QPalma - Optimal Spliced Alignments of Short Sequence Reads

Fabio De Bona[1]
Stephan Ossowski[2]
Korbinian Schneeberger[2]
Gunnar Rätsch[1]
(1) Friedrich Miescher Laboratory of the Max Planck Society
(2) Max Planck Institute for Developmental Biology
Tübingen, Germany

July 18, 2008

Transcriptome Analysis with NG Sequencing

- Techniques are promising for:
  - Discovery of new genes,
  - Identification of alternative splice forms, ...
- 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

  $\Rightarrow$ Alignments more error prone
- Improve alignments by using more information:
  - Accurate splice site models
  - Intron length model
  - Quality scores model

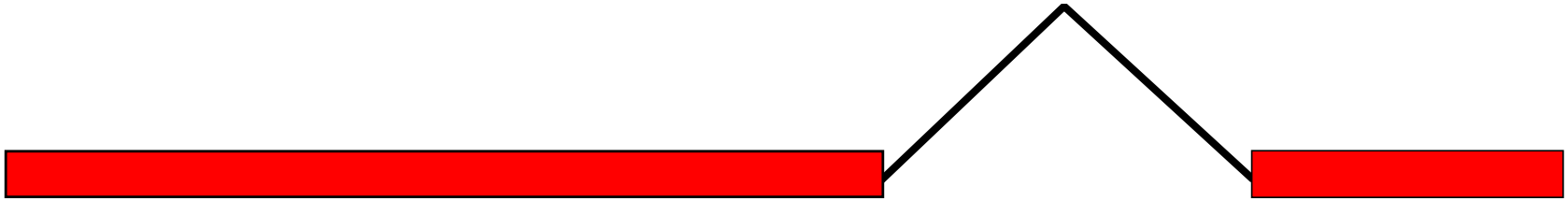
Transcriptome Analysis with NG Sequencing

- Techniques are promising for:
  - Discovery of new genes,
  - Identification of alternative splice forms, ...
- 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

  $\Rightarrow$ 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

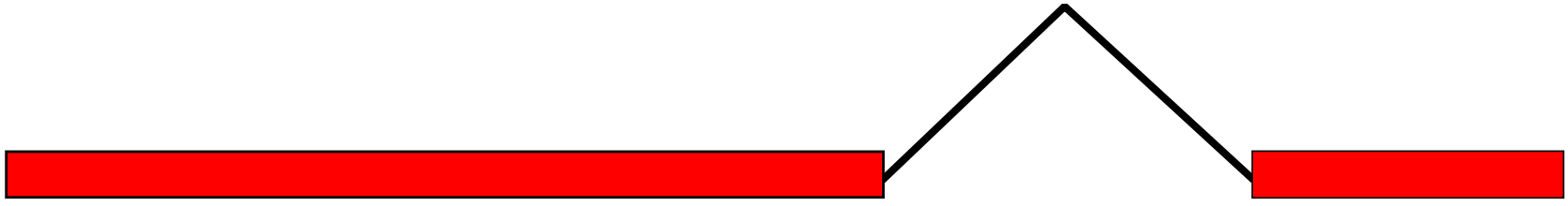
Spliced vs. Unspliced Alignments



```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT....CCGTAGAATTGACTGTGTTG...
   GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC
    CAAACCAGTGACCTGACTACTACGTCGTAACGTACA
     AAACCAGTGACCTGACTACTACGTCGTAACGTACAC
      AACCAGTGACCTGACTACTACGTCGTAACGTACACG
```

- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads

Spliced vs. Unspliced Alignments



```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT....CCGTAGAATTGACTGTGTTG...
   GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC
    CAAACCAGTGACCTGACTACTACGTCGTAACGTACA
     AAACCAGTGACCTGACTACTACGTCGTAACGTACAC
      AACCAGTGACCTGACTACTACGTCGTAACGTACACG
       ACCAGTGACCTGACTACTACGTCGTAACGTACACG ———————————— A
        CCAGTGACCTGACTACTACGTCGTAACGTACACG ———————————— AA
         CAGTGACCTGACTACTACGTCGTAACGTACACG ———————————— AAT
          AGTGACCTGACTACTACGTCGTAACGTACACG ———————————— AATT
           GTGACCTGACTACTACGTCGTAACGTACACG ———————————— AATTG
            TGACCTGACTACTACGTCGTAACGTACACG ———————————— AATTGA
```

- 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

A Pipeline for Efficient Large Scale Alignment

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

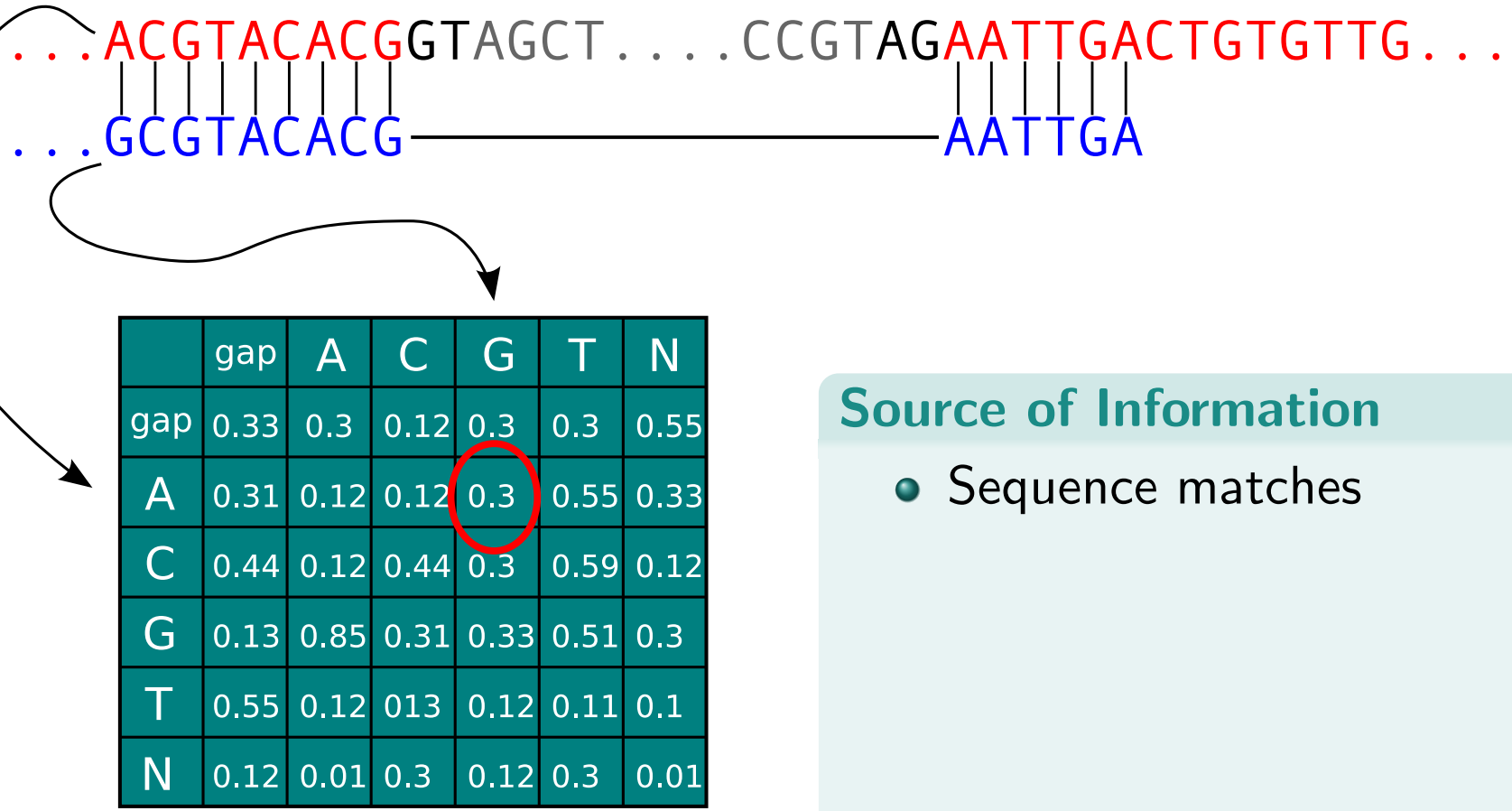
A Pipeline for Efficient Large Scale Alignment

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

Pipeline Workflow

1. Find seed regions
   $\approx$ 4h for $2,6 \times 10^6$ reads
2. First run an approximation of the full model
   $\approx$ 17min for $2,2 \times 10^6$ reads
3. Use the full model for the candidate spliced reads
   $\approx$ 8h for $\sim 442,000$ reads

Extended Smith-Waterman Algorithm



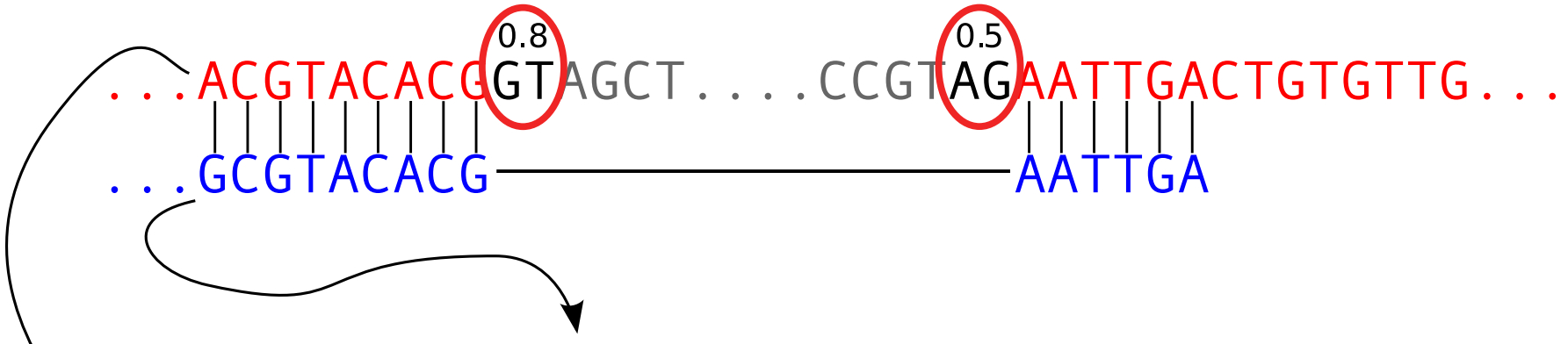
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

- Sequence matches

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

Extended Smith-Waterman Algorithm



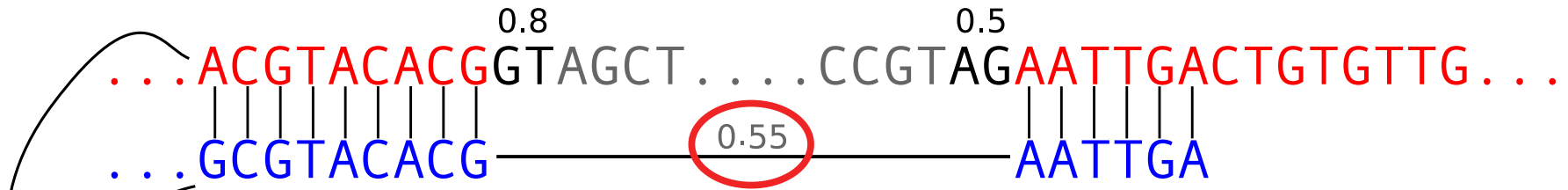
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

- Sequence matches
- Computational splice site predictions

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

Extended Smith-Waterman Algorithm



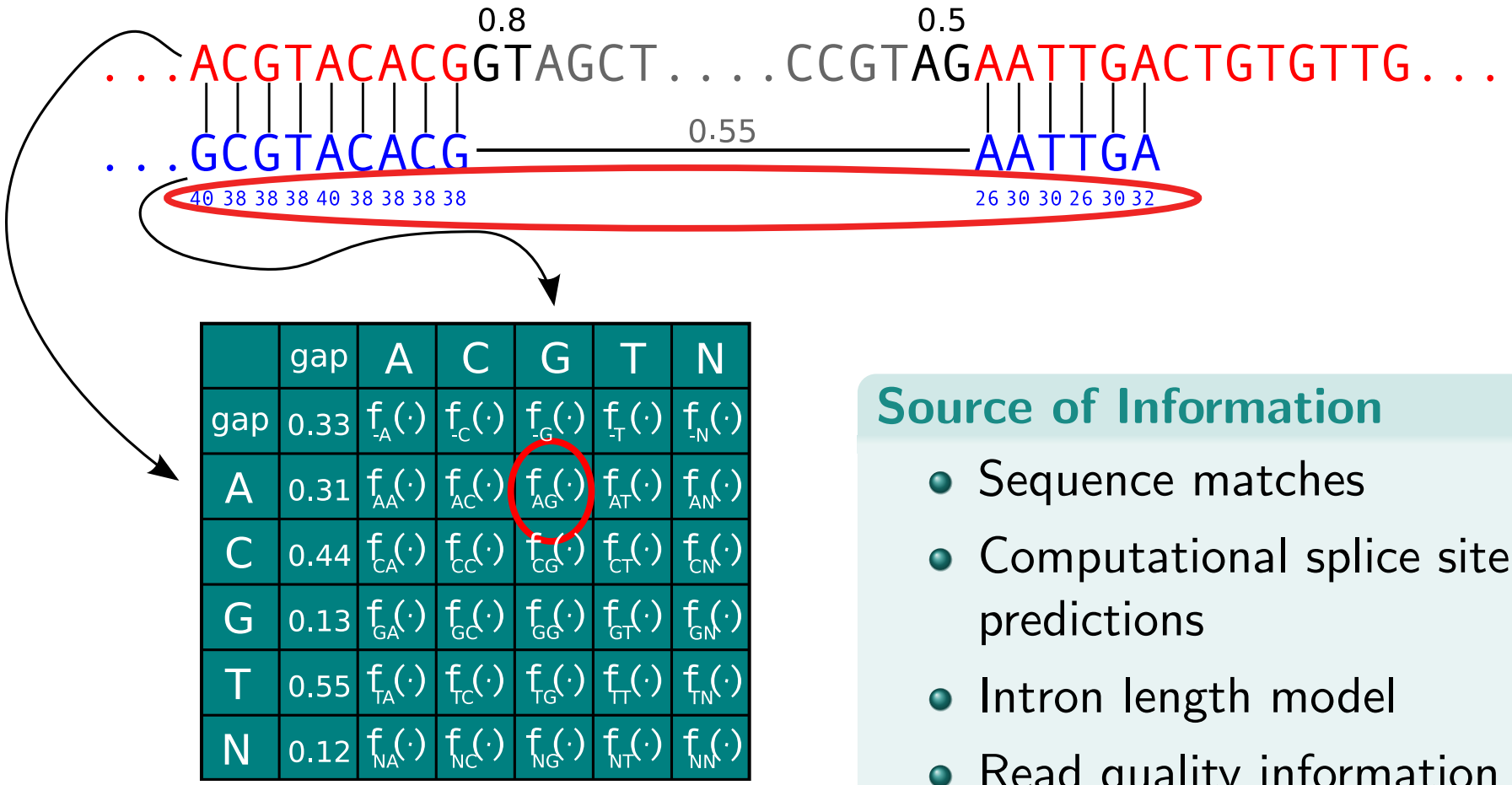
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

- Sequence matches
- Computational splice site predictions
- Intron length model

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

Extended Smith-Waterman Algorithm



Source of Information

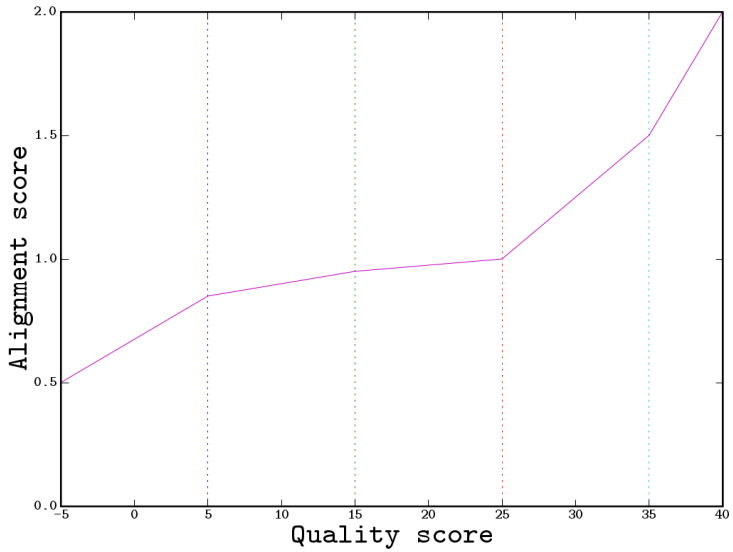
- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

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

Extended Smith-Waterman Algorithm



	gap	A	C	G	T	N
gap	0.33	$f_{-A}(\cdot)$	$f_{-C}(\cdot)$	$f_{-G}(\cdot)$	$f_{-T}(\cdot)$	$f_{-N}(\cdot)$
A	0.31	$f_{AA}(\cdot)$	$f_{AC}(\cdot)$	$f_{AG}(\cdot)$	$f_{AT}(\cdot)$	$f_{AN}(\cdot)$
C	0.44	$f_{CA}(\cdot)$	$f_{CC}(\cdot)$	$f_{CG}(\cdot)$	$f_{CT}(\cdot)$	$f_{CN}(\cdot)$
G	0.13	$f_{GA}(\cdot)$	$f_{GC}(\cdot)$	$f_{GG}(\cdot)$	$f_{GT}(\cdot)$	$f_{GN}(\cdot)$
T	0.55	$f_{TA}(\cdot)$	$f_{TC}(\cdot)$	$f_{TG}(\cdot)$	$f_{TT}(\cdot)$	$f_{TN}(\cdot)$
N	0.12	$f_{NA}(\cdot)$	$f_{NC}(\cdot)$	$f_{NG}(\cdot)$	$f_{NT}(\cdot)$	$f_{NN}(\cdot)$



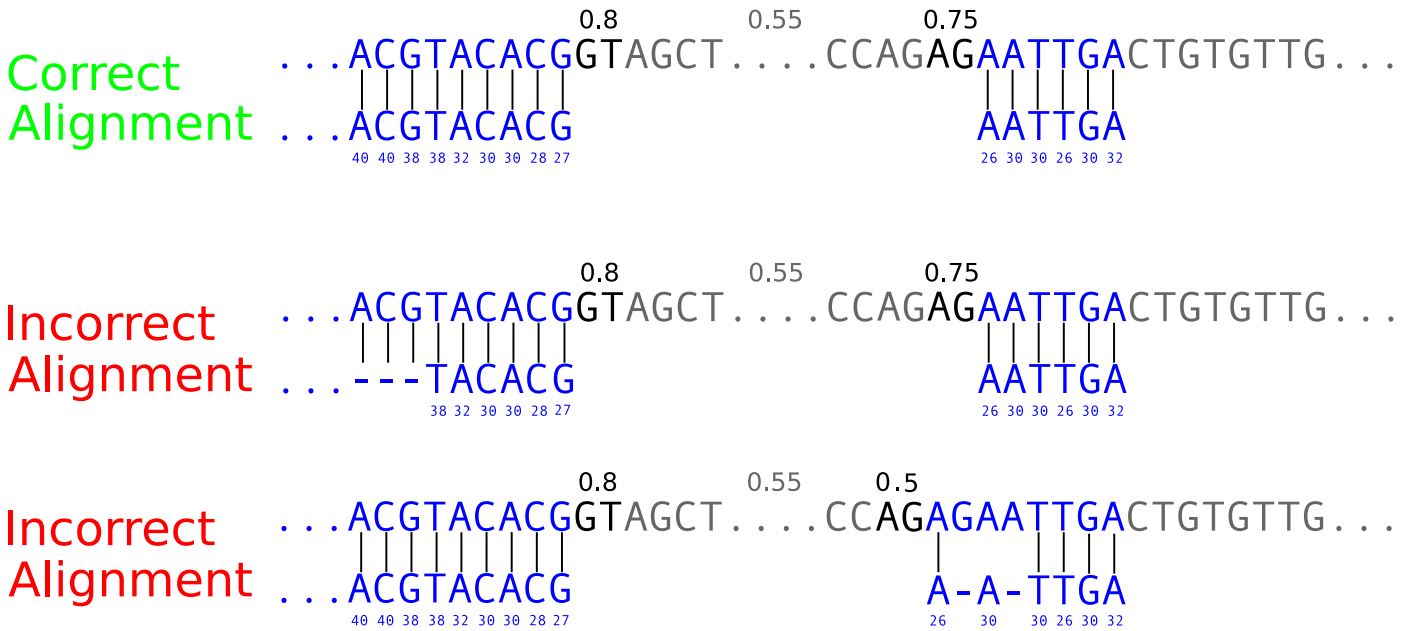
Quality scoring $f : (\Sigma \times \mathbb{R}) \times \Sigma \to \mathbb{R}$

Scoring Parameter Inference

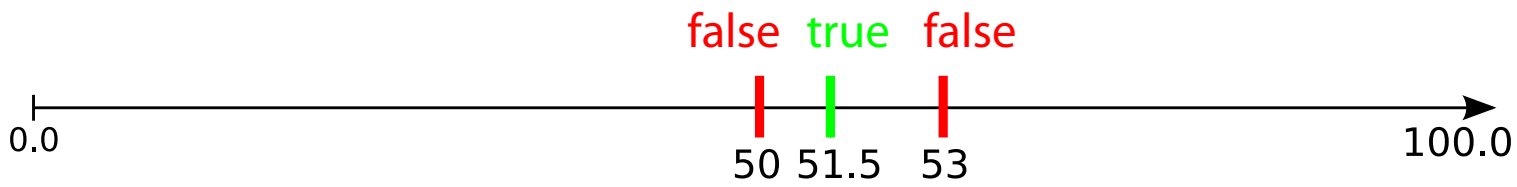
- What are optimal parameters?
- How do we jointly optimize 336 parameters?

Scoring Parameter Inference

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

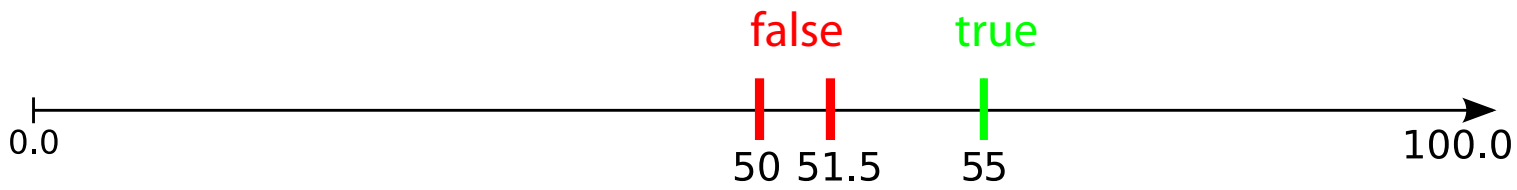


Cartoon



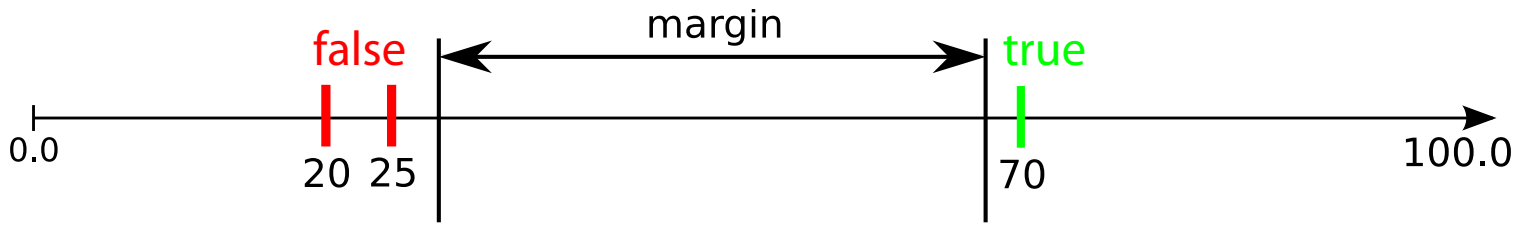
- Correct alignment is **not** highest scoring one

Cartoon



- Can we do better?

Cartoon



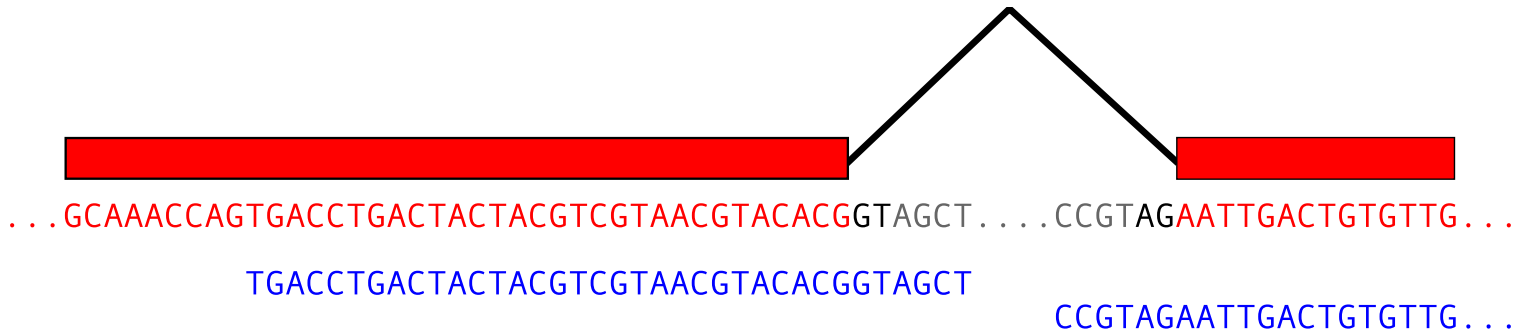
- Use a technique motivated by "large-margin" methods
- Idea: Enforce a margin between correct and incorrect examples
- One has to solve a huge quadratic problem

What is the training data?

- How do we obtain correct alignments?
- Use genomic sequencing data and annotation

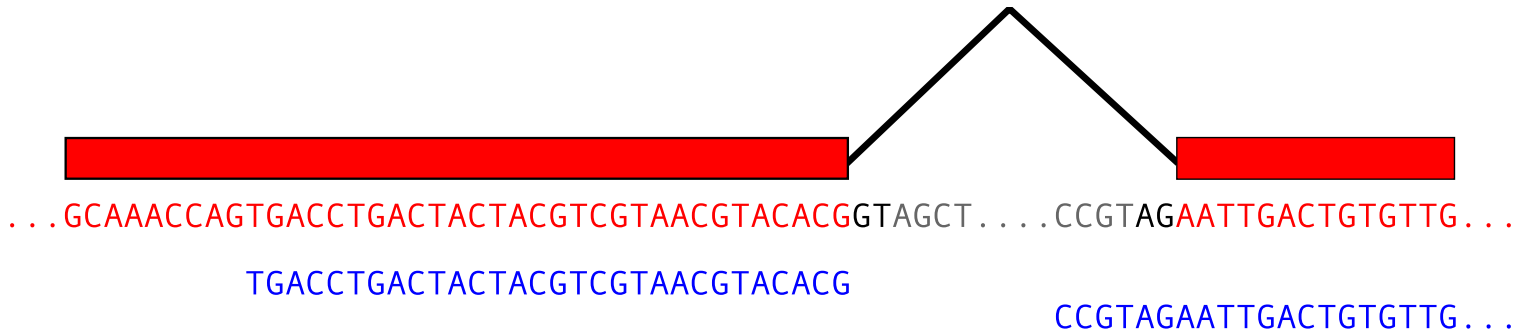
$\Rightarrow$ Simulate transcriptome data

Generation of artificially spliced reads



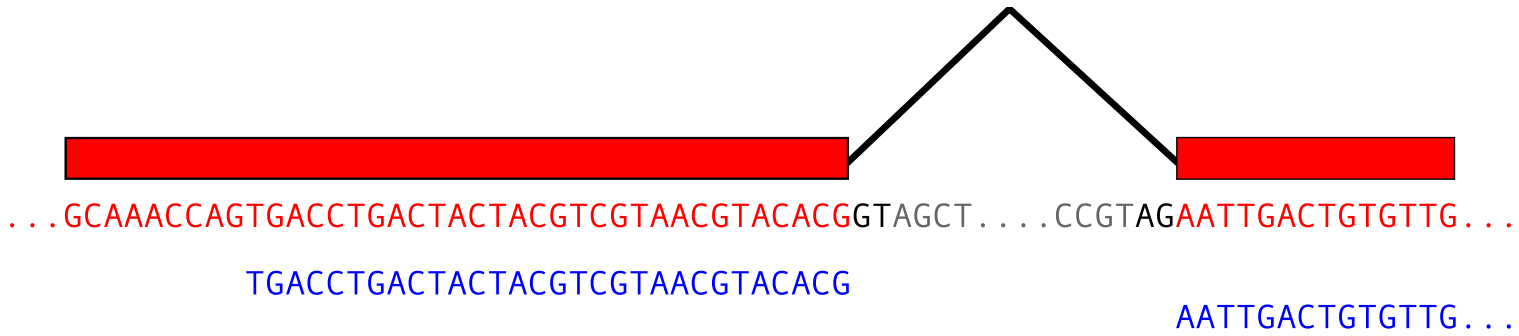
$\Rightarrow$ Identify reads overlapping into introns

Generation of artificially spliced reads



$\Rightarrow$ Remove overlapping parts

Generation of artificially spliced reads



$\Rightarrow$ Remove overlapping parts

Generation of artificially spliced reads



...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT....CCGTAGAATTGACTGTGTTG...

TGACCTGACTACTACGTCGTAACGTACACGAATTGACTGTGTTG...

$\Rightarrow$ Combine both exonic parts

Generation of artificially spliced reads



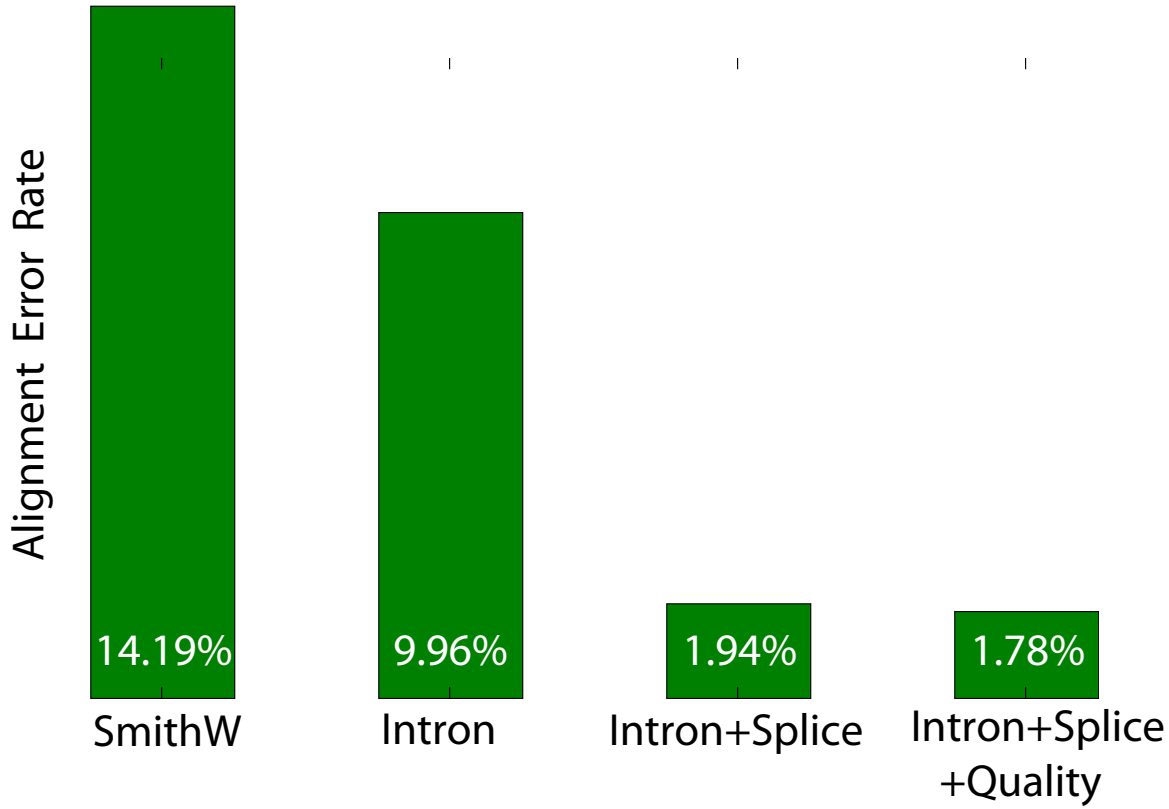
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT....CCGTAGAATTGACTGTGTTG...

TGACCTGACTACTACGTCGTAACGTACACGAATTGA

$\Rightarrow$ Truncate to desired length

First Experiment - Does Quality Info Help?

- Trained on $10,000$ artificially spliced genome reads (Arabidopsis thaliana)
- Tested on $30,000$



Results on Whole Genome Alignments

We evaluated on 2.98×10^6 reads where $\sim 10\%$ were spliced. Results:

- Error rate on spliced reads 5.2%
- Combined spliced/unspliced error rate is 1.7%

De Bona et al. ECCB 2008

Summary / Conclusion

- Challenging data: short error prone reads
- Special learning algorithm
- Successfully combined many information sources
- Fast enough for whole genome/transcriptome sequencing
- Can be applied to data from other sequencing platforms (454)

Summary / Conclusion

- Challenging data: short error prone reads
- Special learning algorithm
- Successfully combined many information sources
- Fast enough for whole genome/transcriptome sequencing
- Can be applied to data from other sequencing platforms (454)

Improvements / future directions

- Speed-up dynamic program
- Tune parameters for seed region finding
- Incorporate information on artifacts

Acknowledgements

Coauthors

- Stephan Ossowski
- Korbinian Schneeberger
- Gunnar Rätsch

Weigel Lab – Data Generation

- Detlef Weigel
- Richard Clark
- Christa Lanz

Stipend

- Friedrich Miescher Lab of the Max Planck Society

Thank you!

fabio@tuebingen.mpg.de

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

