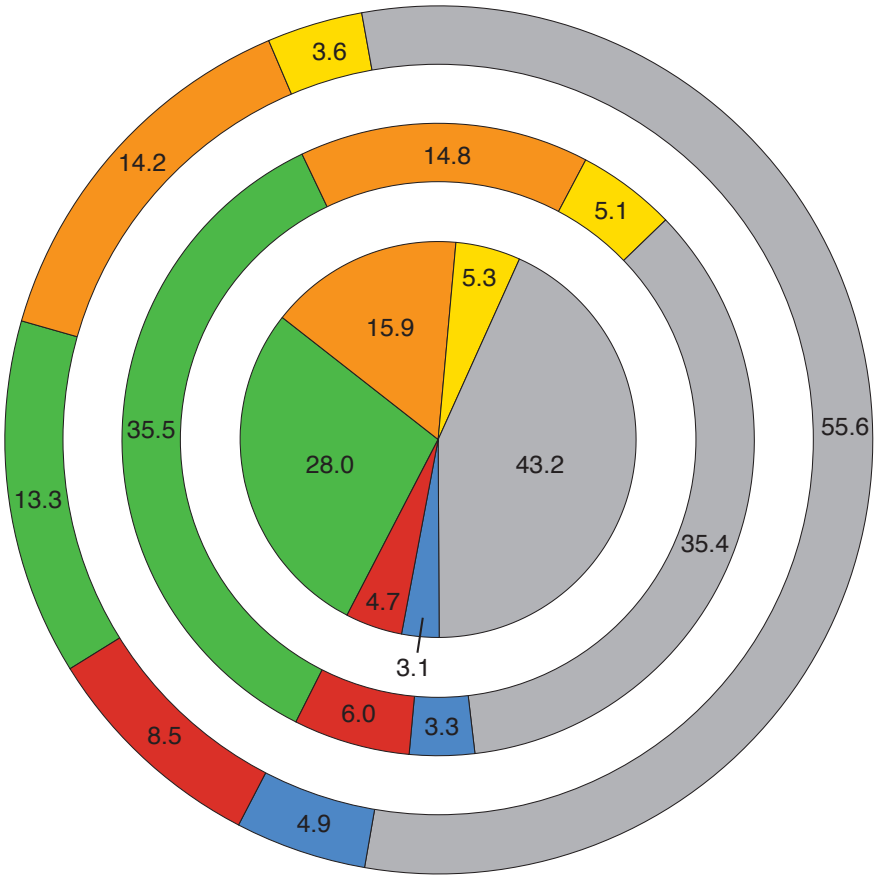
Complementing SNP Calls



Fraction of called/covered polymorphisms (test set):

	SNP calling (MB+ML)	Region predictor
SNPs	∼32%	∼41%
SNPs both methods	∼65%	
Deletions (per base)		∼71%
Insertions		∼39%

Polymorphism Distribution



Sequence type	SNP no.	bp in PSPs
Coding	230,186	3,828,824
Intron	96,069	4,090,306
Untranslated RNA	32,771	1,029,948
Intergenic	229,335	16,046,419
Pseudogene	21,552	1,413,369
Transposon	38,657	2,462,213

Inner circle: Genomic bases
Inner ring: SNP predictions
Outer ring: Bases in PRPs

Coding regions are underrepresented for PRPs while transposons are overrepresented.

Clark et al., Science, Zeller et al., submitted, 2007

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 site dinucl.)
  - 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 site dinucl.)
  - 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 consensus changes

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

Polymorphic Regions Predictions

- Overlap coding regions of 16,692 genes in at least one ecotype
- 50% overlap in 1,910 genes, 95% overlap in 743 genes
- 122 coding sequence deletions validated by dideoxy sequencing*

* Verified in collaboration with laboratory of Joseph Ecker

Effects on Genes

SNPs (MB+ML) leading to consensus changes

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

Polymorphic Regions Predictions

- Overlap coding regions of 16,692 genes in at least one ecotype
- 50% overlap in 1,910 genes, 95% overlap in 743 genes
- 122 coding sequence deletions validated by dideoxy sequencing*

* Verified in collaboration with laboratory of Joseph Ecker

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, 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, 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, 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, human?)
- Study variations using mRNA tiling arrays
  - Expression levels, splicing, ...

Acknowledgments

Friedrich Miescher Laboratory
- Gabriele Schweikert
- Georg Zeller

The Salk Institute, CA
- Paul Shinn
- Joseph Ecker

MPI for Biological Cybernetics
- Bernhard Schölkopf

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

ZBIT, University of Tübingen
- Daniel Huson

Perlegen Sciences, CA
- Kelly Frazer

Overview

1. Introduction to Bioinformatics
   - Basic Biology and Central Dogma
   - Typical Data Types
   - Common Analysis Tasks
2. Sequence Analysis (with SVMs)
   - String Kernels
   - Large Scale Data Structures
   - Heterogeneous Data
3. Structured Output Learning
   - Hidden Markov Models & Dynamic Programming
   - Discriminative Approaches (CRFs & HMSVMs)
   - Large Scale Approaches
4. Some Applications
   - Spliced Alignments (PALMA)
   - Gene Finding (mGene)
   - Analysis of Resequencing Arrays (SNPs and Polymorphic regions)

Summary

- There exist many different data types
  - Structures, sequences, graphs
- Mainly talked about kernels for sequences (check out Jean-Philippe Vert's lecture on graphs, cf. `http://tinyurl.com/3bg7f2`)
- There exist data structures that allow to solve classification problems with 10 million data points
- Often the problem is more than just classification/regression
  - Dealing with incomplete/heterogeneous data
  - Semi-supervised learning
  - Ranking
  - Predicting structures
- Bioinformatics is full of interesting applications
  - Show that your algorithm works on a relevant problem
  - Get ideas of new inference problems
  - Contribute to the solution of important life science problems
- **Thank you for your attention!**

Summary

- There exist many different data types
  - Structures, sequences, graphs
- Mainly talked about kernels for sequences (check out Jean-Philippe Vert's lecture on graphs, cf. `http://tinyurl.com/3bg7f2`)
- There exist data structures that allow to solve classification problems with 10 million data points
- Often the problem is more than just classification/regression
  - Dealing with incomplete/heterogeneous data
  - Semi-supervised learning
  - Ranking
  - Predicting structures
- Bioinformatics is full of interesting applications
  - Show that your algorithm works on a relevant problem
  - Get ideas of new inference problems
  - Contribute to the solution of important life science problems
- **Thank you for your attention!**

Summary

- There exist many different data types
  - Structures, sequences, graphs
- Mainly talked about kernels for sequences (check out Jean-Philippe Vert's lecture on graphs, cf. `http://tinyurl.com/3bg7f2`)
- There exist data structures that allow to solve classification problems with 10 million data points
- Often the problem is more than just classification/regression
  - Dealing with incomplete/heterogeneous data
  - Semi-supervised learning
  - Ranking
  - Predicting structures
- Bioinformatics is full of interesting applications
  - Show that your algorithm works on a relevant problem
  - Get ideas of new inference problems
  - Contribute to the solution of important life science problems
- **Thank you for your attention!**

Summary

- There exist many different data types
  - Structures, sequences, graphs
- Mainly talked about kernels for sequences (check out Jean-Philippe Vert's lecture on graphs, cf. `http://tinyurl.com/3bg7f2`)
- There exist data structures that allow to solve classification problems with 10 million data points
- Often the problem is more than just classification/regression
  - Dealing with incomplete/heterogeneous data
  - Semi-supervised learning
  - Ranking
  - Predicting structures
- Bioinformatics is full of interesting applications
  - Show that your algorithm works on a relevant problem
  - Get ideas of new inference problems
  - Contribute to the solution of important life science problems
- **Thank you for your attention!**

Summary



- There exist many different data types
  - Structures, sequences, graphs
- Mainly talked about kernels for sequences (check out Jean-Philippe Vert's lecture on graphs, cf. `http://tinyurl.com/3bg7f2`)
- There exist data structures that allow to solve classification problems with 10 million data points
- Often the problem is more than just classification/regression
  - Dealing with incomplete/heterogeneous data
  - Semi-supervised learning
  - Ranking
  - Predicting structures
- Bioinformatics is full of interesting applications
  - Show that your algorithm works on a relevant problem
  - Get ideas of new inference problems
  - Contribute to the solution of important life science problems
- **Thank you for your attention!**

Summary

- There exist many different data types
  - Structures, sequences, graphs
- Mainly talked about kernels for sequences (check out Jean-Philippe Vert's lecture on graphs, cf. `http://tinyurl.com/3bg7f2`)
- There exist data structures that allow to solve classification problems with 10 million data points
- Often the problem is more than just classification/regression
  - Dealing with incomplete/heterogeneous data
  - Semi-supervised learning
  - Ranking
  - Predicting structures
- Bioinformatics is full of interesting applications
  - Show that your algorithm works on a relevant problem
  - Get ideas of new inference problems
  - Contribute to the solution of important life science problems
- **Thank you for your attention!**

Summary

- There exist many different data types
  - Structures, sequences, graphs
- Mainly talked about kernels for sequences (check out Jean-Philippe Vert's lecture on graphs, cf. `http://tinyurl.com/3bg7f2`)
- There exist data structures that allow to solve classification problems with 10 million data points
- Often the problem is more than just classification/regression
  - Dealing with incomplete/heterogeneous data
  - Semi-supervised learning
  - Ranking
  - Predicting structures
- Bioinformatics is full of interesting applications
  - Show that your algorithm works on a relevant problem
  - Get ideas of new inference problems
  - Contribute to the solution of important life science problems
- **Thank you for your attention!** Questions!?