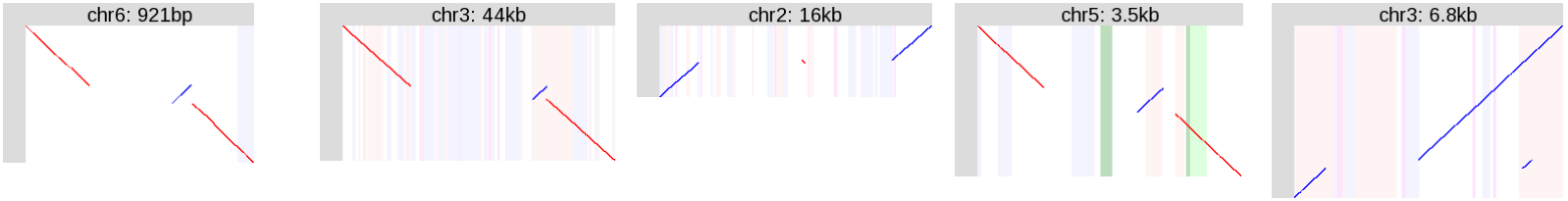
Finding and aligning related regions of sequences

Martin C. Frith
Artificial Intelligence Research Center, AIST
Graduate School of Frontier Sciences, University of Tokyo
AIST-Waseda University CBBD-OIL
2017-09-08

https://sites.google.com/site/frithbioinfo/



Contents

- Background
- What are we really trying to do?
- Probability-based alignment
- Moar alignment!
- Determining rates of substitution, insertion & deletion
- Alignment ambiguity
- Alignment with duplications & rearrangements
- Aligning spliced RNA or cDNA to a genome

Contents

- Background
- What are we really trying to do?
- Probability-based alignment
- Moar alignment!
- Determining rates of substitution, insertion & deletion
- Alignment ambiguity
- Alignment with duplications & rearrangements
- Aligning spliced RNA or cDNA to a genome

Diverse genetic sequence data

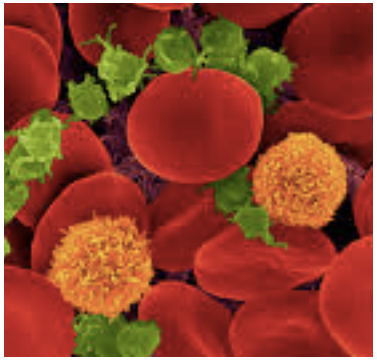
What kinds of microbes are in here?

Or in here?

What's wrong with his DNA?

Why so long-lived?

ctagcggtattattgcct
aaattgctattatggttctg
gctattatgatgtagtaa
tctctgattatatgata
ctcgttatatattttaaaa
cccgggggggtatatattaaa
aaaatattattatattaaaaaa



What genes are active in each cell?

Infectious diseases

Ancient DNA

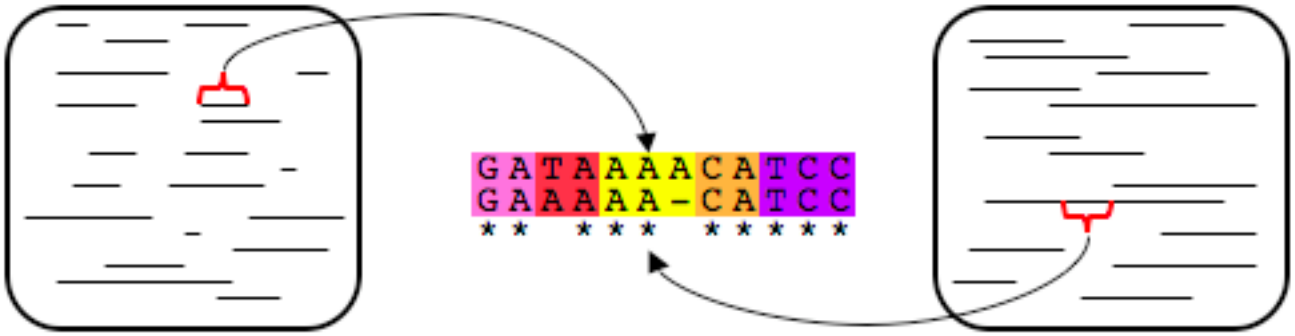
Religious relics?

Diverse genetic sequence data

- DNA sequencing machines are not (yet) perfect!
- They can only produce:
  - Short DNA “reads”, with errors
- Or:
  - Longer DNA reads, with more errors

How do we analyze these sequences?

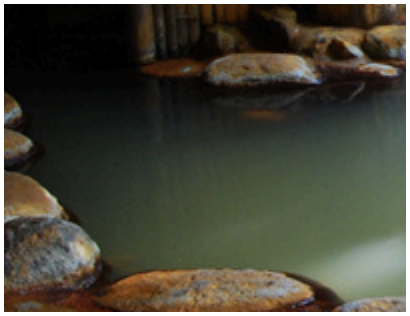
The main way is by comparing and aligning them



Diverse types of sequence data

Diverse sequence comparison tasks

What kinds of microbes are there?



```
ctagcggtattattgcct
   aaattgctattatggttctg
gctattatgatgtagtaa
      tctctgattatatgata
          ctcgttatatattttaaaa
            cccgggggggtatatattaaa
               aaaatattattatattaaaaaa
```

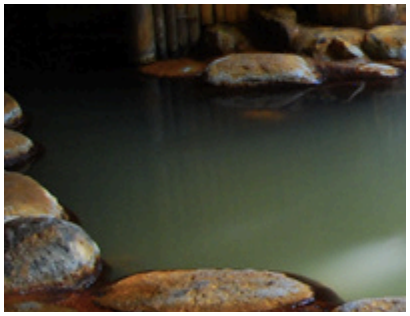
Compare to a database of known genomes

```
   atatatatatat--ccgt
   ||||.||..||...||||
…cgatat-tacttactgccgttgc…
   …gctataaaaggtctctggagaaa…
```

Compare to a database of known proteins

```
   atatatatatattagccgt
   |||...||| |||...|||
LysPheAlaPro-ProGlyGlyAla…
CysTrpTrpAlaGlyAlaPro…
```

What kinds of microbes are there?



```
ctagcggtattattgcct
   aaattgctattatggttctg
gctattatgatgtagtaa
      tctctgattatatgata
          ctcgttatatattttaaaa
            cccgggggggtatatattaaa
               aaaatattattatattaaaaaa
```

Compare to a database of known genomes

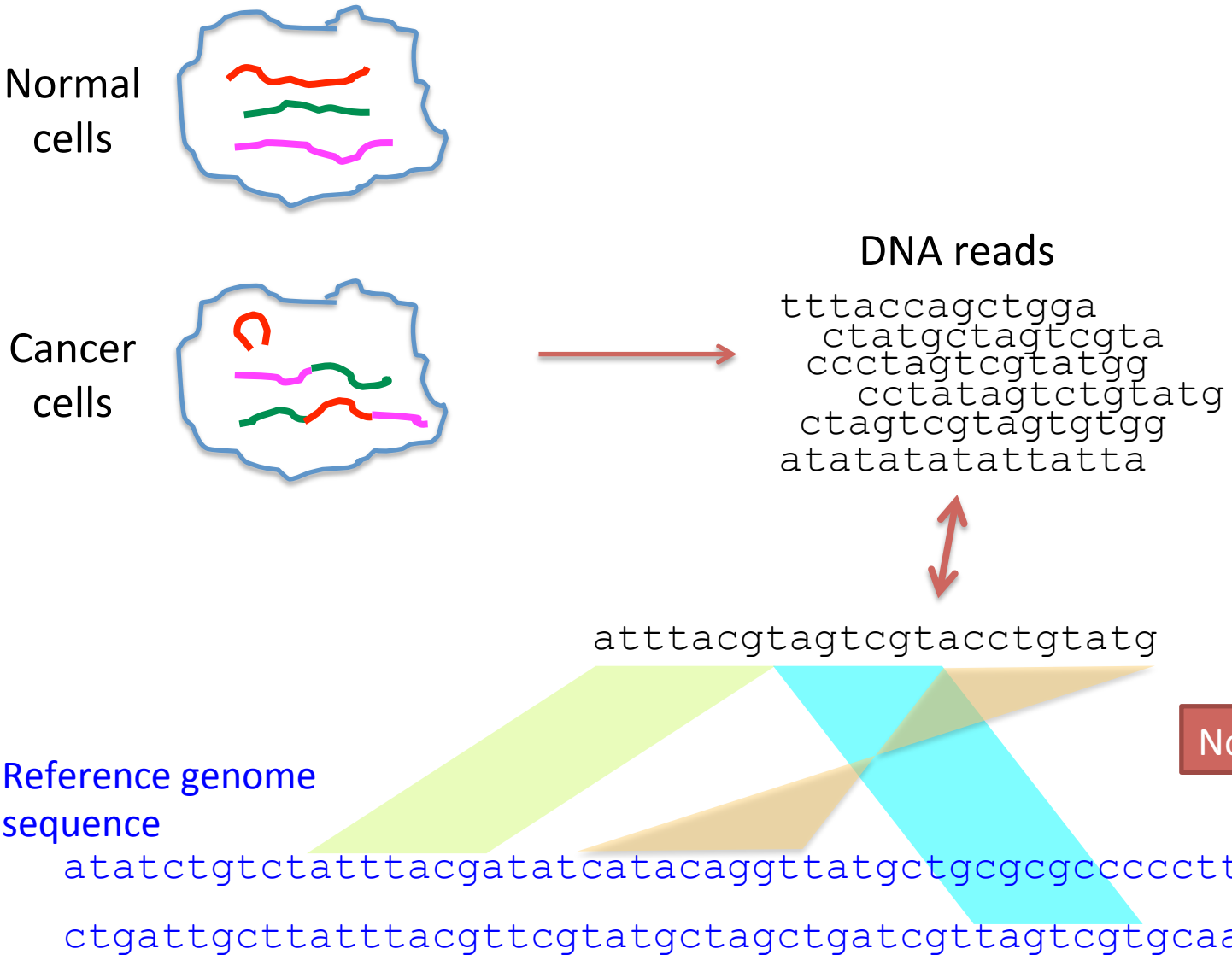
```
   atatatatatat--ccgt
   ||||.||..||...||||
…cgatat-tacttactgccgttgc…
   …gctataaaaggtctctggagaaa…
```

Compare to a database of known proteins

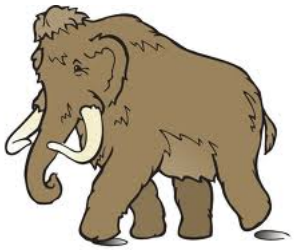
frame-shift

```
   atatatatatattagccgt
   |||...|||  |||...|||
LysPheAlaPro-ProGlyGlyAla…
CysTrpTrpAlaGlyAlaPro…
```

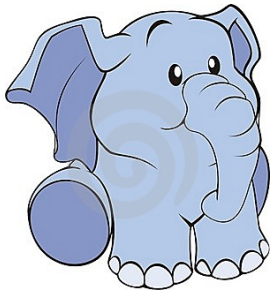
Finding DNA rearrangements in cancer



Ancient DNA



Mammoth DNA reads



Elephant genome

aactgatgctgta

ctatgctagtgtgc

cgcaaaagtgta

ccggctatattcg

gcggcgtagctg

cccgatgctagta

gtcgtagctgat

atgatcgtagct

gcggcgtagctg

gcggcgtagctg

aaaaaaaaaaaa

```
cgca--tgtgta
||||    |||||
cgcaaaagtgta
```

```
atgctatttctat
||||  |||  ||||
atgcaatt-ctat
```

tgtagtcgtagctatgctagctgtcgtaatatattatatgctagttactcgaac

ctgctagctagctgcgtcgtctcttattatatatatatattgcgcgcgtgtc

ctgatcgtacgtagctagtcgtcggcgcgcgatgctgtgta

cgctgatgctgtgtagtgatcgtgatgctgcagtgc

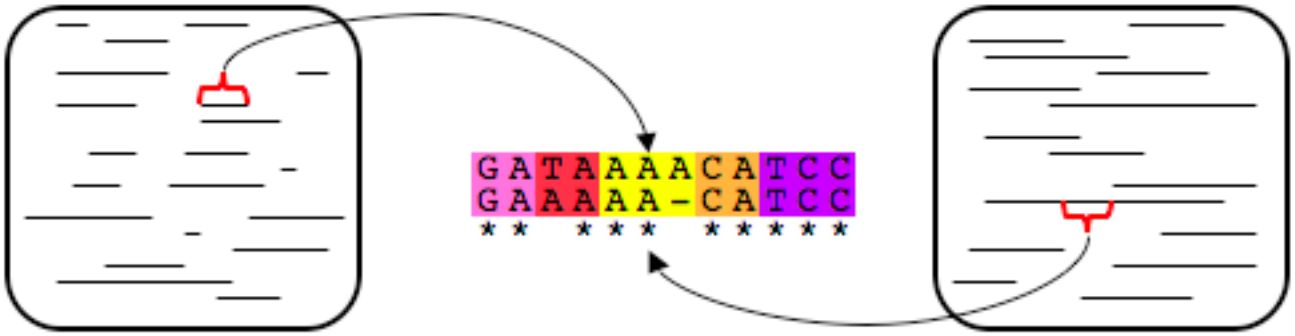
ctgctcgtcgtatattatatctgtcgtagctgt

tgtgatcgtatattatatctgctgcgtatg

ctgatcgtagctagctttttattgccc

How do we analyze these sequences?

The main way is by comparing and aligning them



Diverse types of sequence data

Diverse sequence comparison tasks

Contents

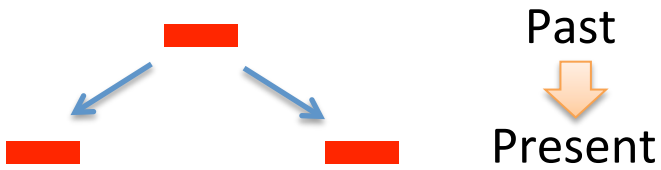
- Background
- What are we really trying to do?
- Probability-based alignment
- Moar alignment!
- Determining rates of substitution, insertion & deletion
- Alignment ambiguity
- Alignment with duplications & rearrangements
- Aligning spliced RNA or cDNA to a genome

What are we really trying to do?

- Find similar sequences?
- Find homologous sequences?
  - What about paralogous sequences?

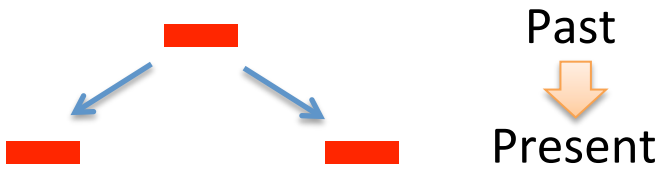
Homology

Homology: descent from a common ancestor

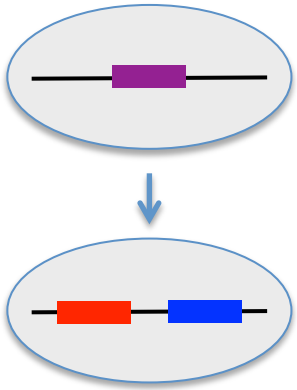


Homology

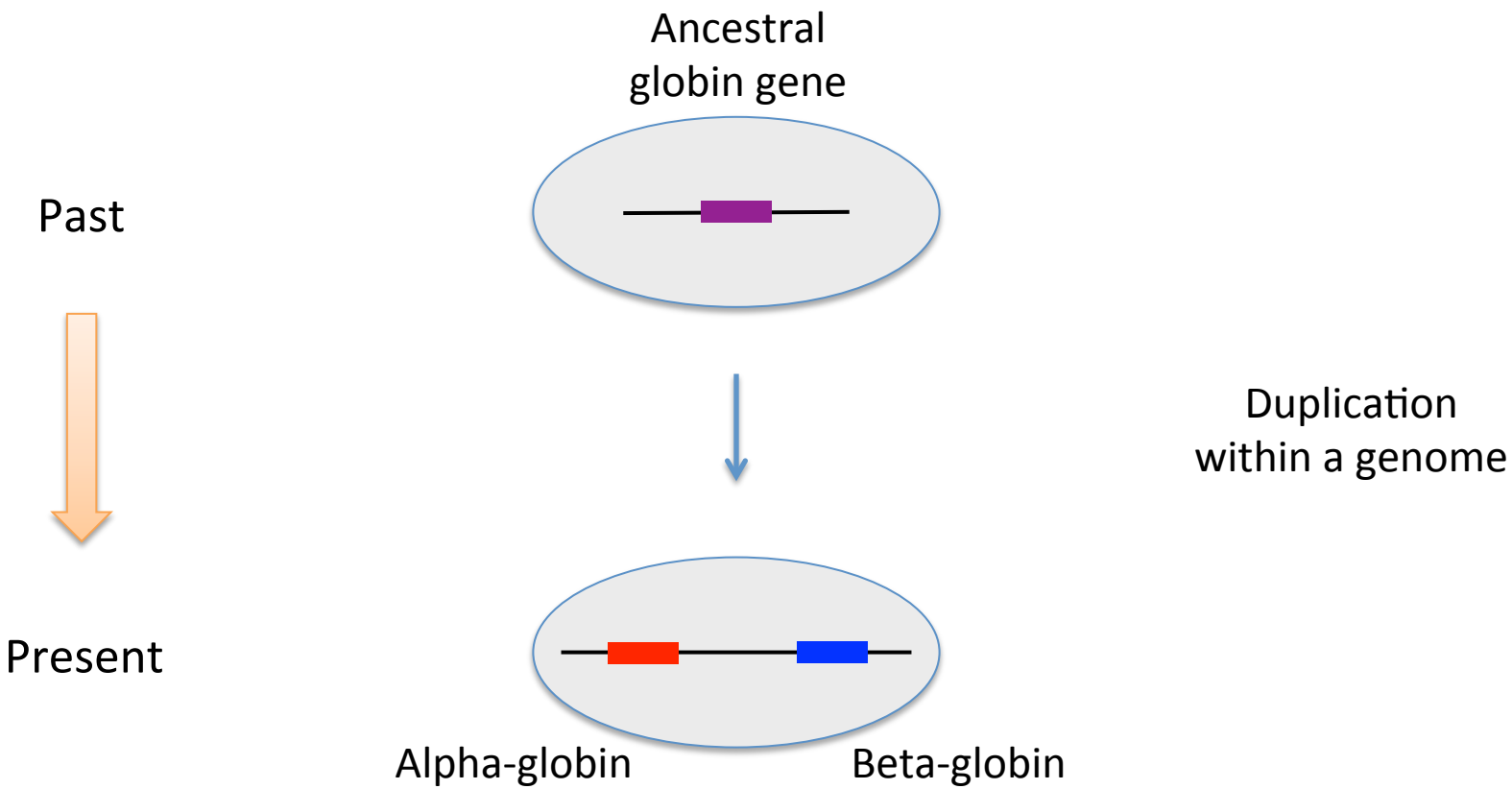
Homology: descent from a common ancestor



Paralogy: descent from most recent common ancestor by duplication within a genome



Example



Paralogs

Compare DNA from a patient to a reference genome

gctattatgatgtagtaa
tctctgattatatgata
ctcgttatatattttaaaa
cccgggggggtatatattaaa
aaaatattattatattaaaaaa

DNA reads

A DNA read from alpha-globin: cgcggtataagctataaggtta

Referenc e genome sequence: ctgattgcttatttacgttcgtatgctagctgatcgtagtcgtcgagcttatcgtgggc

Alpha-globin

Beta-globin

Here, we want to avoid paralogs.
Homology is necessary but not sufficient.

Unknown bacteria

```
ctagcggtattattgcct
   aaattgctattatggttctg
gctattatgatgtagtaa
      tctctgattatatgata
          ctcgttatatattttaaaa
            cccgggggggtatatattaaa
               aaaatattattatattaaaaaa
                  ..........
```

Here, we don't try to avoid paralogs.
(Maybe because it seems too hard?)

Homology is necessary and sufficient.

Compare to a database of known proteins

```
   atatatatatattagccgt
   |||...||| |||...|||
LysPheAlaPro-ProGlyGlyAla…
CysTrpTrpAlaGlyAlaPro…
```

What are we really trying to do?

- Find similar sequences?
- Find homologous sequences?
  - What about paralogous sequences?
- *It depends on our specific task*

Contents

- Background
- What are we really trying to do?
- Probability-based alignment
- Moar alignment!
- Determining rates of substitution, insertion & deletion
- Alignment ambiguity
- Alignment with duplications & rearrangements
- Aligning spliced RNA or cDNA to a genome

How do we find and align related sequences?

- The most accurate way is by using probabilities
- This is not new
  - In 20-year-old textbooks
- Unfortunately, many people don't do this!

Probability-based alignment

Step 1: Determine probabilities (rates) of substitution, deletion & insertion

	a	c	g	t
a	.29	.0019	.019	.0017
c	.0019	.18	.00064	.0040
g	.019	.00064	.18	.0013
t	.0017	.0040	.0013	.29

For example, between one set of human DNA reads and a reference human genome

Deletion: open = 0.067 extend = 0.44
Insertion: open = 0.017 extend = 0.48

Probability-based alignment

Step 1: Determine probabilities (rates) of substitution, deletion & insertion

	a	c	g	t
a	.29	.0019	.019	.0017
c	.0019	.18	.00064	.0040
g	.019	.00064	.18	.0013
t	.0017	.0040	.0013	.29

For example, between one set of human DNA reads and a reference human genome

These probabilities are for human nanopore (R9.4) reads. The rate of A↔G errors is quite high.

Deletion: open = 0.067 extend = 0.44
Insertion: open = 0.017 extend = 0.48

Probability-based alignment

Step 1: Determine probabilities (rates) of substitution, deletion & insertion

	a	c	g	t
a	.29	.0019	.019	.0017
c	.0019	.18	.00064	.0040
g	.019	.00064	.18	.0013
t	.0017	.0040	.0013	.29

For example, between one set of human DNA reads and a reference human genome

These probabilities are for human nanopore (R9.4) reads. The rate of A↔G errors is quite high.

Deletion: open = 0.067 extend = 0.44
Insertion: open = 0.017 extend = 0.48

Align sequences *while considering these probabilities*.
Prefer alignments with higher probability.

```
ctatgccacgtgaggtgtggc          a c g t g - - a g g
                               | |   | |     | | |
attacatgctaggccac              a c a t g c t a g g
```

Score-based alignment

Step 1: Choose scores for substitution, deletion & insertion

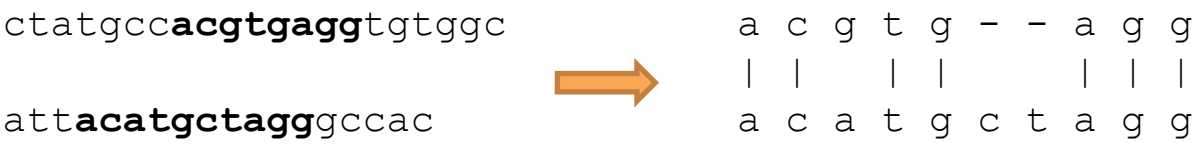
	a	c	g	t
a	5	-16	-8	-12
c	-16	7	-20	-10
g	-8	-20	7	-15
t	-12	-10	-15	7

Deletion: open = -7 extend = -3
Insertion: open = -14 extend = -2

You will often see this.

It is mathematically equivalent to probability-based alignment.

Prefer alignments with higher score



Alignment score = **sum** of match, mismatch, gap scores

Probability-based alignment

Step 1: Determine probabilities (rates) of substitution, deletion & insertion

	a	c	g	t
a	.29	.0019	.019	.0017
c	.0019	.18	.00064	.0040
g	.019	.00064	.18	.0013
t	.0017	.0040	.0013	.29

For example, between one set of human DNA reads and a reference human genome

These probabilities are for human nanopore (R9.4) reads. The rate of A↔G errors is quite high.

Deletion: open = 0.067 extend = 0.44
Insertion: open = 0.017 extend = 0.48

Align sequences *while considering these probabilities*.
Prefer alignments with higher probability.

```
ctatgccacgtgaggtgtggc          a c g t g - - a g g
                               | |   | |     | | |
attacatgctaggccac              a c a t g c t a g g
```

Probability-based alignment

Step 1: Determine probabilities (rates) of substitution, deletion & insertion

	a	c	g	t
a	.29	.0019	.019	.0017
c	.0019	.18	.00064	.0040
g	.019	.00064	.18	.0013
t	.0017	.0040	.0013	.29

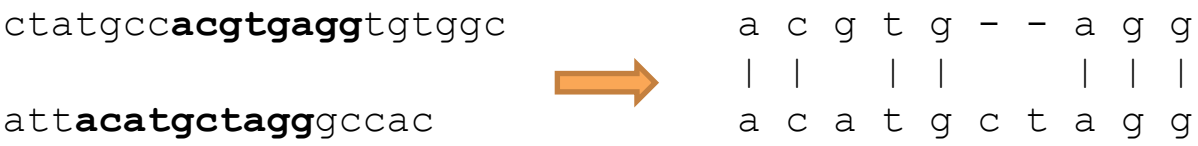
For example, between one set of human DNA reads and a reference human genome

These probabilities are for human nanopore (R9.4) reads. The rate of A↔G errors is quite high.

Deletion: open = 0.067 extend = 0.44
Insertion: open = 0.017 extend = 0.48

This is an accurate method, **if** we use probabilities that fit our data

Align sequences *while considering these probabilities*.
Prefer alignments with higher probability.



Examples of special probabilities

- Plasmodium falciparum (malaria)
  - DNA is very AT-rich: 80% A+T, 20% G+C
- Bisulfite-converted DNA
  - A method for detecting DNA methylation
  - Produces DNA reads with biased C/T composition
- PAR-CLIP
  - A method for finding RNA-protein interactions
  - Produces DNA reads with altered probabilities

Probability-based alignment

Unknown bacteria

ctagcggtattattgcct

aaattgctattatggttctg

gctattatgatgtagtaa

tctctgattatatgata

ctcgttatatattttaaaa

cccgggggtatatattaaa

aaaatattattatattaaaaaa

Query sequences

Compare to a database of known genomes

gctgtatatgctgctattgctgta…

cgattatatatattagtatt…

Probability-based alignment

Unknown bacteria

ctagcggtattattgcct
aaattgctattatggttctg
gctattatgatgtagtaa
tctctgattatatgata
ctcgttatatattttaaaa
cccgggggtatatattaaa
aaaatattattatattaaaaaa

Query sequences

Compare to a database of known genomes

gctgtatatgctgctattgctgta…

cgattatatatattagtatt…

Some query sequences have close relatives in the database

Other query sequences only have distant relatives in the database

The probabilities (of substitution, deletion & insertion) vary!

Probability-based alignment

Unknown bacteria

ctagcggtattattgcct
aaattgctattatggttctg
gctattatgatgtagtaa
tctctgattatatgata
ctcgttatatattttaaaa
cccgggggtatatattaaa
aaaatattattatattaaaaaa

Query sequences

Compare to a database of known genomes

gctgtatatgctgctattgctgta…
cgattatatatattagtatt…

Some query sequences have close relatives in the database

Other query sequences only have distant relatives in the database

The probabilities (of substitution, deletion & insertion) vary!

The same problem occurs in protein sequence search (e.g. BLAST)

Usual “solution”: use a compromise set of probabilities

Contents

- Background
- What are we really trying to do?
- Probability-based alignment
- Moar alignment!
- Determining rates of substitution, insertion & deletion
- Alignment ambiguity
- Alignment with duplications & rearrangements
- Aligning spliced RNA or cDNA to a genome

A strange scoring scheme

	a	c	g	t
a	9	-1	-1	-1
c	-1	9	-1	-1
g	-1	-1	9	-1
t	-1	-1	-1	9

- Alignment score = **sum** of match, mismatch, gap scores

A strange scoring scheme

	a	c	g	t
a	9	-1	-1	-1
c	-1	9	-1	-1
g	-1	-1	9	-1
t	-1	-1	-1	9

- Alignment score = **sum** of match, mismatch, gap scores
- *Completely random* DNA has 1 match per 4 bases
- So this scoring scheme will align random DNA!

Moar alignment!

	a	c	g	t
a	9	-1	-1	-1
c	-1	9	-1	-1
g	-1	-1	9	-1
t	-1	-1	-1	9


```
ctgatcatttgcgacga
    ||||  |||
cctattatacatt-gcgtgctgat
```

IdiotAligner

```
ctgatcatttgcgacga
 |xxx||||  |||
cctattatacatt-gcgtgctgat
```

My aligner aligns more bases!
It must be better!

Moar alignment!

- This may seem obvious, but…
- In practice, people are often tempted by aligning as much as possible
  - As many reads as possible
  - As many bases as possible
- Be careful of this temptation
  - It may lead you to **IdiotAligner**

Contents

- Background
- What are we really trying to do?
- Probability-based alignment
- Moar alignment!
- Determining rates of substitution, insertion & deletion
- Alignment ambiguity
- Alignment with duplications & rearrangements
- Aligning spliced RNA or cDNA to a genome

Probability-based alignment

Step 1: Determine probabilities (rates) of substitution, deletion & insertion

	a	c	g	t
a	.29	.0019	.019	.0017
c	.0019	.18	.00064	.0040
g	.019	.00064	.18	.0013
t	.0017	.0040	.0013	.29

For example, between one set of human DNA reads and a reference human genome

Deletion: open = 0.067 extend = 0.44
Insertion: open = 0.017 extend = 0.48

Probability-based alignment

Step 1: Determine probabilities (rates) of substitution, deletion & insertion

How?

last-train

Training alignment parameters for arbitrary sequencers with LAST-TRAIN

Michiaki Hamada,[1,2,3,*] Yukiteru Ono,[4] Kiyoshi Asai[3,5] and Martin C. Frith[2,3,5,*]

Bioinformatics Advance Access published December 30, 2016

Probability-based alignment

Step 1: Determine probabilities (rates) of substitution, deletion & insertion

How?

last-train

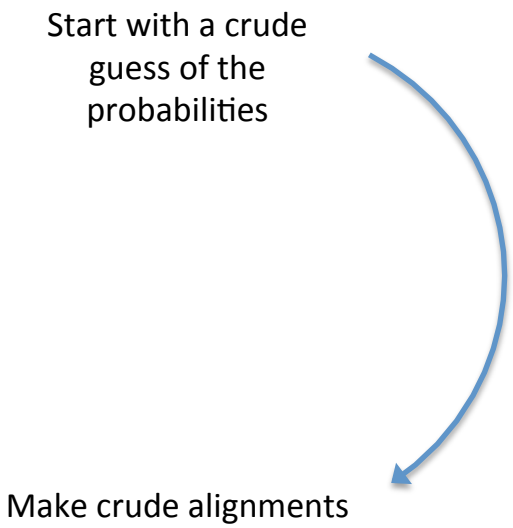
"last-train always works well, by magic": not true.

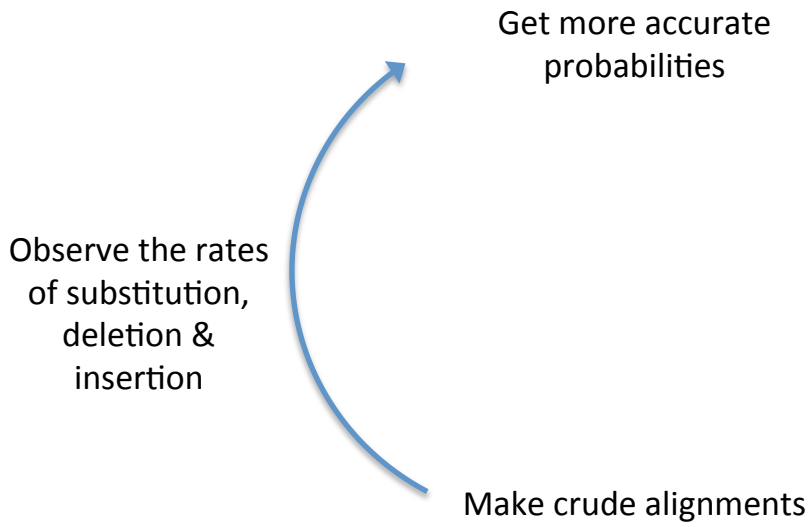
Best to understand roughly how it works.

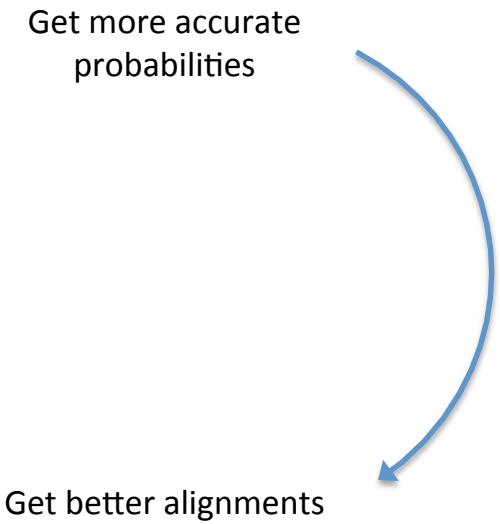
Training alignment parameters for arbitrary sequencers with LAST-TRAIN

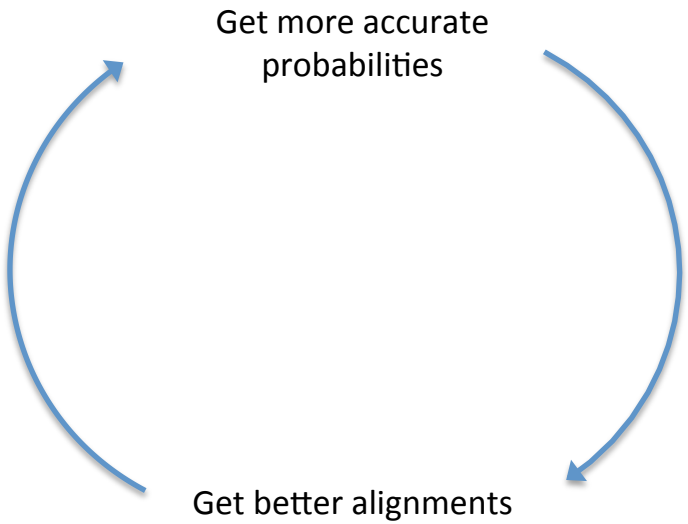
Michiaki Hamada,[1,2,3,*] Yukiteru Ono,[4] Kiyoshi Asai[3,5] and Martin C. Frith[2,3,5,*]

Bioinformatics Advance Access published December 30, 2016


Start with a crude guess of the probabilities
Make crude alignments

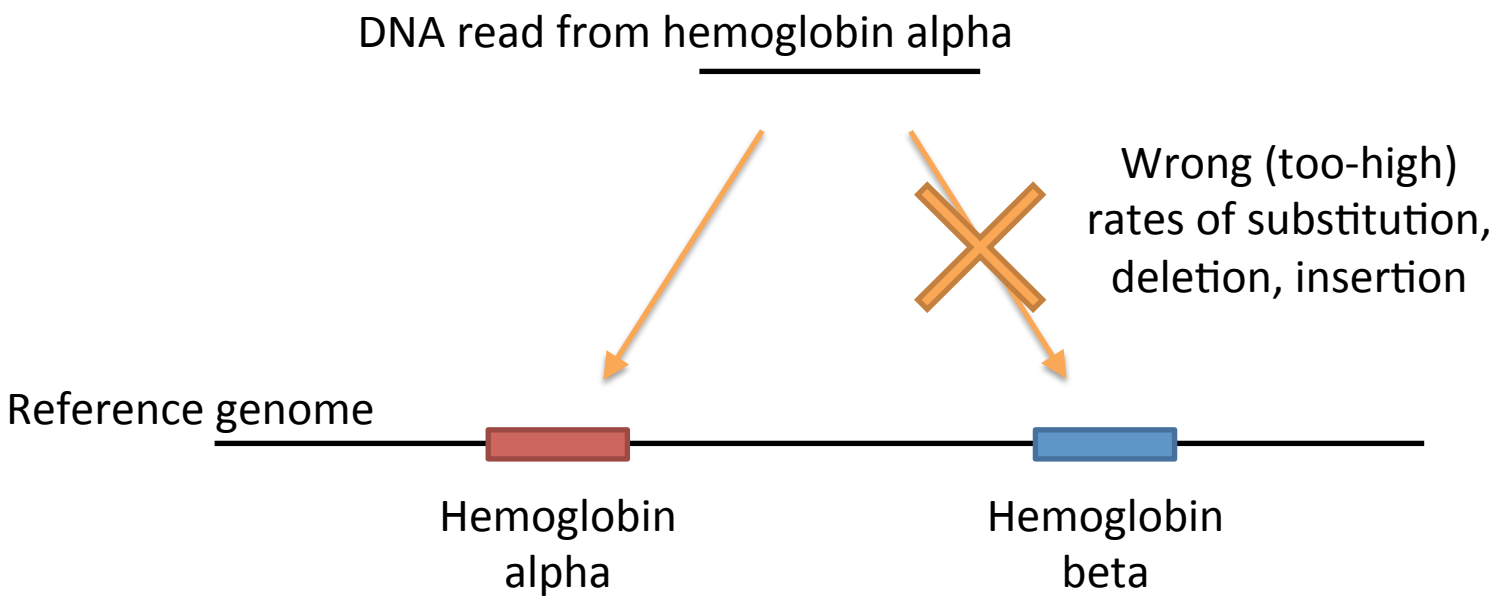

Get more accurate probabilities
Observe the rates of substitution, deletion & insertion
Make crude alignments


Get more accurate probabilities
Get better alignments



Keep iterating until the result stops changing

Paralogs should be avoided



`last-train` only uses the most-similar alignment for (each part of) each DNA read

How well does `last-train` work in practice?

- Human DNA reads versus human genome:
  - Seems to work very well
- DNA reads from a jellyfish versus badly-assembled genome of another jellyfish:
  - My colleague is trying to do this. It might work.
- Metagenomic DNA reads versus microbe DNA database:
  - I have not tested last-train in this situation.

Contents

- Background
- What are we really trying to do?
- Probability-based alignment
- Moar alignment!
- Determining rates of substitution, insertion & deletion
- Alignment ambiguity
- Alignment with duplications & rearrangements
- Aligning spliced RNA or cDNA to a genome

Alignment ambiguity

```
ctagctaaccgtatcgtgggc
|||||  |    |||||   | ||
ctagcca---gtatctagtgc
```

Or

```
ctagctaaccgtatcgtgggc
|||||    | |||||   | ||
ctagc---cagtatctagtgc
```

?

Per-column probabilities

…	g	c	a	t	c	c	t	t	g	g	g	t	c	t	c	g	a	c	a	t	…
…	g	c	c	t	c	g	t	t	a	g	a	-	-	t	a	g	a	t	a	g	…
	.99	.99	.99	.95	.93	.92	.90	.79	.55	.33	.16	.22	.49	.55	.59	.71	.93	.97	.98	.99	

- Can be calculated by considering the probabilities of alternative alignments

Per-column probabilities

…	g	c	a	t	c	c	t	t	g	g	g	t	c	t	c	g	a	c	a	t	…
…	g	c	c	t	c	g	t	t	a	g	a	-	-	t	a	g	a	t	a	g	…
	.99	.99	.99	.95	.93	.92	.90	.79	.55	.33	.16	.22	.49	.55	.59	.71	.93	.97	.98	.99	

- Can be calculated by considering the probabilities of alternative alignments
- Per-column probabilities help in finding genetic variants (e.g. SNPs) accurately
  - e.g. by `last-genotype` (https://github.com/mcfrith/last-genotype)

Contents

- Background
- What are we really trying to do?
- Probability-based alignment
- Moar alignment!
- Determining rates of substitution, insertion & deletion
- Alignment ambiguity
- Alignment with duplications & rearrangements
- Aligning spliced RNA or cDNA to a genome

What kinds of sequence change occur?

Substitution

catgtctccccta

↓

catgtcttcctcta

Deletion

cctgtatatatgctataa

↓

cctgtactataa

Duplication

aaaatctgtattg

↓

aaaatctgaatctgtattg

"Spontaneous generation": rare

catgtctcccta

↓

catgtctcagactagccta

Re-positioning

catgtctggcgattagtcccta

↓

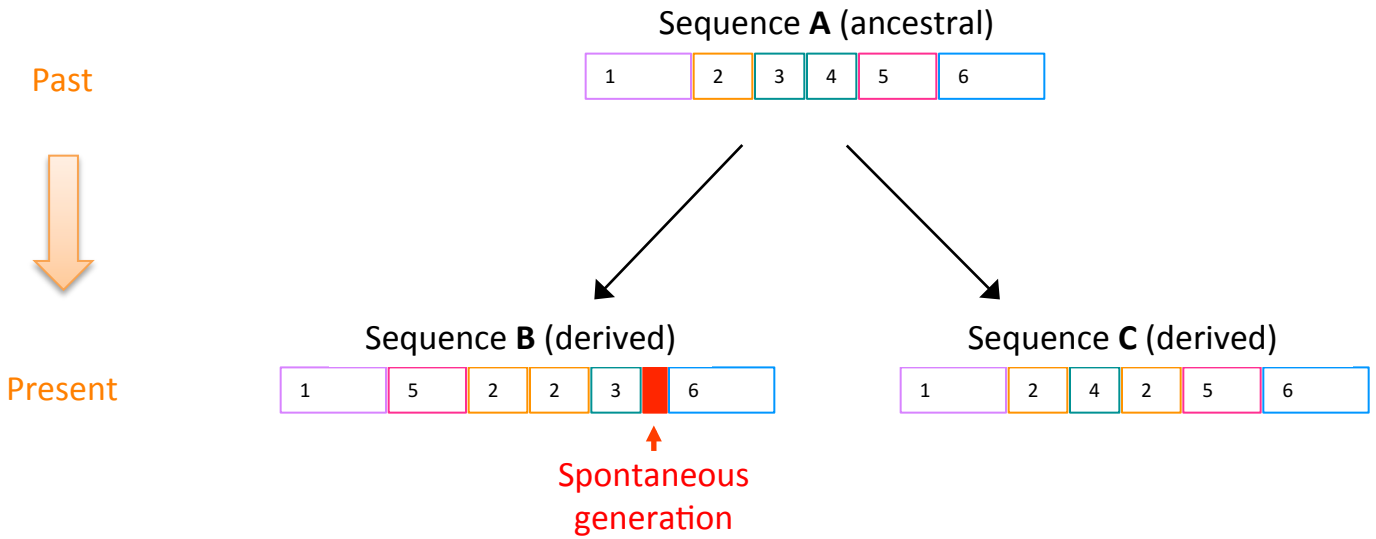
catgtgattagtctggcccta

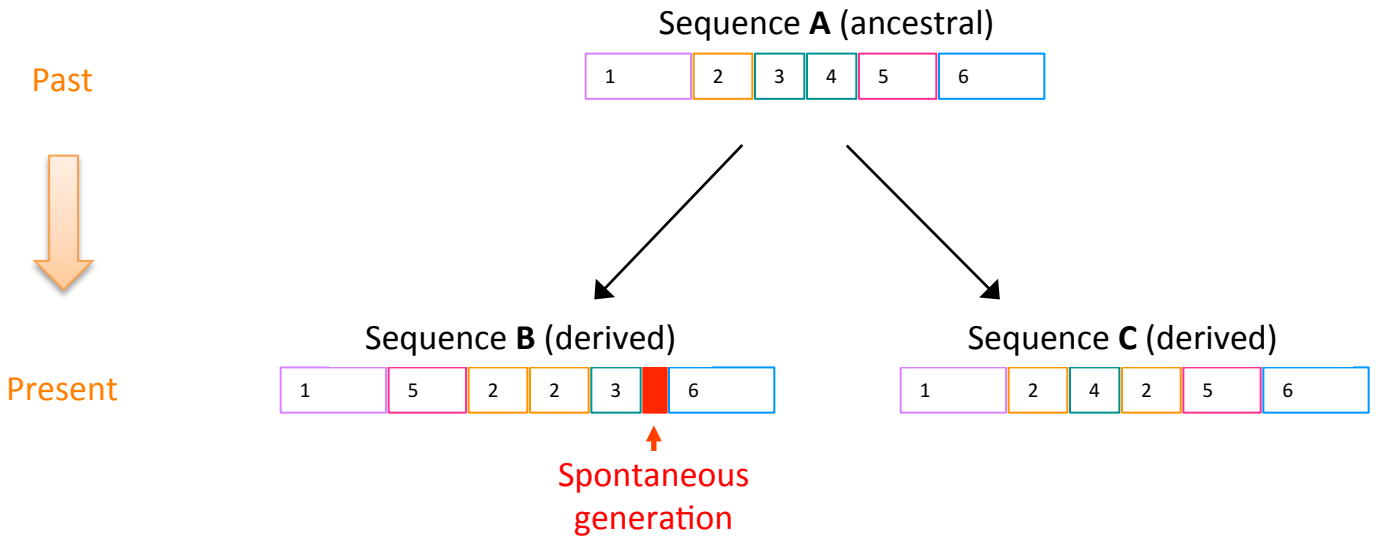
Virus insertion

tagacagctag

↓

tagacaattcgcgataggtagctag


Sequence **A** (ancestral)
Past
1
2
3
4
5
6
Sequence **B** (derived)
Sequence **C** (derived)
Present
1
5
2
2
3
6
1
2
4
2
5
6
Spontaneous generation



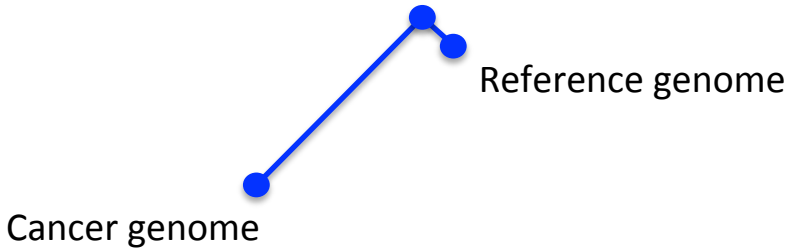
Align **B** to **C** (derived to derived): hard.
Deletions and duplications in both.

Align **A** to **B** (ancestral to derived): easier!

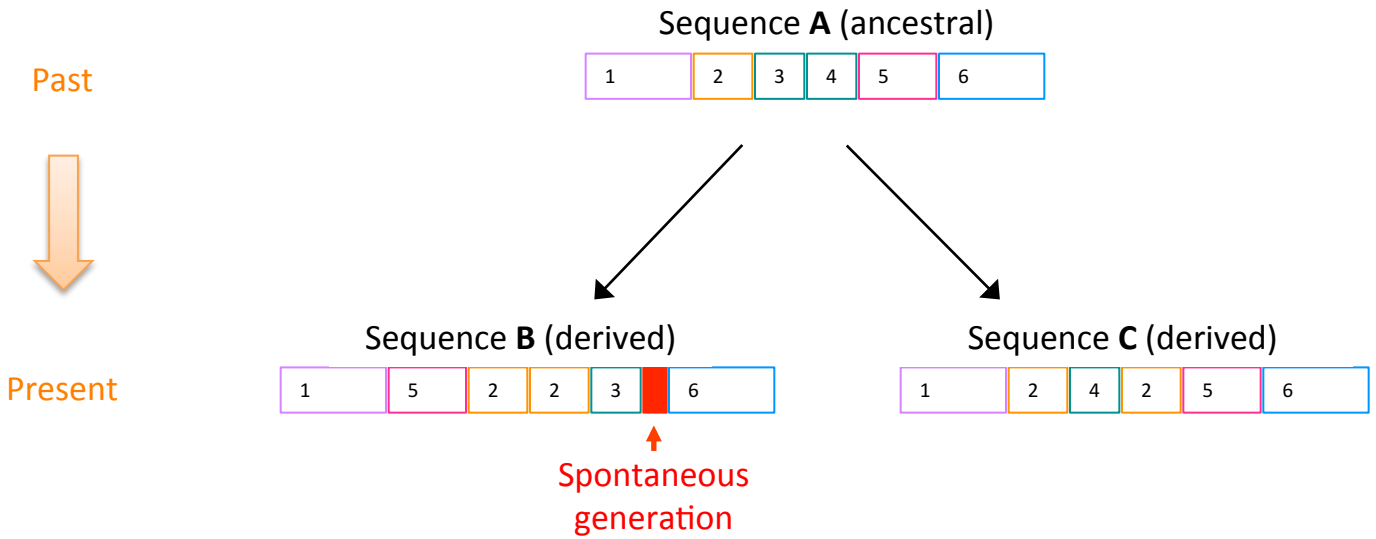
Ancestral sequence has
no lineage-specific deletions,
no lineage-specific duplications.

Therefore: (almost) every part of the derived sequence is descended from a unique part of the ancestral sequence

Example: cancer DNA



To a good approximation, the reference genome is **ancestral**

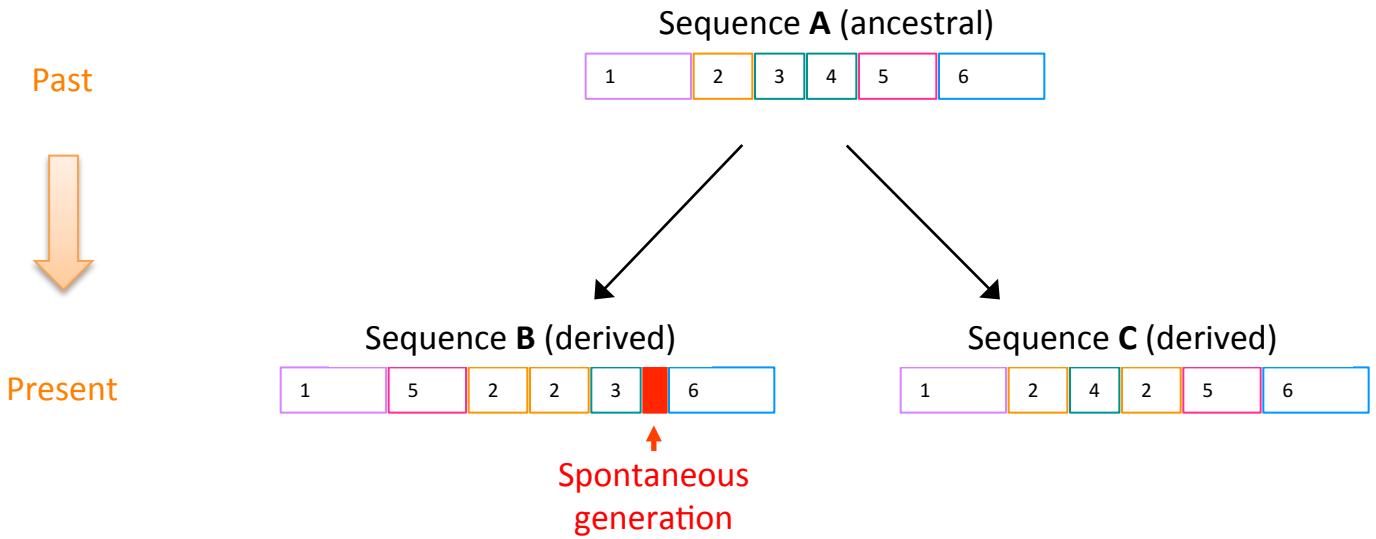


Align **B** to **C** (derived to derived): hard.
Deletions and duplications in both.

Align **A** to **B** (ancestral to derived): easier!

Ancestral sequence has
no lineage-specific deletions,
no lineage-specific duplications.

Therefore: (almost) every part of the derived sequence is descended from a unique part of the ancestral sequence



Align **B** to **C** (derived to derived): hard.
Deletions and duplications in both.

Align **A** to **B** (ancestral to derived): easier!

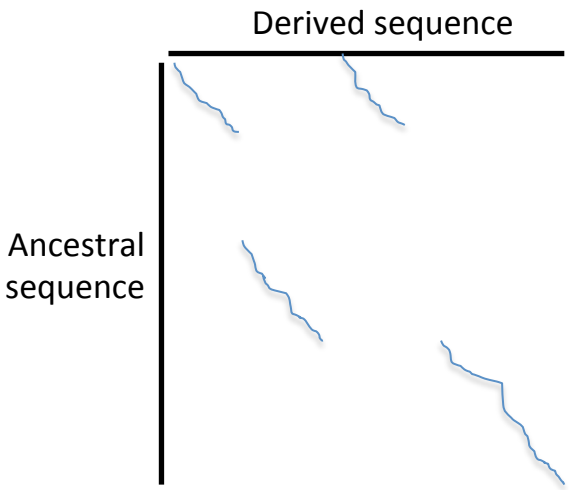
Find optimal division of the derived sequence into parts, and optimal alignment of each part to the ancestor

Ancestral sequence has
no lineage-specific deletions,
no lineage-specific duplications.

Therefore: (almost) every part of the derived sequence is descended from a unique part of the ancestral sequence

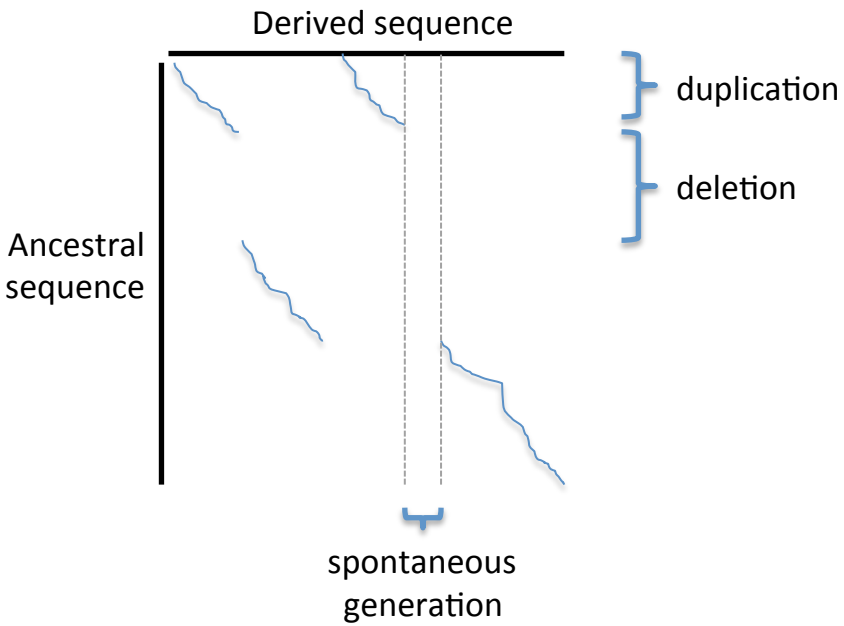
Method: last-split

Finds optimal division of the derived sequence into parts,
and most-likely alignment of each part



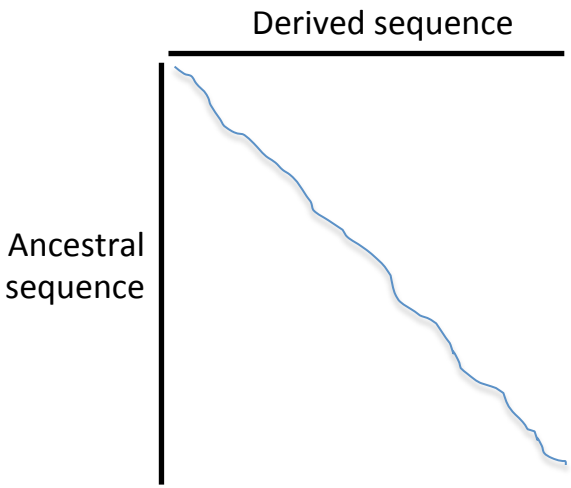
Method: last-split

Finds optimal division of the derived sequence into parts,
and most-likely alignment of each part



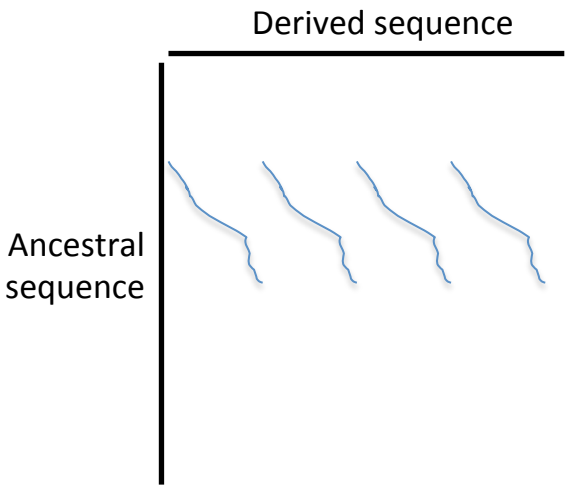
Method: last-split

Finds optimal division of the derived sequence into parts,
and most-likely alignment of each part

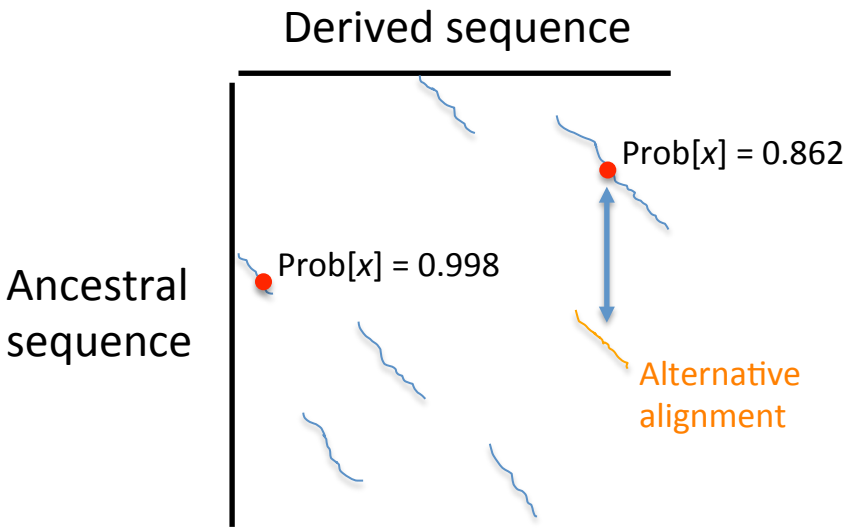


Method: last-split

Finds optimal division of the derived sequence into parts,
and most-likely alignment of each part



Probability that each aligned position is correct



This indicates the reliability / unambiguity of each alignment part

This is important because genomes have many similar duplicated/repeated sequences

Test data

- Nanopore DNA reads from one human (30x genome coverage): https://github.com/nanopore-wgs-consortium/NA12878

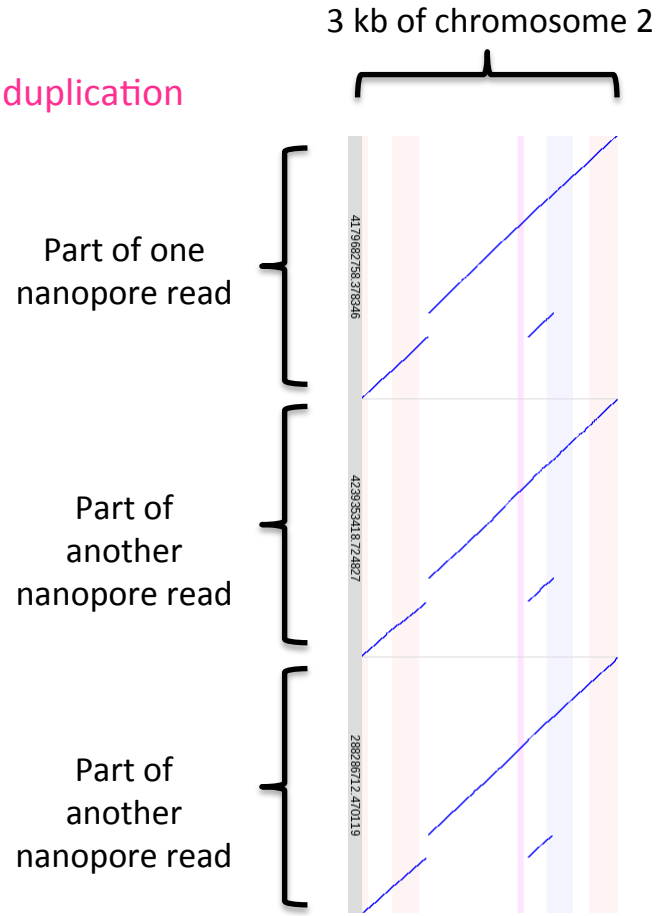
- Align these DNA reads to reference human genome
  - Look for rearrangements

- Problem: the reference genome is not ancestral

- Solution: compare to chimpanzee and gorilla genomes.
  Only trust rearrangements where the human reference has the same arrangement as chimpanzee or gorilla
  (so the human reference **is** ancestral)

Example:

Non-tandem duplication



Vertical stripes

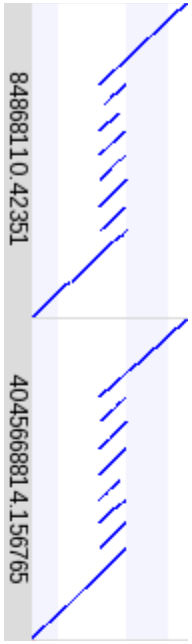
Purple = simple sequence

Pink/blue = transposons

Example:

Tandem heptuplication

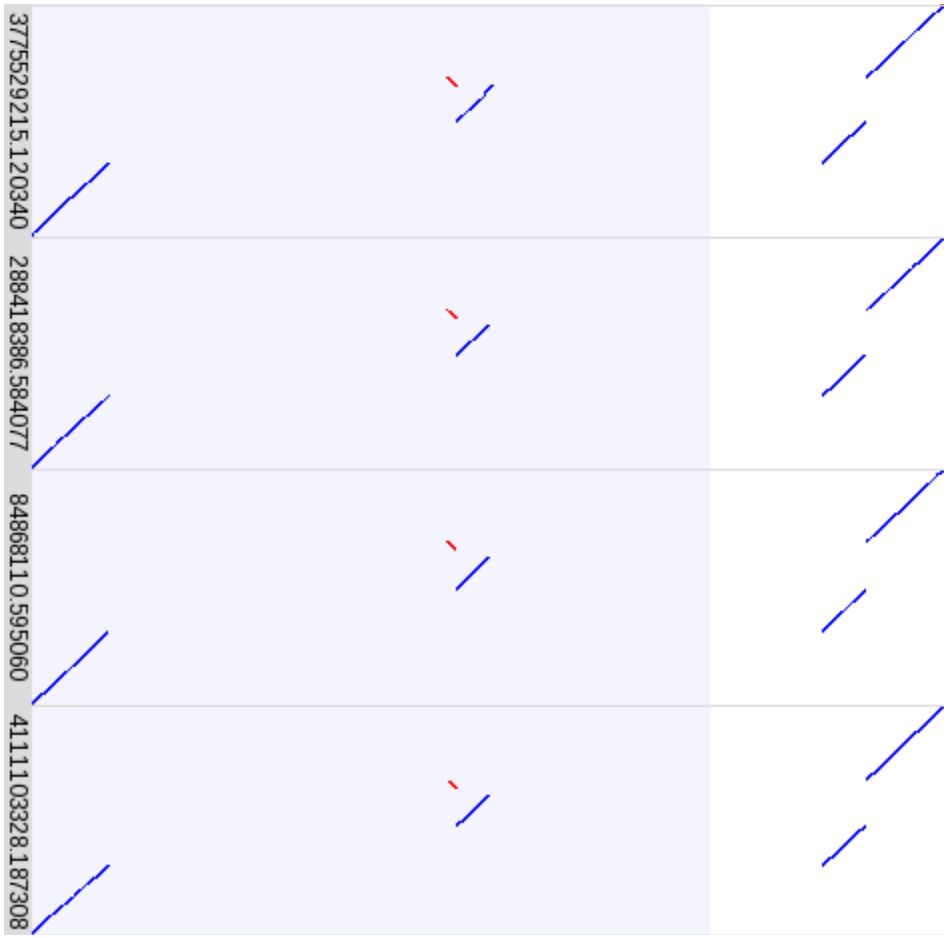
500 bp of chromosome 3



Example:

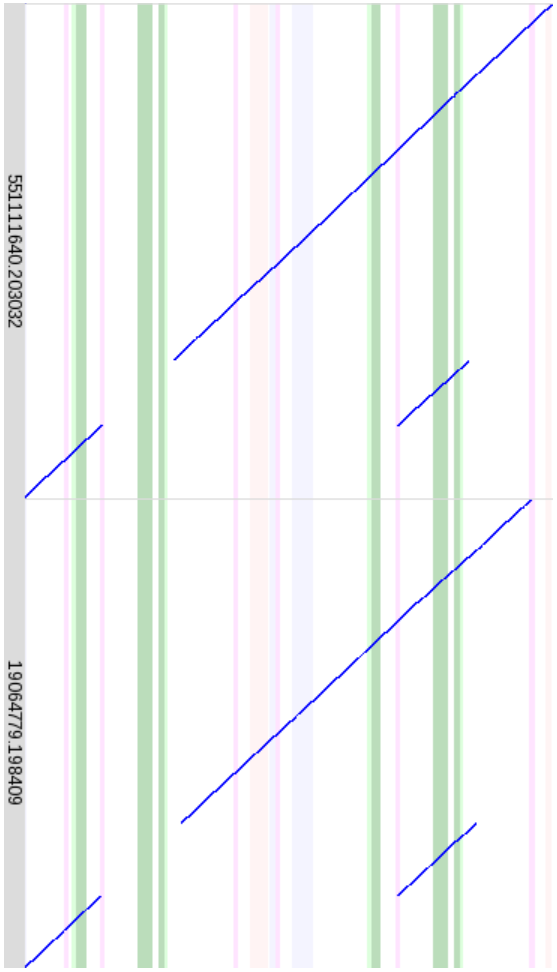
Shatter-and-rejoin

5.5 kb of chromosome 4



Example:
Gene conversion

Hemoglobin gamma 1 Hemoglobin gamma 2



Vertical stripes
Purple = simple sequence
Pink/blue = transposons
Green = exon

I think no other long-read aligner can get this right

New Results

A bestiary of localized sequence rearrangements in human DNA

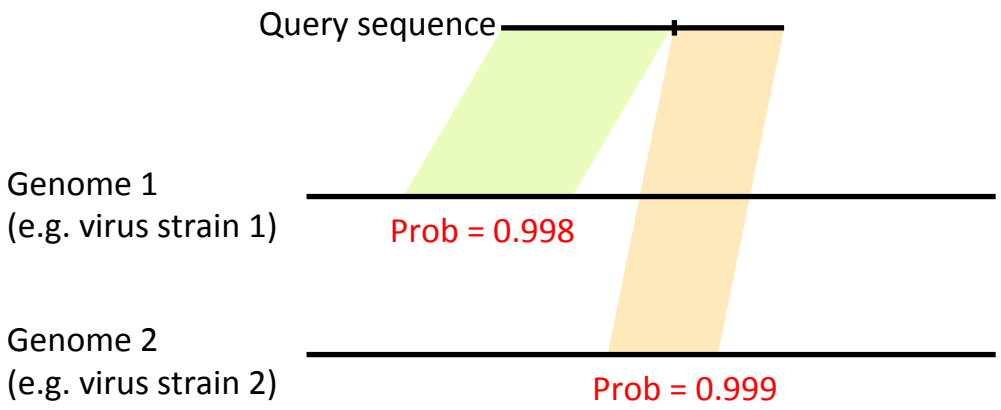
Martin Frith, Sofia Khan

doi: https://doi.org/10.1101/175943

Abstract

Genomes mutate and evolve in ways simple (substitution or deletion of bases) and complex (e.g. chromosome shattering). We do not fully understand what types of complex mutation occur, and we cannot routinely characterize arbitrarily-complex mutations in a high-throughput, genome-wide manner. Long-read DNA sequencing methods (e.g. PacBio, nanopore) are promising for this task, because one read may encompass a whole complex mutation. We describe an analysis pipeline to characterize arbitrarily-complex "local" mutations, i.e. intrachromosomal mutations encompassed by one DNA read. We apply it to nanopore and PacBio reads from one human cell line (NA12878), and survey sequence rearrangements, both real and artifactual. Almost all the real rearrangements belong to recurring patterns or motifs: the most common is tandem multiplication (e.g. heptuplication), but there are also complex patterns such as localized shattering, which resembles DNA damage by radiation. Gene conversions are identified, including one between hemoglobin gamma genes. This study demonstrates a way to find intricate rearrangements with any number of duplications, deletions, and repositionings. It demonstrates a probability-based method to resolve ambiguous rearrangements involving highly similar sequences, as occurs in gene conversion. We present a catalog of local rearrangements in one human cell line, and show which rearrangement patterns do, and do not, occur.

Finding chimeric sequences, e.g. viral recombination



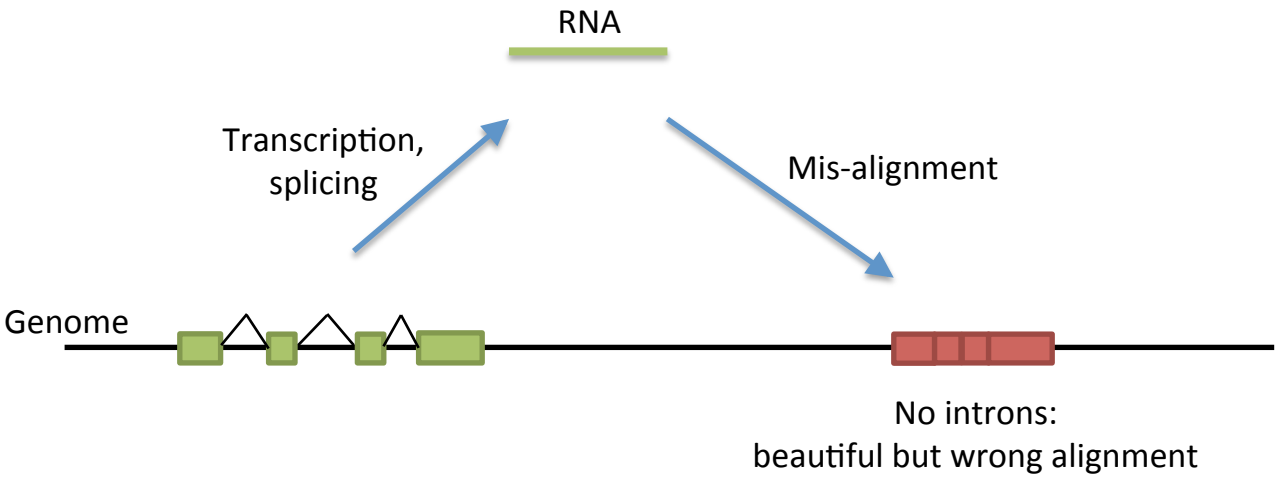
Contents

- Background
- What are we really trying to do?
- Probability-based alignment
- Moar alignment!
- Determining rates of substitution, insertion & deletion
- Alignment ambiguity
- Alignment with duplications & rearrangements
- Aligning spliced RNA or cDNA to a genome

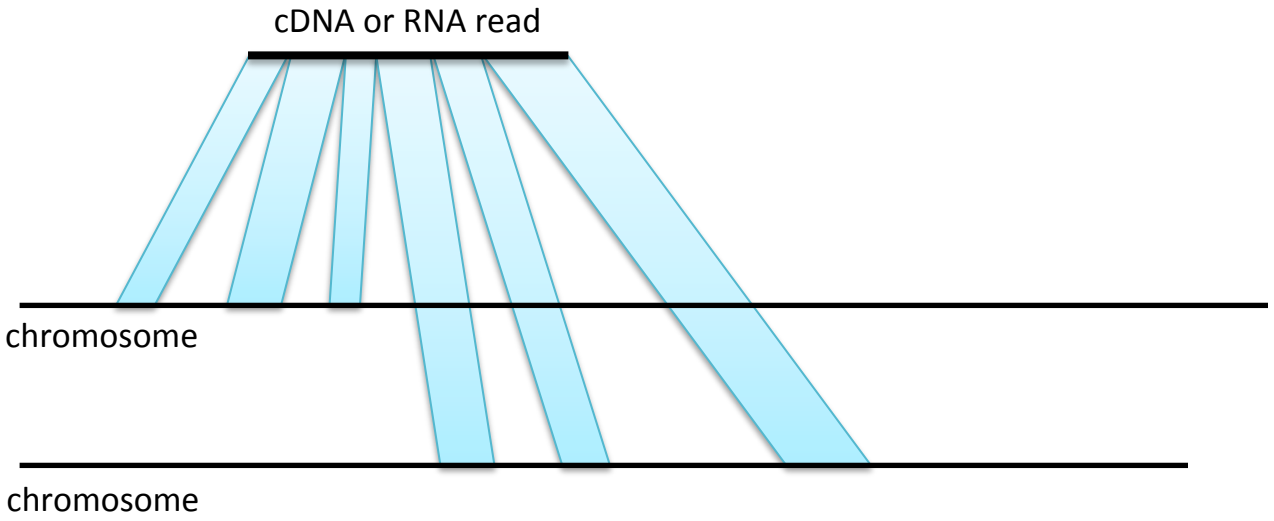
Aligning spliced RNA or cDNA to a genome

- last-split
  - Basically the same method
    - Genome is ancestral, RNA is derived & rearranged
  - Here, the method prefers typical exon-intron structure with GT-AG signals (higher probability)
    - But allows arbitrary rearrangements (e.g. gene fusion)

Problem: processed pseudogenes



Gene fusions



Both parts have splicing → probably reliable (not processed pseudogene)

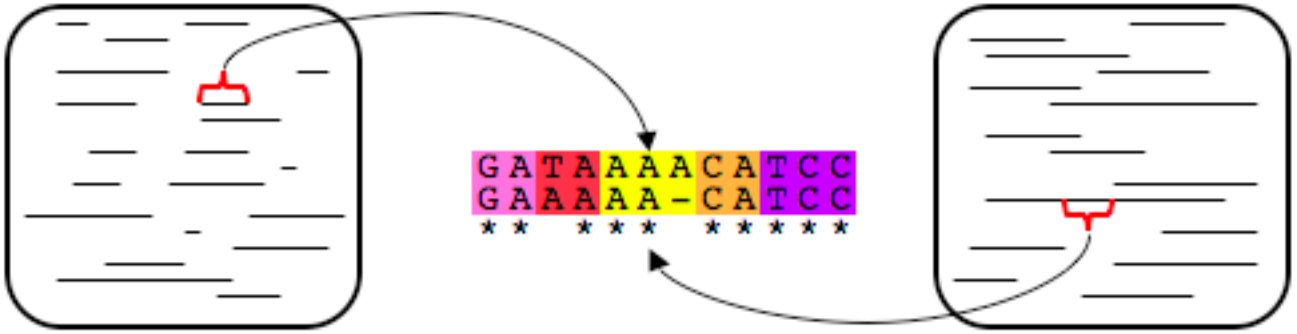
Some tools for post-alignment analysis

- https://github.com/mcfrith/local-rearrangements
  - Draws pictures of rearrangements
- https://github.com/mcfrith/last-rna
  - Scripts for analyzing alignments of cDNA or RNA to a genome
- New and experimental
  - Questions & suggestions welcome!

Our software

LAST

Since 2008



http://last.cbrc.jp/

- **Finds and aligns similar regions**
- Huge data OK
- Low similarity OK
- DNA-protein with frameshifts OK
- Long or short sequences OK
- Split or spliced alignment OK
- Biased sequences OK
  - E.g. malaria: 80% A+T
  - E.g. bisulfite-converted
- Per-column probabilities
- Can use sequence quality data
- Simple-sequence filtering that works

Join us!

GRADUATE SCHOOL OF FRONTIER SCIENCES
THE UNIVERSITY OF TOKYO

http://www.cbms.k.u-tokyo.ac.jp/english/