Methods for Transcriptome Analysis with Tiling Arrays and mRNA-Seq

Gunnar Rätsch

Friedrich Miescher Laboratory
Max Planck Society
Tübingen, Germany

Talk at the University of Toronto
July 17, 2008
Research Topics

1. Machine learning methods
   ⇒ Develop fast, accurate and interpretable learning methods

2. Genome annotation
   ⇒ Predict features encoded on DNA

3. Biological networks
   ⇒ Understand interactions between gene products

4. Analysis of polymorphisms
   ⇒ Discover polymorphisms and associate them with phenotypes
Research Topics

1. Machine learning methods
   ⇒ Develop fast, accurate and interpretable learning methods

2. Genome annotation
   ⇒ Predict features encoded on DNA

3. Biological networks
   ⇒ Understand interactions between gene products

4. Analysis of polymorphisms
   ⇒ Discover polymorphisms and associate them with phenotypes
Research Topics

1. Machine learning methods
   \(\Rightarrow\) *Develop fast, accurate and interpretable learning methods*

2. Genome annotation
   \(\Rightarrow\) *Predict features encoded on DNA*

3. Biological networks
   \(\Rightarrow\) *Understand interactions between gene products*

4. Analysis of polymorphisms
   \(\Rightarrow\) *Discover polymorphisms and associate them with phenotypes*
Research Topics

1. Machine learning methods
   ⇒ Develop fast, accurate and interpretable learning methods

2. Genome annotation
   ⇒ Predict features encoded on DNA

3. Biological networks
   ⇒ Understand interactions between gene products

4. Analysis of polymorphisms
   ⇒ Discover polymorphisms and associate them with phenotypes
Machine Learning Methods
⇒ Develop fast, accurate and interpretable learning methods

1. Large scale sequence classification
   with Sonnenburg (Fraunhofer, Berlin) & Schölkopf (MPI Biol. Cybernetics)

2. Analysis and explanation of learning result
   with Sonnenburg (Fraunhofer, Berlin)

3. Sequence segmentation
   with Altun (MPI Biol. Cybernetics)

[e.g. Sonnenburg et al., 2007, Rätsch et al., 2006, Rätsch and Sonnenburg, 2007]
Machine Learning Methods
⇒ Develop fast, accurate and interpretable learning methods

1. Large scale sequence classification
   with Sonnenburg (Fraunhofer, Berlin) & Schölkopf (MPI Biol. Cybernetics)

2. Analysis and explanation of learning result
   with Sonnenburg (Fraunhofer, Berlin)

3. Sequence segmentation
   with Altun (MPI Biol. Cybernetics)

[e.g. Sonnenburg et al., 2007, Rätsch et al., 2006, Rätsch and Sonnenburg, 2007]
Machine Learning Methods
⇒ Develop fast, accurate and interpretable learning methods

1. Large scale sequence classification
   with Sonnenburg (Fraunhofer, Berlin) & Schölkopf (MPI Biol. Cybernetics)

2. Analysis and explanation of learning result
   with Sonnenburg (Fraunhofer, Berlin)

3. Sequence segmentation
   with Altun (MPI Biol. Cybernetics)

[e.g. Sonnenburg et al., 2007, Rätsch et al., 2006, Rätsch and Sonnenburg, 2007]
Genome annotation
⇒ Predict features encoded on DNA

1. Ab initio gene finding and prediction of alternative splicing
   1. *C. remanei/briggsae/japonica/brenneri* with Stein (CSHL)
   2. *P. pacificus* with Sommer (MPI Developmental Biology)
   3. Many fungal genomes with Güldener (MIPS)
   4. *V. carteri* with Hallmann (U. Bielefeld)

2. Transcriptome tiling arrays
   with Weigel (MPI Developmental Biology)

3. Alignment methods for short read sequencing
   with Weigel (MPI Developmental Biology)

4. Prediction of RNA subcellular localization and secondary structure
   [e.g. Rätsch et al., 2007, Zeller et al., 2008b, De Bona et al., 2008]
Genome annotation
⇒ Predict features encoded on DNA

1. Ab initio gene finding and prediction of alternative splicing
   - C. remanei/briggsae/japonica/brenneri with Stein (CSHL)
   - P. pacificus with Sommer (MPI Developmental Biology)
   - Many fungal genomes with Güldener (MIPS)
   - V. carteri with Hallmann (U. Bielefeld)
   - Future: A. lyrata, D. melanogaster, D. rerio, human, ... 

2. Transcriptome tiling arrays
   with Weigel (MPI Developmental Biology)

3. Alignment methods for short read sequencing
   with Weigel (MPI Developmental Biology)

4. Prediction of RNA subcellular localization and secondary structure
   [e.g. Rätsch et al., 2007, Zeller et al., 2008b, De Bona et al., 2008]
Genome annotation
⇒ Predict features encoded on DNA

1. Ab initio gene finding and prediction of alternative splicing
   - C. remanei/briggsae/japonica/brenneri with Stein (CSHL)
   - P. pacificus with Sommer (MPI Developmental Biology)
   - Many fungal genomes with Güldener (MIPS)
   - V. carteri with Hallmann (U. Bielefeld)
   - Future: A. lyrata, D. melanogaster, D. rerio, human, ...

2. Transcriptome tiling arrays
   with Weigel (MPI Developmental Biology)

3. Alignment methods for short read sequencing
   with Weigel (MPI Developmental Biology)

4. Prediction of RNA subcellular localization and secondary structure
   [e.g. Rätsch et al., 2007, Zeller et al., 2008b, De Bona et al., 2008]
Genome annotation
⇒ Predict features encoded on DNA

1. Ab initio gene finding and prediction of alternative splicing
   - C. remanei/briggsae/japonica/brenneri with Stein (CSHL)
   - P. pacificus with Sommer (MPI Developmental Biology)
   - Many fungal genomes with Güldener (MIPS)
   - V. carteri with Hallmann (U. Bielefeld)
   - Future: A. lyrata, D. melanogaster, D. rerio, human, ...

2. Transcriptome tiling arrays
   with Weigel (MPI Developmental Biology)

3. Alignment methods for short read sequencing
   with Weigel (MPI Developmental Biology)

4. Prediction of RNA subcellular localization and secondary structure
   [e.g. Rätsch et al., 2007, Zeller et al., 2008b, De Bona et al., 2008]
Biological networks
⇒ Understand interactions between gene products

1. Identification of Transcription factor targets
   with Lohmann (MPI Developmental Biology)

2. Network motif discovery
   with Tsuda (MPI Biol. Cybernetics) and Dittman (MIPS)

3. Future: Quantitative modeling of networks
   [e.g. Georgii et al., 2008, Schultheiss et al., 2008]
Biological networks

⇒ Understand interactions between gene products

1. Identification of Transcription factor targets
   with Lohmann (MPI Developmental Biology)

2. Network motif discovery
   with Tsuda (MPI Biol. Cybernetics) and Dittman (MIPS)

3. Future: Quantitative modeling of networks
   [e.g. Georgii et al., 2008, Schultheiss et al., 2008]
Biological networks
⇒ Understand interactions between gene products

1. Identification of Transcription factor targets
   with Lohmann (MPI Developmental Biology)

2. Network motif discovery
   with Tsuda (MPI Biol. Cybernetics) and Dittman (MIPS)

3. Future: Quantitative modeling of networks
   [e.g. Georgii et al., 2008, Schultheiss et al., 2008]
Array-based resequencing for polymorphism discovery

1. *A. thaliana* with Weigel & Schölkopf (MPI Biol. Cybernetics)
2. *O. sativa* with Rice consortium & Weigel (MPI Devel. Biology)
3. *M. musculus* with Eskin (UCLA)

Future: Genome-wide association studies/environmental effects

1. *A. thaliana* with Weigel (MPI Developmental Biology)
2. Human diseases with Lawrence (U. Manchester) and Tsuda (MPI Biol. Cybernetics)

[e.g. Clark et al., 2007, Zeller et al., 2008a]
Analysis of Polymorphisms
⇒ Predict polymorphisms and associate them with phenotypes

1. Array-based resequencing for polymorphism discovery
   1. *A. thaliana* with Weigel & Schölkopf (MPI Biol. Cybernetics)
   2. *O. sativa* with Rice consortium & Weigel (MPI Devel. Biology)
   3. *M. musculus* with Eskin (UCLA)

2. Future: Genome-wide association studies/environmental effects
   1. *A. thaliana* with Weigel (MPI Developmental Biology)
   2. Human diseases with Lawrence (U. Manchester) and Tsuda (MPI Biol. Cybernetics)

[e.g. Clark et al., 2007, Zeller et al., 2008a]
1 Transcriptome analysis with tiling arrays (50%)
   ⇒ Identification of transcribed regions & alternative splicing

2 Spliced Alignments of Short Reads (40%)
   ⇒ Accurate alignments using side information

3 Gene Finding with Tiling Arrays & mRNA-seq (10%)
   ⇒ Transcriptome measurements improve gene predictions
Tiling Arrays for Transcriptome Analysis

- Whole-genome quantitative measurements
- Cost-effective
  ⇒ Replicates affordable, many tissues / mutants / conditions
- Unbiased
  ⇒ Do not rely on annotations or known cDNAs
Whole-genome quantitative measurements
Cost-effective
  ⇒ Replicates affordable, many tissues / mutants / conditions
Unbiased
  ⇒ Do not rely on annotations or known cDNAs
Tiling Arrays for Transcriptome Analysis

- Whole-genome quantitative measurements
- Cost-effective
  \[\Rightarrow\] Replicates affordable, many tissues / mutants / conditions
- Unbiased
  \[\Rightarrow\] Do not rely on annotations or known cDNAs
Intensities are Noisy Measurements

Systematic bias induced by probe sequence effects
⇒ model effect for normalization
Intensities are Noisy Measurements

Systematic bias induced by probe sequence effects
⇒ model effect for normalization
Intensity Depends on Probe Sequence

Results for the hybridization of polyadenylated RNA root tissue samples from *Arabidopsis thaliana*.

Previously proposed: Sequence Quantile Normalization (SQN)
[Royce et al., 2007]

[Zeller et al., 2008b]
Assume constant transcript intensities $\bar{y}_i$ (median estimate)

Learn intensity deviation from transcript intensity $\delta_i := y_i - \bar{y}_i$

Model effect depending on probe sequence $x_i$ and $y_i$:

$$f(x_i, y_i) \approx \delta_i$$ using quantilized linear regression

[Zeller et al., 2008b]
Assume constant transcript intensities $\overline{y}_i$ (median estimate)

Learn intensity deviation from transcript intensity $\delta_i := y_i - \overline{y}_i$

Model effect depending on probe sequence $x_i$ and $y_i$:
$f(x_i, y_i) \approx \delta_i$ using quantilized linear regression

[Zeller et al., 2008b]
Assume constant transcript intensities $\bar{y}_i$ (median estimate)

Learn intensity deviation from transcript intensity $\delta_i := y_i - \bar{y}_i$

Model effect depending on probe sequence $x_i$ and $y_i$:

$$f(x_i, y_i) \approx \delta_i$$

using quantilized linear regression
Global thresholding of probe intensities
⇒ bi-partition into exonic and intronic/intergenic probes

[Zeller et al., 2008b, Eichner et al., 2008]
Global thresholding of probe intensities
⇒ bi-partition into exonic and intronic/intergenic probes

[Zeller et al., 2008b, Eichner et al., 2008]
Exon/Background Probe Separation

- **Global thresholding**

![Graph showing sensitivity and false positive rate comparison between global thresholding, raw intensity, and transcript-normalized methods.](image)

- **Support Vector Machines (SVMs)** to discriminate exons from introns

Segments typically consist of several probes ⇒ drastically improved separation

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw intensity</td>
<td>0.778</td>
</tr>
<tr>
<td>Transcript-normalized</td>
<td>0.838</td>
</tr>
</tbody>
</table>

[Zeller et al., 2008b, Eichner et al., 2008]
Exon/Background Probe Separation

- Global thresholding

- Support Vector Machines (SVMs) to discriminate exons from introns

Segments typically consist of several probes
⇒ drastically improved separation

[Zeller et al., 2008b, Eichner et al., 2008]
Global thresholding

Support Vector Machines (SVMs) to discriminate exons from introns

Segments typically consist of several probes ⇒ drastically improved separation

[Zeller et al., 2008b, Eichner et al., 2008]
**Goal:** Identify exon/intron segments that show different intensities than *other exons/introns* in at least one analyzed sample.

[Eichner et al., 2008]
**Goal:** Identify exon/intron segments that are differentially spliced in the analyzed samples.

[Eichner et al., 2008]
Alternative vs. Differential Splicing
A Comparison with EST/cDNA-based Information

ROC curves for intron retention

Gene expression

100
90
80
70
60
50
40
30
20
10
0

0 10 20 30 40 50 60 70 80 90 100

Sensitivity [%]
10 20
1 − Specificity [%]
10 20 30 40 50 60 70 80 90 100

Differential splicing
Alternative splicing

Gene expression

high
low

[Eichner et al., 2008]
Tiling Array Segmentation

Goal: Characterize each probe as either intergenic, exonic or intronic

- observed intensity
- annotated exonic
- annotated intronic
- “ideal noise-free intensity”

Margin-based segmentation of tiling array data (mSTAD) extends a segmentation method by Huber et al. [2006]
- very flexible noise model
- accounts for spliced transcripts
- parameters are learned on tiling array data from regions of known transcripts
**Goal:** Characterize each probe as either intergenic, exonic or intronic.

Margin-based segmentation of tiling array data (mSTAD) extends a segmentation method by Huber et al. [2006] with:

- A very flexible noise model
- Accounting for spliced transcripts
- Parameters learned on tiling array data from regions of known transcripts
Tiling Array Segmentation

Goal: Characterize each probe as either intergenic, exonic or intronic

margin-based segmentation of tiling array data (mSTAD) extends a segmentation method by Huber et al. [2006]

- very flexible noise model
- accounts for spliced transcripts
- parameters are learned on tiling array data from regions of known transcripts
Tiling Array Segmentation

Goal: Characterize each probe as either intergenic, exonic or intronic

Learn to associate a state with each probe given its hybridization signal and local context

\[ Q = 20 \text{ discrete expression levels} \]

Use regions around annotated genes (TAIR7) for training.

Similar to GenRate model [Frey et al., 2006]

[Zeller et al., 2008b]
Tiling Array Segmentation

Goal: Characterize each probe as either intergenic, exonic or intronic

Learn to associate a state with each probe given its hybridization signal and local context

\[ \mathbf{Q} = 20 \text{ discrete expression levels} \]

Use regions around annotated genes (TAIR7) for training.

Similar to GenRate model [Frey et al., 2006]

[Zeller et al., 2008b]
Tiling Array Segmentation

Goal: Characterize each probe as either intergenic, exonic or intronic

Learn to associate a state with each probe given its hybridization signal and local context

\[ Q = 20 \text{ discrete expression levels} \]

Use regions around annotated genes (TAIR7) for training.

Similar to GenRate model [Frey et al., 2006]
Tiling Array Segmentation

Goal: Characterize each probe as either intergenic, exonic or intronic

Learn to associate a state with each probe given its hybridization signal and local context

$Q = 20$ discrete expression levels

Use regions around annotated genes (TAIR7) for training.

Similar to GenRate model [Frey et al., 2006] [Zeller et al., 2008b]
Segmentation Accuracy

Sensitivity

Specificity

[Seller et al., 2008b]
Comparison to Affymetrix’s Transfrags

[Laubinger et al., 2008b]
Discovering New Transcripts

- Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.
- 37 of 47 (>75%) RT-PCR validations successful.

http://www.weigelworld.org/resources/microarray/at-tax

[Laubinger et al., 2008b]
Discovering New Transcripts

- Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.
- 37 of 47 (>75%) RT-PCR validations successful.

http://www.weigelworld.org/resources/microarray/at-tax

[Laubinger et al., 2008b]
Discovering New Transcripts

- Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.
- 37 of 47 (>75%) RT-PCR validations successful.

http://www.weigelworld.org/resources/microarray/at-tax

[Laubinger et al., 2008b]
Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.

37 of 47 (>75%) RT-PCR validations successful.

http://www.weigelworld.org/resources/microarray/at-tax

[Laubinger et al., 2008b]
Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.

37 of 47 (>75%) RT-PCR validations successful.

http://www.weigelworld.org/resources/microarray/at-tax [Laubinger et al., 2008b]
Discovering New Transcripts

Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.

37 of 47 (>75%) RT-PCR validations successful.

http://www.weigelworld.org/resources/microarray/at-tax [Laubinger et al., 2008b]
Discovering New Transcripts

Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.

37 of 47 (75%) RT-PCR validations successful.

http://www.weigelworld.org/resources/microarray/at-tax [Laubinger et al., 2008b]
Outlook: Incorporate Sequence Information

- Incorporate sequence-based splice site predictions into **mSTAD**
  ⇒ improved recognition of exon-intron boundaries
  ⇒ no bias against non-coding transcripts

- Use tiling array data as feature for *ab initio* gene finder **mGene**
  ⇒ highly accurate gene predictions for (protein-coding) genes with expression support.
Outlook: Incorporate Sequence Information

- Incorporate sequence-based splice site predictions into mSTAD
  ⇒ improved recognition of exon-intron boundaries
  ⇒ no bias against non-coding transcripts

- Use tiling array data as feature for *ab initio* gene finder mGene
  ⇒ highly accurate gene predictions for (protein-coding) genes with expression support.
Next Generation Sequencing

- Produces huge amounts of data
- Competes with Sanger sequencing and tiling arrays

- Differences to Sanger sequencing:
  - Much faster and cost effective per base
  - Much more and shorter fragments
  - Much more errors

- Genome (re-)sequencing
  - Identification of polymorphisms
  - De novo genome sequencing
  - ...

- Transcriptome sequencing
  - Discovery of new genes
  - Identification of alternative splice forms
  - ...

Gunnar Rätsch (FML, Tübingen)
Next Generation Sequencing

- Produces huge amounts of data
- Competes with Sanger sequencing and tiling arrays

**Differences to Sanger sequencing:**
- Much faster and cost effective per base
- Much more and *shorter* fragments
- Much more *errors*

- Genome (re-)sequencing
  - Identification of polymorphisms
  - *De novo* genome sequencing
  - ...

- Transcriptome sequencing
  - Discovery of new genes
  - Identification of alternative splice forms
  - ...
Next Generation Sequencing

- Produces huge amounts of data
- Competes with Sanger sequencing and tiling arrays

Differences to Sanger sequencing:
- Much faster and cost effective per base
- Much more and shorter fragments
- Much more errors

Genome (re-)sequencing
- Identification of polymorphisms
- De novo genome sequencing
- ...

Transcriptome sequencing
- Discovery of new genes
- Identification of alternative splice forms
- ...
Next Generation Sequencing

- Produces huge amounts of data
- Competes with Sanger sequencing and tiling arrays

Differences to Sanger sequencing:
- Much faster and cost effective per base
- Much more and shorter fragments
- Much more errors

Genome (re-)sequencing
- Identification of polymorphisms
- *De novo* genome sequencing
  - ...

Transcriptome sequencing
- Discovery of new genes
- Identification of alternative splice forms
  - ...
Next Generation Sequencing

- Produces huge amounts of data
- Competes with Sanger sequencing and tiling arrays

Differences to Sanger sequencing:
- Much faster and cost effective per base
- Much more and *shorter* fragments
- Much more *errors*

Genome (re-)sequencing
- Identification of polymorphisms
- *De novo* genome sequencing
- ...

Transcriptome sequencing
- Discovery of new genes
- Identification of alternative splice forms
- ...
Spliced vs. Unspliced Alignments

- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads
Spliced vs. Unspliced Alignments

- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads

...GCAAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT...CCAATGACTGTTG...
Spliced vs. Unspliced Alignments

- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads
Spliced vs. Unspliced Alignments

Challenge

Develop learning method that accurately aligns all reads by appropriately combining the available information.

- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads
Alignment Scoring Function

Classical scoring $f : \Sigma \times \Sigma \rightarrow \mathbb{R}$

Source of Information
- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

<table>
<thead>
<tr>
<th></th>
<th>gap</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>gap</td>
<td>0.33</td>
<td>0.3</td>
<td>0.12</td>
<td>0.3</td>
<td>0.3</td>
<td>0.55</td>
</tr>
<tr>
<td>A</td>
<td>0.31</td>
<td>0.12</td>
<td>0.12</td>
<td>0.3</td>
<td>0.55</td>
<td>0.33</td>
</tr>
<tr>
<td>C</td>
<td>0.44</td>
<td>0.12</td>
<td>0.44</td>
<td>0.3</td>
<td>0.59</td>
<td>0.12</td>
</tr>
<tr>
<td>G</td>
<td>0.13</td>
<td>0.85</td>
<td>0.31</td>
<td>0.33</td>
<td>0.51</td>
<td>0.3</td>
</tr>
<tr>
<td>T</td>
<td>0.55</td>
<td>0.12</td>
<td>0.13</td>
<td>0.12</td>
<td>0.11</td>
<td>0.1</td>
</tr>
<tr>
<td>N</td>
<td>0.12</td>
<td>0.01</td>
<td>0.3</td>
<td>0.12</td>
<td>0.3</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Alignment Scoring Function

Classical scoring $f : \Sigma \times \Sigma \rightarrow \mathbb{R}$

Source of Information
- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

Table:

<table>
<thead>
<tr>
<th></th>
<th>gap</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>gap</td>
<td>0.33</td>
<td>0.3</td>
<td>0.12</td>
<td>0.3</td>
<td>0.3</td>
<td>0.55</td>
</tr>
<tr>
<td>A</td>
<td>0.31</td>
<td>0.12</td>
<td>0.12</td>
<td>0.3</td>
<td>0.55</td>
<td>0.33</td>
</tr>
<tr>
<td>C</td>
<td>0.44</td>
<td>0.12</td>
<td>0.44</td>
<td>0.3</td>
<td>0.59</td>
<td>0.12</td>
</tr>
<tr>
<td>G</td>
<td>0.13</td>
<td>0.85</td>
<td>0.31</td>
<td>0.33</td>
<td>0.51</td>
<td>0.3</td>
</tr>
<tr>
<td>T</td>
<td>0.55</td>
<td>0.12</td>
<td>0.13</td>
<td>0.12</td>
<td>0.11</td>
<td>0.1</td>
</tr>
<tr>
<td>N</td>
<td>0.12</td>
<td>0.01</td>
<td>0.3</td>
<td>0.12</td>
<td>0.3</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Alignment Scoring Function

Classical scoring $f : \Sigma \times \Sigma \rightarrow \mathbb{R}$

Source of Information
- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information
Alignment Scoring Function

Quality scoring $f : (\Sigma \times \mathbb{R}) \times \Sigma \rightarrow \mathbb{R}$

Source of Information
- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information
Alignment Scoring Function

Quality scoring $f : (\Sigma \times \mathbb{R}) \times \Sigma \rightarrow \mathbb{R}$
Solving the Inverse Alignment Problem

- How do we jointly optimize the 336 parameters?
- What are optimal parameters?

Example: three possible alignments
Solving the Inverse Alignment Problem

- How do we jointly optimize the 336 parameters?
- What are optimal parameters?

Example: three possible alignments

Correct Alignment

```
```

Incorrect Alignment

```
...ACGTACACG T A C A C G
```

Incorrect Alignment

```
```

Incorrect Alignment

```
...ACGTACACG
```

Gunnar Rätsch (FML, Tübingen)  Transcriptome Analysis with Arrays and mRNA-Seq  U Toronto, July 17, 2008  24 / 41
1. Correct alignment is **not** highest scoring one
2. Better parameters: now it is highest scoring. Can we do better?
Scoring of the three alignments:

1. Correct alignment is **not** highest scoring one
2. Better parameters: now it is highest scoring. Can we do better?
Scoring of the three alignments:

- Idea: Enforce a margin between correct and incorrect examples
- One has to solve a large quadratic optimization problem
First Experiment

Generate set of artificially spliced reads
- Genomic reads with quality information
- Genome annotation for artificially splicing the reads
- Use 10,000 reads for training and 30,000 for testing

De Bona et al. [2008]
A Pipeline for Efficient Alignments

1 Run-time complexity of alignment $O(m \cdot n)$
2 Many reads will be fully contained in an exon

- Can we find smaller seed regions to align to?
- How do we discriminate between spliced/unspliced reads?

Pipeline Workflow (Example with $\approx 2.6$ million reads)

1 Find seed regions
   ($\approx 4h$ for 2,586,170 reads; 179 reads/second)
2 First run an approximation of the full model
   ($\approx 17min$ for 2,180,858 reads; 417 reads/second)
3 Use the full model for the candidate spliced reads
   ($\approx 8h$ for 441,579; 15 reads/second)

De Bona et al. [2008]
A Pipeline for Efficient Alignments

1. Run-time complexity of alignment $O(m \cdot n)$
2. Many reads will be fully contained in an exon

- Can we find smaller seed regions to align to?
- How do we discriminate between spliced/unspliced reads?

Pipeline Workflow (Example with $\approx 2.6$ million reads)

1. Find seed regions
   ($\approx 4h$ for 2,586,170 reads; 179 reads/second)
2. First run an approximation of the full model
   ($\approx 17min$ for 2,180,858 reads; 417 reads/second)
3. Use the full model for the candidate spliced reads
   ($\approx 8h$ for 441,579; 15 reads/second)

De Bona et al. [2008]
# A Pipeline for Efficient Alignments

1. Run-time complexity of alignment $O(m \cdot n)$
2. Many reads will be fully contained in an exon

- Can we find smaller seed regions to align to?
- How do we discriminate between spliced/unspliced reads?

## Pipeline Workflow (Example with $\approx 2.6$ million reads)

1. Find seed regions
   ($\approx 4h$ for 2,586,170 reads; 179 reads/second)
2. First run an approximation of the full model
   ($\approx 17min$ for 2,180,858 reads; 417 reads/second)
3. Use the full model for the candidate spliced reads
   ($\approx 8h$ for 441,579; 15 reads/second)

De Bona et al. [2008]
So far:

- Adapted to Illumina 1G Genome Analyzer
  ⇒ works similarly for other platforms
- Evaluation on artificially spliced reads
  ⇒ how does it work in the real-world?

Working on:

- Getting it faster
- Include seed-finding in learning
- Methods for constructing splice graphs
Outlook: Methods for NG Sequencing

So far:
- Adapted to Illumina 1G Genome Analyzer
  ⇒ works similarly for other platforms
- Evaluation on artificially spliced reads
  ⇒ how does it work in the real-world?

Working on:
- Getting it faster
- Include seed-finding in learning
- Methods for constructing splice graphs
Given a DNA sequence $x \in \{\text{'A'}, \text{'C'}, \text{'G'}, \text{'T'}\}^L$

Find the correct **label sequence** $y = y_1 y_2 \ldots y_L$

($y_i \in Y = \{\text{'intergenic'}, \text{'exon'}, \text{'intron'}, \ldots \}$)
Standard Approach: HMMs

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
Standard Approach: HMMs

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
States correspond to sequence signals
- Depends on recognition of signals on the DNA
Transitions correspond to segments
- Model length and content of segment
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

We use 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

We use 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

We use Support Vector Machines
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

_We use Support Vector Machines_
Example: Predictions in UCSC Browser
Example: Predictions in UCSC Browser

WormBase Gene Annotations

chr l: 110900 111000 111100 111200 111300 111400 111500 111600 111700 111800 111900 112000 112100 112200

Schweikert et al. [2008]
mGene learns how to combine signal and content predictions for accurate gene structure prediction.

- Based on state-of-the-art machine learning
- May use additional sources of information
- Winner in the nGASP competition (Cat. 1-3)

Example: Predictions in UCSC Browser

Schweikert et al. [2008]
Results of nGASP Competition (Cat. 1)
(Training and Testing on 10% of the C. elegans Genome)
Transcriptome Measurements for Improved Gene Finding

Ideas: Improve mGene by using

1. tiling array measurements as “content”-sensor track
2. base pair read coverage as “content”-sensor track
3. aligned spliced reads as high-confidence intron predictions

So far for A. thaliana: (preliminary)

1. ab initio mGene
   transcript level performance: \( \text{mean}(\text{SN}, \text{SP}) = 74.3\% \)
2. Tiling array measurements in several tissues/conditions
   transcript level performance: \( \text{mean}(\text{SN}, \text{SP}) = 78.3\% \)
3. mRNA-seq (15x) for base pair read coverage
   transcript level performance: \( \text{mean}(\text{SN}, \text{SP}) = 76.5\% \)

Behr et al. [2008]
Transcriptome Measurements for Improved Gene Finding

I n d e a s: Improve mGene by using

1. tiling array measurements as “content”-sensor track
2. base pair read coverage as “content”-sensor track
3. aligned spliced reads as high-confidence intron predictions

S o f a r f o r A. thaliana: (preliminary)

1. \textit{ab initio} mGene
   transcript level performance: \textbf{mean}(SN, SP) = 74.3%
2. Tiling array measurements in several tissues/conditions
   transcript level performance: \textbf{mean}(SN, SP) = 78.3%
3. mRNA-seq (15x) for base pair read coverage
   transcript level performance: \textbf{mean}(SN, SP) = 76.5%

Behr et al. [2008]
Transcriptome Measurements for Improved Gene Finding

Ideas: Improve mGene by using

1. tiling array measurements as “content”-sensor track
2. base pair read coverage as “content”-sensor track
3. aligned spliced reads as high-confidence intron predictions

So far for A. thaliana: (preliminary)

1. *ab initio* mGene
   transcript level performance: \( \text{mean}(SN, SP) = 74.3\% \)
2. Tiling array measurements in several tissues/conditions
   transcript level performance: \( \text{mean}(SN, SP) = 78.3\% \)
3. mRNA-seq (15x) for base pair read coverage
   transcript level performance: \( \text{mean}(SN, SP) = 76.5\% \)

Behr et al. [2008]
Transcriptome Measurements for Improved Gene Finding

**Ideas:** Improve mGene by using

1. tiling array measurements as “content”-sensor track
2. base pair read coverage as “content”-sensor track
3. aligned spliced reads as high-confidence intron predictions

**So far for A. thaliana:** (preliminary)

1. *ab initio* mGene
   transcript level performance: \( \text{mean}(\text{SN}, \text{SP}) = 74.3\% \)
2. Tiling array measurements in several tissues/conditions
   transcript level performance: \( \text{mean}(\text{SN}, \text{SP}) = 78.3\% \)
3. mRNA-seq (15x) for base pair read coverage
   transcript level performance: \( \text{mean}(\text{SN}, \text{SP}) = 76.5\% \)

Behr et al. [2008]
Conclusions

- Analysis of Tiling Array Data
  - Proper normalization helps downstream analyses
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions

- Short Read Alignments
  - Integrates splice site predictions & quality information
  - Novel technique to learn how to combine information

- Transcriptome Measurements
  - Lead to improved gene finding
  - Allow us to validate our assumptions in gene finding
  - Give rise to interesting computational challenges
  - Help to uncover the full complexity of transcriptomes
Conclusions

- Analysis of Tiling Array Data
  - Proper normalization helps downstream analyses
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions

- Short Read Alignments
  - Integrates splice site predictions & quality information
  - Novel technique to learn how to combine information

- Transcriptome Measurements
  - Lead to improved gene finding
  - Allow us to validate our assumptions in gene finding
  - Give rise to interesting computational challenges
  - Help to uncover the full complexity of transcriptomes
Conclusions

- Analysis of Tiling Array Data
  - Proper normalization helps downstream analyses
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions

- Short Read Alignments
  - Integrates splice site predictions & quality information
  - Novel technique to learn how to combine information

- Transcriptome Measurements
  - Lead to improved gene finding
  - Allow us to validate our assumptions in gene finding
  - Give rise to interesting computational challenges
  - Help to uncover the full complexity of transcriptomes
## Acknowledgments

<table>
<thead>
<tr>
<th>Tiling Arrays</th>
<th>Short Read Alignments</th>
<th>Gene Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>• <strong>Georg Zeller</strong> (FML &amp; MPI)</td>
<td>• <strong>Fabio De Bona</strong> (FML)</td>
<td>• <strong>Gabi Schweikert</strong> (FML)</td>
</tr>
<tr>
<td>• <strong>Johannes Eichner</strong> (FML)</td>
<td>• Stephan Ossowskii (MPI)</td>
<td>• <strong>Jonas Behr</strong> (FML)</td>
</tr>
<tr>
<td>• <strong>Sascha Laubinger</strong> (MPI)</td>
<td>• Korbinian Schneeberger (MPI)</td>
<td>• Alex Zien (FML &amp; FIRST)</td>
</tr>
<tr>
<td>• Stefan Henz (MPI)</td>
<td></td>
<td>• Georg Zeller (FML &amp; MPI)</td>
</tr>
<tr>
<td>• Detlef Weigel (MPI)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

More Information

- [http://www.fml.mpg.de](http://www.fml.mpg.de)
- Slides are available online

Thank you!
Acknowledgments

Tiling Arrays
- **Georg Zeller** (FML & MPI)
- **Johannes Eichner** (FML)
- **Sascha Laubinger** (MPI)
- Stefan Henz (MPI)
- Detlef Weigel (MPI)

Short Read Alignments
- **Fabio De Bona** (FML)
- Stephan Ossowski (MPI)
- Korbinian Schneeberger (MPI)

Gene Finding
- **Gabi Schweikert** (FML)
- **Jonas Behr** (FML)
- Alex Zien (FML & FIRST)
- Georg Zeller (FML & MPI)

More Information
- [http://www.fml.mpg.de](http://www.fml.mpg.de)
- Slides are available online

Thank you!
## Acknowledgments

### Tiling Arrays
- **Georg Zeller** (FML & MPI)
- **Johannes Eichner** (FML)
- **Sascha Laubinger** (MPI)
- Stefan Henz (MPI)
- Detlef Weigel (MPI)

### Short Read Alignments
- **Fabio De Bona** (FML)
- Stephan Ossowski (MPI)
- Korbinian Schneeberger (MPI)

### Gene Finding
- **Gabi Schweikert** (FML)
- **Jonas Behr** (FML)
- Alex Zien (FML & FIRST)
- Georg Zeller (FML & MPI)

### More Information
- [http://www.fml.mpg.de](http://www.fml.mpg.de)
- Slides are available online

---

Thank you!


