QPalma - Optimal Spliced Alignments of Short Sequence Reads

Fabio De Bona\textsuperscript{1}
Stephan Ossowski\textsuperscript{2}
Korbinian Schneeberger\textsuperscript{2}
Gunnar Rätsch\textsuperscript{1}

(1) Friedrich Miescher Laboratory of the Max Planck Society
(2) Max Planck Institute for Developmental Biology
Tübingen, Germany

September 25, 2008
Next Generation (NG) sequencing technologies produce huge amounts of sequencing data.

Difference to Sanger sequencing:
- Much cheaper and faster
- Much more and shorter fragments ⇒ “reads”
- Much more errors
Next Generation (NG) sequencing technologies produce huge amounts of sequencing data.

Difference to Sanger sequencing:
- Much cheaper and faster
- Much more and shorter fragments $\Rightarrow$ “reads”
- Much more errors

Genome / transcriptome sequencing

Data can be used via local alignments:
- Discovery of new genes,
- Identification of alternative splice forms, . . .
QPalma’s aim: Accurately Align transcriptome reads to genomic sequences

- Genomic read mapping already challenging
- Transcriptome read mapping is even more difficult:
  - Spliced reads stem from several genomic regions
  - Short reads getting even shorter due to splicing
  - Alignments more error prone

Improve alignments by using more information:
- Accurate splice site models
- Intron length model
- Quality scores model

Idea: Use a machine learning method to infer optimal parameters
QPalma’s aim: Accurately Align transcriptome reads to genomic sequences

- Genomic read mapping already challenging

- Transcriptome read mapping is even more difficult:
  - Spliced reads stem from several genomic regions
  - Short reads getting even shorter due to splicing
    - Alignments more error prone

- Improve alignments by using more information:
  - Accurate splice site models
  - Intron length model
  - Quality scores model
QPalma’s aim: Accurately Align transcriptome reads to genomic sequences

- Genomic read mapping already challenging

- Transcriptome read mapping is even more difficult:
  - Spliced reads stem from several genomic regions
  - Short reads getting even shorter due to splicing
    - Alignments more error prone

- Improve alignments by using more information:
  - Accurate splice site models
  - Intron length model
  - Quality scores model

Idea: Use a **machine learning** method to infer optimal parameters
Spliced vs. Unspliced Alignments

- Find matching region on genome with a few mismatches
Spliced vs. Unspliced Alignments

- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many 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
Extended Smith-Waterman Algorithm

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

Source of Information
- Sequence matches
Extended Smith-Waterman Algorithm

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

Source of Information
- Sequence matches
- Computational splice site predictions
Extended Smith-Waterman Algorithm

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

Source of Information
- Sequence matches
- Computational splice site predictions
- Intron length model
Extended Smith-Waterman Algorithm

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
Extended Smith-Waterman Algorithm

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

<table>
<thead>
<tr>
<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>$f_A()$</td>
<td>$f_C()$</td>
<td>$f_G()$</td>
<td>$f_T()$</td>
</tr>
<tr>
<td>A</td>
<td>0.31</td>
<td>$f_{AA}()$</td>
<td>$f_{AC}()$</td>
<td>$f_{AG}()$</td>
<td>$f_{AT}()$</td>
</tr>
<tr>
<td>C</td>
<td>0.44</td>
<td>$f_{CA}()$</td>
<td>$f_{CC}()$</td>
<td>$f_{CG}()$</td>
<td>$f_{CT}()$</td>
</tr>
<tr>
<td>G</td>
<td>0.13</td>
<td>$f_{GA}()$</td>
<td>$f_{GC}()$</td>
<td>$f_{GG}()$</td>
<td>$f_{GT}()$</td>
</tr>
<tr>
<td>T</td>
<td>0.55</td>
<td>$f_{TA}()$</td>
<td>$f_{TC}()$</td>
<td>$f_{TG}()$</td>
<td>$f_{TT}()$</td>
</tr>
<tr>
<td>N</td>
<td>0.12</td>
<td>$f_{NA}()$</td>
<td>$f_{NC}()$</td>
<td>$f_{NG}()$</td>
<td>$f_{NT}()$</td>
</tr>
</tbody>
</table>
Scoring Parameter Inference

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

How do we jointly optimize 336 parameters?
Correct alignment is not highest scoring one
Correct alignment is highest scoring one
Can we do better?
Use a technique motivated by “large-margin” methods

Idea: Enforce a margin between correct and incorrect examples

One has to solve a huge quadratic problem
How Can We Generate Data for Training?

- How do we obtain true alignments for training QPalma?
- Simulate *realistic* transcriptome reads
How Can We Generate Data for Training?

- How do we obtain true alignments for training QPalma?
- Simulate *realistic* transcriptome reads
  - Consider well-annotated genome: *Arabidopsis thaliana*
  - Use short *genomic* reads (38nt) with natural quality scores
  - Use reference *annotation* (TAIR7)
How Can We Generate Data for Training?

- How do we obtain true alignments for training QPalma?
- Simulate realistic transcriptome reads
  - Consider well-annotated genome: *Arabidopsis thaliana*
  - Use short genomic reads (38nt) with natural quality scores
  - Use reference annotation (TAIR7)
  - Generate artificially spliced reads:

```
...GCAAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT...CCGTAGAATTGACTGTGTGTTG...
TGACCTGACTACTACGTCGTAACGTACACGGTAGCT
CCGTAGAATTGACTGTGTGTTG...
```

⇒ Identify reads overlapping into introns
How Can We Generate Data for Training?

- How do we obtain true alignments for training QPalma?
- Simulate *realistic* transcriptome reads
  - Consider well-annotated genome: *Arabidopsis thaliana*
  - Use short *genomic* reads (38nt) with natural quality scores
  - Use reference *annotation* (TAIR7)
  - Generate artificially spliced reads:

```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT...CCGTAAGAATTGACTGTGTTG...
TGACCTGACTACTACGTCGTAACGTACACG
CCGTAAGAATTGACTGTGTTG...
```

⇒ Remove overlapping parts
How do we obtain true alignments for training QPalma?

Simulate realistic transcriptome reads

- Consider well-annotated genome: *Arabidopsis thaliana*
- Use short genomic reads (38nt) with natural quality scores
- Use reference annotation (TAIR7)
- Generate artificially spliced reads:

```plaintext
GCAAACCAGTCGACGCTCTACTACGTCGTAACGTACACGGTAGCT....CGTAGAATTGACTGTTG...
TGACCTGACTACTACGTCGTAACGTACACGG
AATTGACTGTTG...
```

⇒ Remove overlapping parts
How Can We Generate Data for Training?

- How do we obtain true alignments for training QPalma?
- Simulate *realistic* transcriptome reads
  - Consider well-annotated genome: *Arabidopsis thaliana*
  - Use short *genomic* reads (38nt) with natural quality scores
  - Use reference *annotation* (TAIR7)
  - Generate artificially spliced reads:

```
...GCAAAACCAGTGACCTGACTACTACGTCGTAACGTACACCGTAGCT...CCGTAAGATTGACTGTGTGG...
TGACCTGACTACTACGTCGTAACGTACACGAAATTGACTGTGTGG...
```

⇒ Combine both exonic parts
How Can We Generate Data for Training?

- How do we obtain true alignments for training QPalma?
- Simulate *realistic* transcriptome reads
  - Consider well-annotated genome: *Arabidopsis thaliana*
  - Use short *genomic* reads (38nt) with natural quality scores
  - Use reference *annotation* (TAIR7)
  - Generate artificially spliced reads:

\[
\ldots \text{GCAAAACCAGTGACCTGACTACTACGTCGTAACGTACACCGTAGCT} \ldots \text{CCGTAGAATTGACTGTGTTG} \ldots
\]

\[
\text{TGACCTGACTACTACGTCGTAACGTACACGAAATTGA}
\]

⇒ Truncate to desired length (38nt)
Results on Artificially Spliced Reads

- Given: Short reads and corresponding genomic regions
- Task: Find the correct spliced alignment
Results on Artificially Spliced Reads

- Given: Short reads and corresponding genomic regions
- Task: Find the correct spliced alignment
- Trained on 10,000 artificially spliced genome reads
- Tested on 30,000 spliced reads
Results on Artificially Spliced Reads

- Given: Short reads and corresponding genomic regions
- Task: Find the correct spliced alignment
- Trained on 10,000 artificially spliced genome reads
- Tested on 30,000 spliced reads

<table>
<thead>
<tr>
<th></th>
<th>SmithW</th>
<th>Intron</th>
<th>Intron+Splice</th>
<th>Intron+Splice +Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment Error Rate</td>
<td>14.19%</td>
<td>9.96%</td>
<td>1.94%</td>
<td>1.78%</td>
</tr>
</tbody>
</table>
Error Rate Depends on Intron Position

- Trust introns confirmed by spliced read with $\geq 6$nt overlap
- Pure Smith-Waterman algorithm would need longer overlaps and would still perform worse
Error Rate Depends on Intron Position

1. Trust introns confirmed by spliced read with $\geq 6$ nt overlap
Error Rate Depends on Intron Position

1. Trust introns confirmed by spliced read with \( \geq 6 \) nt overlap

2. Pure Smith-Waterman algorithm would need longer overlaps and would still perform worse
Can We Use QPalma for Whole-Genome Alignments?

So far we had two assumptions:

1. All reads are spliced
2. Genomic region is known
Can We Use QPalma for Whole-Genome Alignments?

So far we had two assumptions:

1. All reads are spliced
2. Genomic region is known

1. Many reads will be fully contained in an exon
2. Direct alignment to genome too expensive ($O(n \cdot m)$)
A Pipeline for Efficient Large Scale Alignment

**Pipeline Workflow**

1. Map unspliced reads & find seed regions
   - Use suffix-array based method to find short read match with at most two mismatches (*Vmatch*, Kurtz et al.; *ShoRe*, Ossowski et al. 2008)
# A Pipeline for Efficient Large Scale Alignment

## Pipeline Workflow

1. **Map unspliced reads & find seed regions**
   - Use suffix-array based method to find short read match with at most two mismatches (*Vmatch*, Kurtz et al.; *ShoRe*, Ossowski et al. 2008)

2. **Recover potentially spliced reads from first mapping**
   - Use fast approximation of QPalma to decide which reads may be spliced (even if they can be mapped well)
A Pipeline for Efficient Large Scale Alignment

Pipeline Workflow

1. Map unspliced reads & find seed regions
   - Use suffix-array based method to find short read match with at most two mismatches (*Vmatch*, Kurtz et al.; *ShoRe*, Ossowski et al. 2008)

2. Recover potentially spliced reads from first mapping
   - Use fast approximation of QPalma to decide which reads may be spliced (even if they can be mapped well)

3. Identify accurate alignments for the candidate spliced reads
   - Use full QPalma model to align remaining reads
Pipeline Workflow

1. Map unspliced reads & find seed regions
   ▶ Use suffix-array based method to find short read match with at most two mismatches (\textit{Vmatch}, Kurtz et al.; \textit{ShoRe}, Ossowski et al. 2008)
   ▶ \(\approx 4\text{h for } 2,6 \times 10^6 \text{ reads}\)

2. Recover potentially spliced reads from first mapping
   ▶ Use fast approximation of QPalma to decide which reads may be spliced (even if they can be mapped well)
   ▶ \(\approx 17\text{min for } 2,2 \times 10^6 \text{ reads}\)

3. Identify accurate alignments for the candidate spliced reads
   ▶ Use full QPalma model to align remaining reads
   ▶ \(\approx 8\text{h for } \sim 442,000 \text{ reads}\)
Application to Natural Transcriptome Reads

RNA-seq reads for *A. thaliana* (provided by Weigel lab, MPI Devel. Biology)

- 4 lanes from Illumina Genome Analyzer 1G ($\approx 50 \times$ coverage)
- 38nt reads, polyA enriched
- Strand unspecific, young leaves
Application to Natural Transcriptome Reads

RNA-seq reads for *A. thaliana* (provided by Weigel lab, MPI Devel. Biology)

- 4 lanes from Illumina Genome Analyzer 1G ($\approx 50\times$ coverage)
- 38nt reads, polyA enriched
- Strand unspecific, young leaves

Initial Read mapping by ShoRe (http://1001genomes.org)
Spliced alignment by QPalma
(http://fml.mpg.de/raetsch/projects/qpalma)
Application to Natural Transcriptome Reads

RNA-seq reads for *A. thaliana* (provided by Weigel lab, MPI Devel. Biology)

- 4 lanes from Illumina Genome Analyzer 1G ($\approx 50 \times$ coverage)
- 38nt reads, polyA enriched
- Strand unspecific, young leaves
- Initial Read mapping by ShoRe (http://1001genomes.org)
- Spliced alignment by QPalma
  (http://fml.mpg.de/raetsch/projects/qpalma)

⇒ $\approx 30$ million unspliced
⇒ $\approx 1$ million spliced reads
What Can We Do With It?

For instance:

- Assemble alignments to obtain splice graphs *(describing alternative splicing)*
  - Only works well for highly covered genes
What Can We Do With It?

For instance:

- **Assemble alignments to obtain splice graphs** *(describing alternative splicing)*
  - Only works well for highly covered genes

- **Use exon/intron coverage tracks to improve gene finding**
  - Considerable improvements appear possible
For instance:

- **Assemble alignments to obtain splice graphs** (describing alternative splicing)
  - Only works well for highly covered genes

- **Use exon/intron coverage tracks to improve gene finding**
  - Considerable improvements appear possible

- **Estimate relative abundances of alternative transcripts**
  - Spliced reads help as they can connect exons
Summary

- New method for accurate *de novo* spliced alignments of short reads
- Challenging data: many short & error prone reads
Summary

- New method for accurate *de novo* spliced alignments of short reads
- Challenging data: many short & error prone reads
- Proposed algorithm for parameter estimation
- Successfully combined many information sources
Summary

- New method for accurate *de novo* spliced alignments of short reads
- Challenging data: many short & error prone reads
- Proposed algorithm for parameter estimation
- Successfully combined many information sources
- Can be applied to data from other sequencing platforms (ABI, …)
- Fast enough for whole genome/transcriptome sequencing
  - Currently, very good for smaller genomes
  - When excluding indels, then also fast for mouse/human (in progress)
Summary

- New method for accurate *de novo* spliced alignments of short reads
- Challenging data: many short & error prone reads
- Proposed algorithm for parameter estimation
- Successfully combined many information sources
- Can be applied to data from other sequencing platforms (ABI, …)
- Fast enough for whole genome/transcriptome sequencing
  - Currently, very good for smaller genomes
  - When excluding indels, then also fast for mouse/human (in progress)

Potential improvements

- Speed-up dynamic program
- Integrated seed region finding and spliced alignments
Acknowledgements

Coauthors
- Stephan Ossowski
- Korbinian Schneeberger
- Gunnar Rätsch

Weigel Lab – DNA- and RNA-seq Data
- Jun Cao
- Richard Clark
- Christa Lanz
- Detlef Weigel

Stipend
- Friedrich Miescher Laboratory of the Max Planck Society
Thank you!

fabio@tuebingen.mpg.de

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

Poster D19
Results on Artificially Spliced Reads

- Given: Short reads and corresponding genomic regions
- Task: Find the correct spliced alignment
Results on Artificially Spliced Reads

- Given: Short reads and corresponding genomic regions
- Task: Find the correct spliced alignment
- Trained on 10,000 artificially spliced genome reads
- Tested on 30,000 spliced reads
Results on Artificially Spliced Reads

- Given: Short reads and corresponding genomic regions
- Task: Find the correct spliced alignment
- Trained on 10,000 artificially spliced genome reads
- Tested on 30,000 spliced reads

<table>
<thead>
<tr>
<th>Quality information</th>
<th>Splice site pred.</th>
<th>Intron length</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>14.19 %</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>-</td>
<td>13.49 %</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>+</td>
<td>9.96 %</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>9.68 %</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>3.16 %</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>-</td>
<td>2.81 %</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>+</td>
<td>1.94 %</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1.78 %</td>
</tr>
</tbody>
</table>

Fabio De Bona (FML, Tübingen)