QPalma - Optimal Spliced Alignments of Short Sequence Reads

Fabio De Bona

Friedrich Miescher Laboratory
Max Planck Society, Tübingen, Germany

July 18, 2008

Friedrich Miescher Laboratory
of the Max Planck Society

Introduction

- Next Generation sequencing technologies produce huge amounts of data
- Difference to "classical" sequencing methods:
  - Much faster
  - Much more and shorter fragments
  - Much more errors
- Genome / transcriptome sequencing
- Data can be used via local alignments:
  - Discovery of new genes,
  - Identification of alternative splice forms, ...

Friedrich Miescher Laboratory
of the Max Planck Society

Introduction

- Next Generation sequencing technologies produce huge amounts of data
- Difference to "classical" sequencing methods:
  - Much faster
  - Much more and shorter fragments
  - Much more errors
- Genome / transcriptome sequencing
- Data can be used via local alignments:
  - Discovery of new genes,
  - Identification of alternative splice forms, ...

Challenge

Complete and *accurate* genome/transcriptome sequencing and alignment in a few days rather than weeks at a fraction of the costs

Friedrich Miescher Laboratory
of the Max Planck Society

Methods

Why not use an existing method?

- Next Generation sequencing data has certain properties
- We want to incorporate different information sources
  - Accurate splice site models
  - Intron length model
  - Quality scores model

Friedrich Miescher Laboratory
of the Max Planck Society

Methods

Why not use an existing method?

- Next Generation sequencing data has certain properties
- We want to incorporate different information sources
  - Accurate splice site models
  - Intron length model
  - Quality scores model

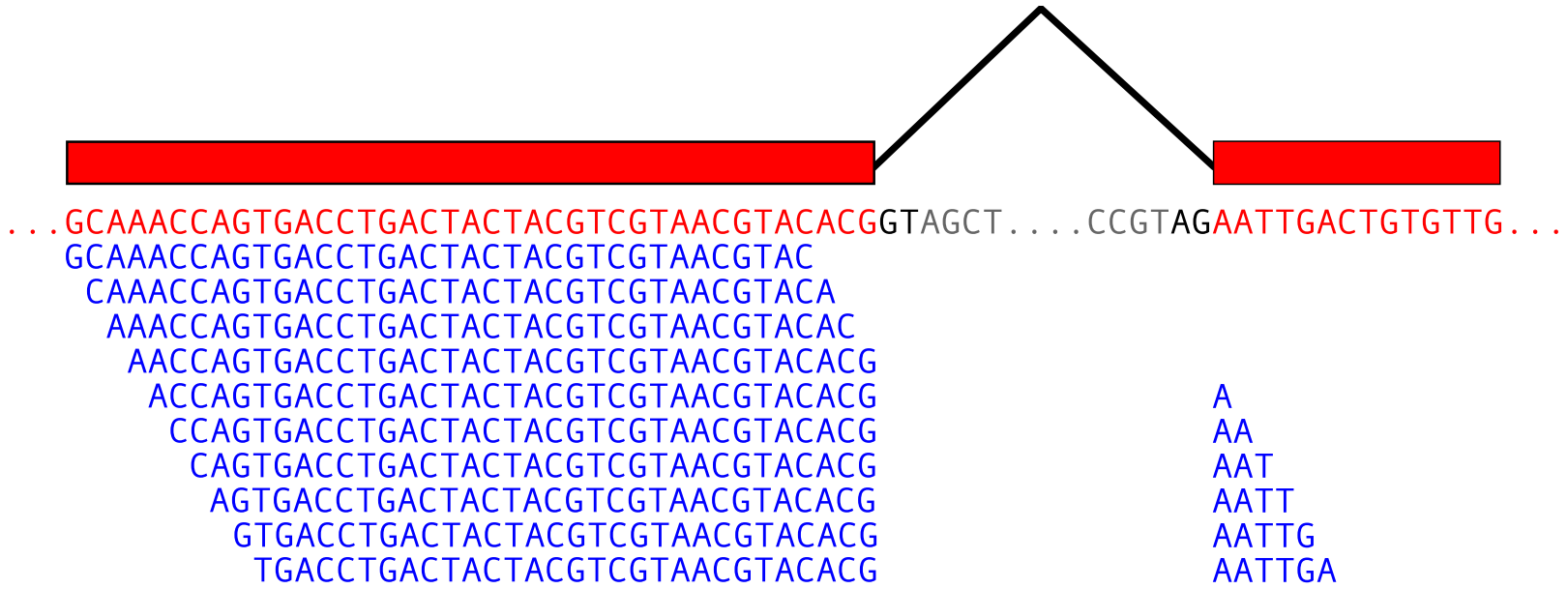
We developed a method which combines

Use a **machine learning** technique to infer optimal parameters

Friedrich Miescher Laboratory
of the Max Planck Society

Two Types of Reads

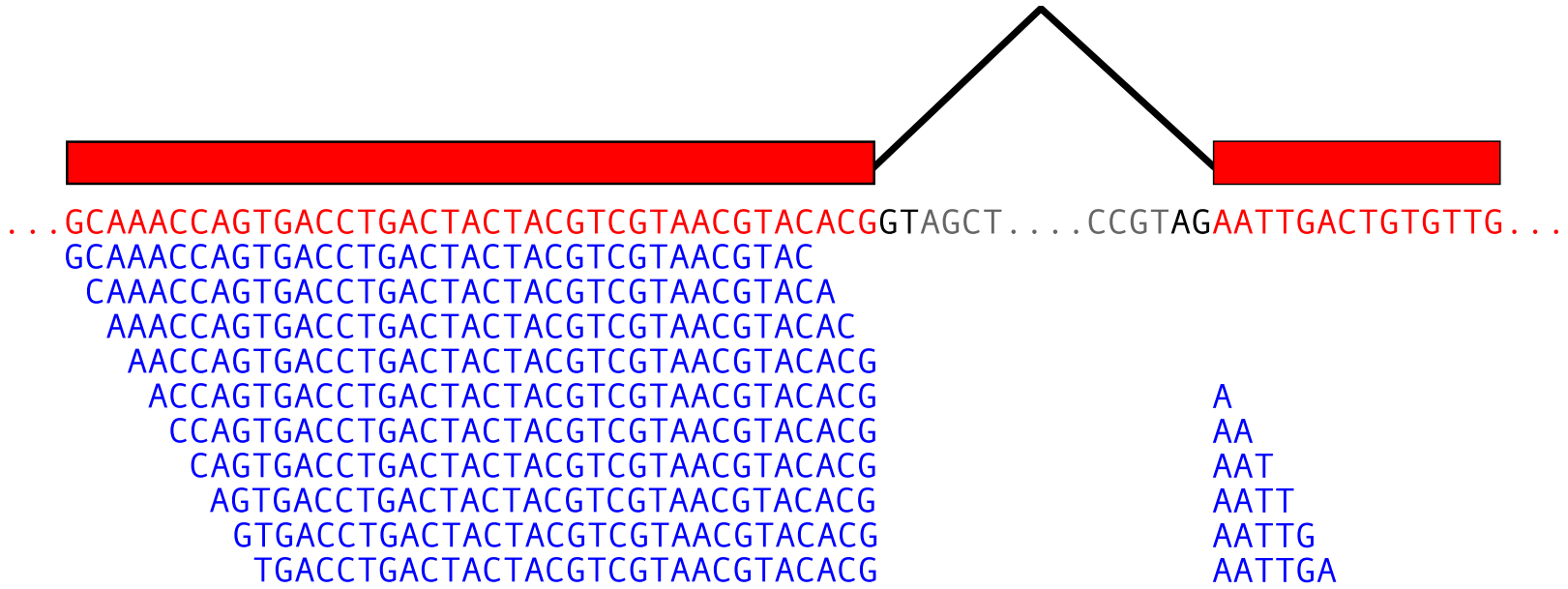
We align transcriptome reads to a reference genome:



Friedrich Miescher Laboratory
of the Max Planck Society

Two Types of Reads

We align transcriptome reads to a reference genome:



- Read lies within exactly one exon: “unspliced read”
- Read overlaps between two or more exons: “spliced read”

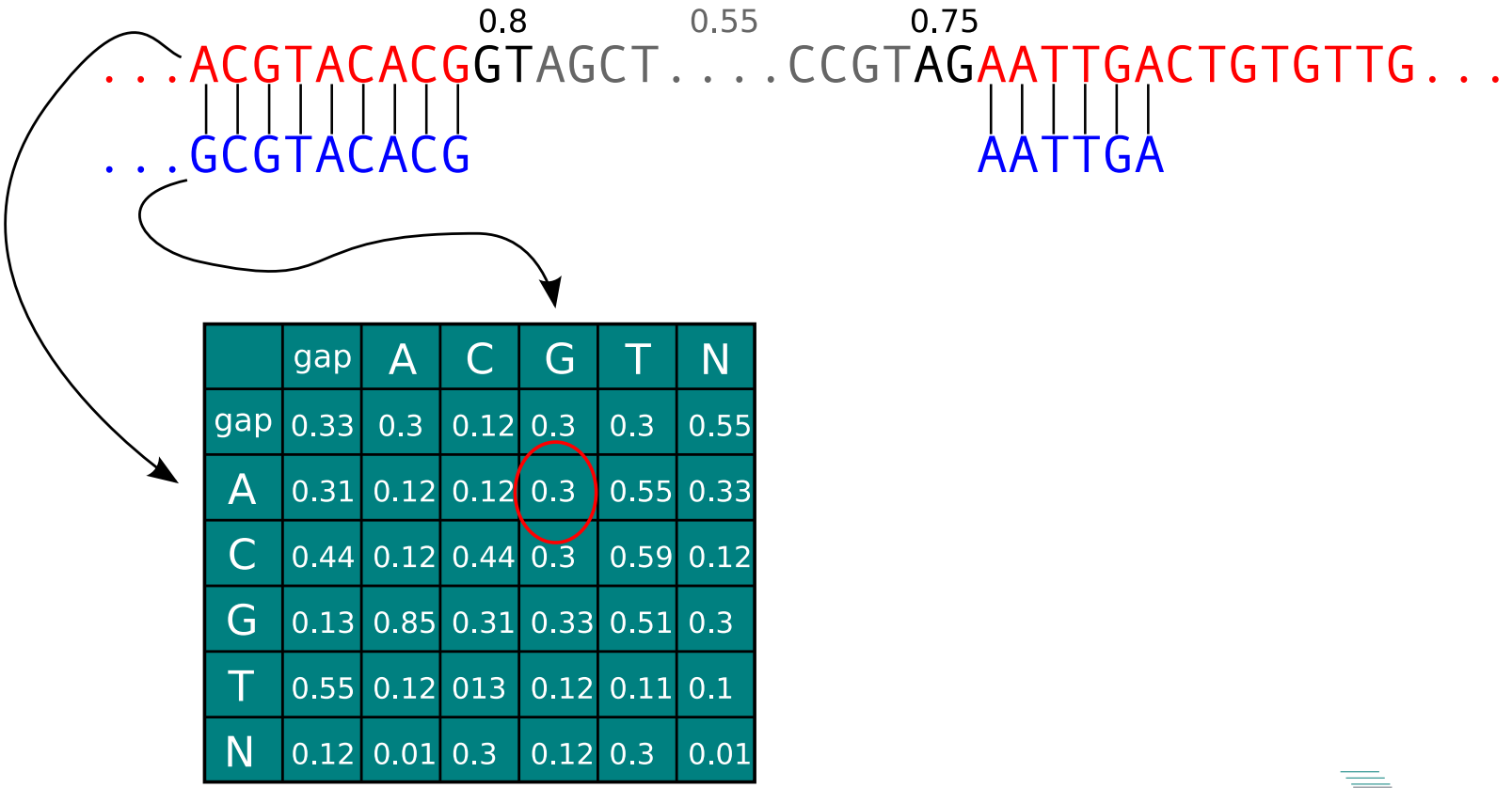
Friedrich Miescher Laboratory
of the Max Planck Society

Alignment Scoring Function

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

Alignment Scoring Function

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



	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

Friedrich Miescher Laboratory
of the Max Planck Society

Reads Have Quality Information

Raw *Illumina Genome Analyzer* data comes with per-base intensities for each position:

Reads Have Quality Information

Raw *Illumina Genome Analyzer* data comes with per-base intensities for each position:

An example read:

ACGTACGTACGTACACAAGTAGACGAACGTAGCTAA

40 40 40 40 38 38 36 36 36 32 32 32 30 25 18 18 40 40 40 40 38 38 36 36 36 32 32 32 30 25 18 18 25 18 18

Friedrich Miescher Laboratory
of the Max Planck Society

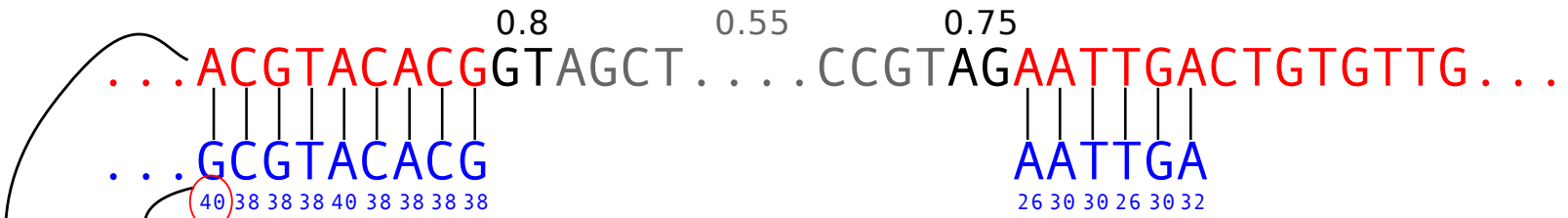
QPalma Scoring Function

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

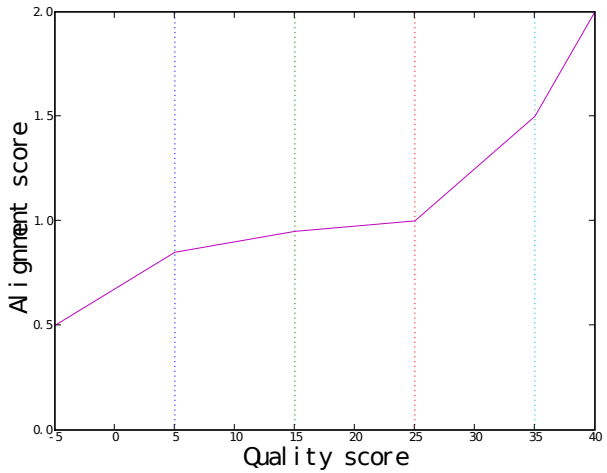
Friedrich Miescher Laboratory
of the Max Planck Society

QPalma Scoring Function

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



	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	$f_{NA}(\cdot)$	$f_{NC}(\cdot)$	$f_{NG}(\cdot)$	$f_{NT}(\cdot)$	$f_{NN}(\cdot)$



Friedrich Miescher Laboratory
of the Max Planck Society

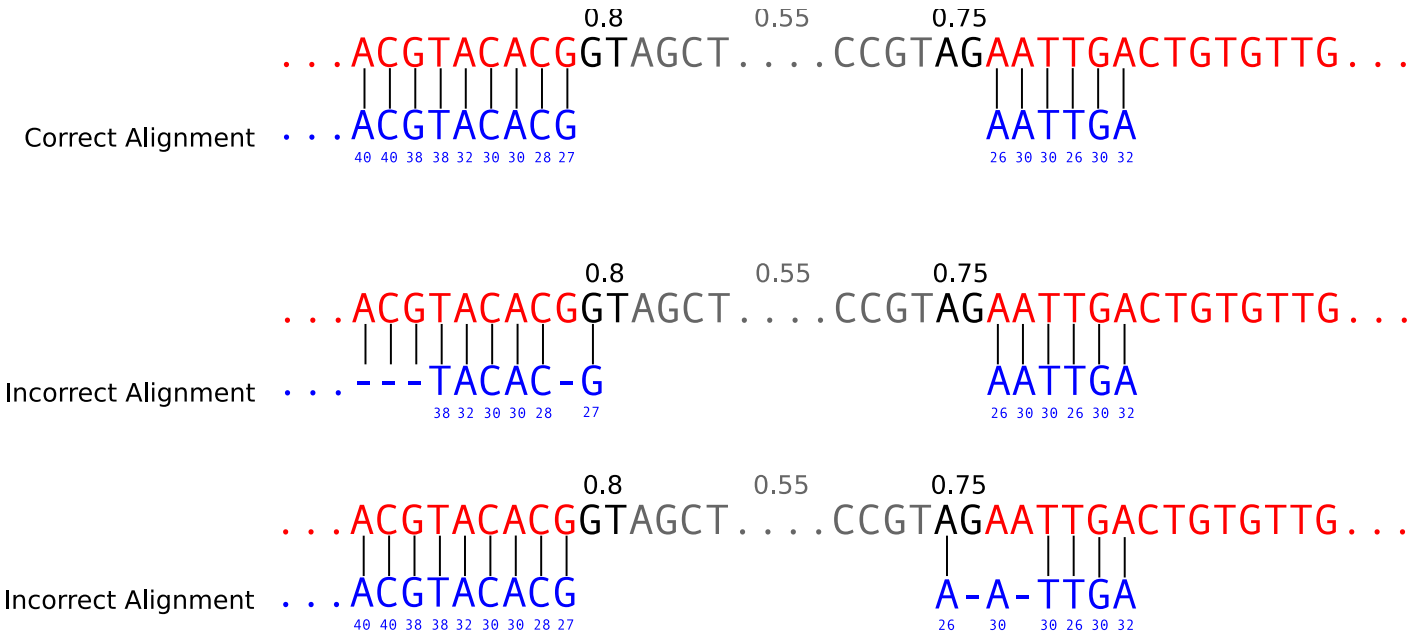
Scoring Parameter Inference

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

Friedrich Miescher Laboratory
of the Max Planck Society

Scoring Parameter Inference

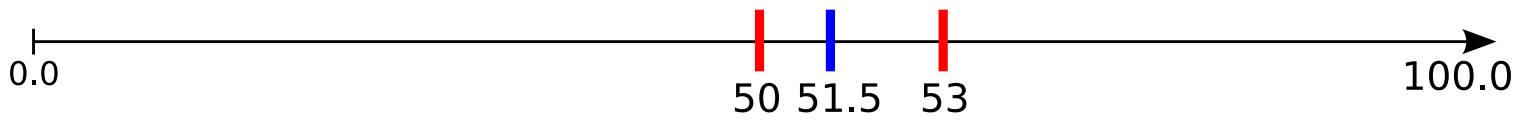
- What are optimal parameters?
- How do we jointly optimize 336 parameters?



Friedrich Miescher Laboratory
of the Max Planck Society

Cartoon

Scoring of the three alignments:

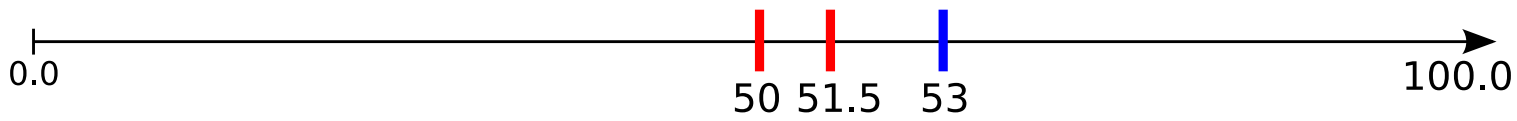


- Correct alignment is **not** highest scoring one

Friedrich Miescher Laboratory
of the Max Planck Society

Cartoon

Scoring of the three alignments:

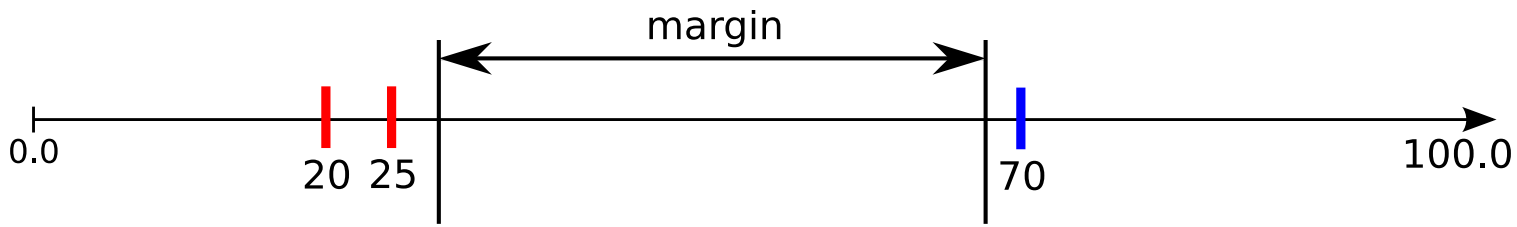


- Correct alignment is **not** highest scoring one
- Can we do better?

Friedrich Miescher Laboratory
of the Max Planck Society

Cartoon

Scoring of the three alignments:

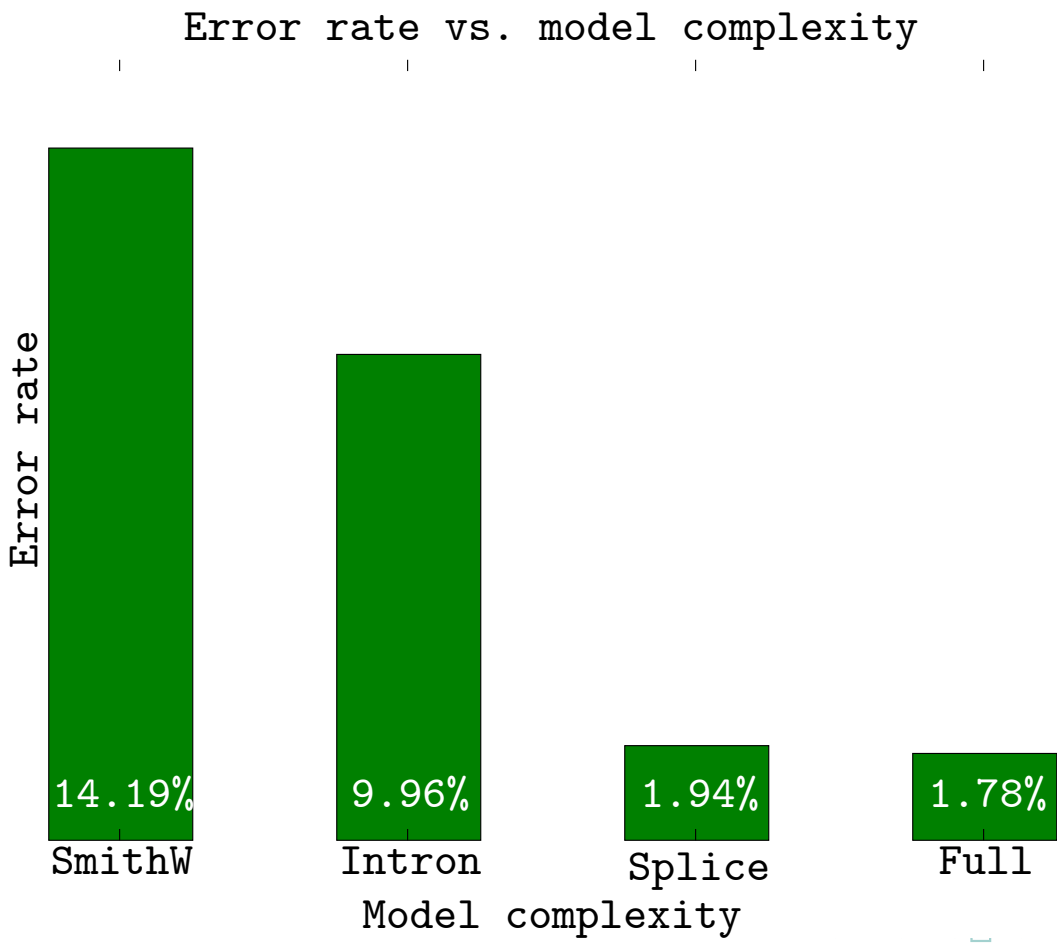


- Use a technique motivated by "large-margin" methods
- Idea: Enforce a margin between correct and incorrect examples
- One has to solve a huge quadratic problem

Friedrich Miescher Laboratory
of the Max Planck Society

First Experiment - Does Quality Info Help?

- Trained on $10,000$ artificially spliced genome reads
- Tested on $30,000$



Friedrich Miescher Laboratory
of the Max Planck Society

A Pipeline for Efficient Large Scale Alignment

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

A Pipeline for Efficient Large Scale Alignment

1. Runtime 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?

A Pipeline for Efficient Large Scale Alignment

1. Runtime 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

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

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

Friedrich Miescher Laboratory
of the Max Planck Society

Summary / Conclusion

- Next Generation sequencing is the future
- Data challenges: short error prone reads
- Special learning algorithm
- Successfully combined many information sources
- Fast enough for whole genome sequencing
- Can be easily extend to other sequencing models (454)

Friedrich Miescher Laboratory
of the Max Planck Society

Summary / Conclusion

- Next Generation sequencing is the future
- Data challenges: short error prone reads
- Special learning algorithm
- Successfully combined many information sources
- Fast enough for whole genome sequencing
- Can be easily extend to other sequencing models (454)

Improvements / future directions

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

Friedrich Miescher Laboratory
of the Max Planck Society

Acknowledgements

- Stephan Ossowski
- Korbinian Schneeberger
- Dr. Gunnar Rätsch
- *Weigel Lab* – Detlef Weigel

- Friedrich Miescher Lab of the Max Planck Society
- Max Planck Institute for Developmental Biology, Tübingen, Germany

Friedrich Miescher Laboratory
of the Max Planck Society

Thank you!



fabio@tuebingen.mpg.de

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