Introduction to Read-Based Alignment

Michael C. Zody, Ph.D.
Workshop on Genomics
Cesky Krumlov
January 12, 2016

Aligning to a Reference

- Aligning sequences is a classic problem
  - Early bioinformatic problem
  - Very similar to older text matching problems
- Several algorithms exist
  - Tradeoffs of speed versus accuracy, sensitivity
- Sequencing throughput creates new problems
  - Short reads have less information than long seqs
  - Data volume requires faster processing per read

Example of alignment

Read:

TCAACTCTGCCAACACCTTCCTCCTCCAGGAAGCACTCCTGGATTTCCCTCTTGCCAACAAGATTCTGGGAGGGCA

Genome:

ATAAAATGGCCAAAATTAACTAGAAGGTGAGTAGAAACTTAAATAAACTAATTACCATTGATGAGAAAAAAAATC
TGCCACTGAAAAAGGCACCCGGTCCAGAGGGTTTCATGAGCGGGAACTGTAGAAACCTTTCGAATTCAACTCTGC
CAACACCTTCCTCCTCCAGGAAGCACTCCTGGATTTCCCTCTTGCCAACAAGATTCTGGGAGGGCAGCTCCTCCA
ACATGCCCCCAACAGCTCTCTGCAGACATATCATATCATATCATATCTTCCATACCATAACTGCCATGCCATACA

Example of alignment

Read:

TCAACTCTGCCAACACCTTCCTCCTCCAGGAAGCACTCCTGGATTTCCCTCTTGCCAACAAGATTCTGGGAGGGCA

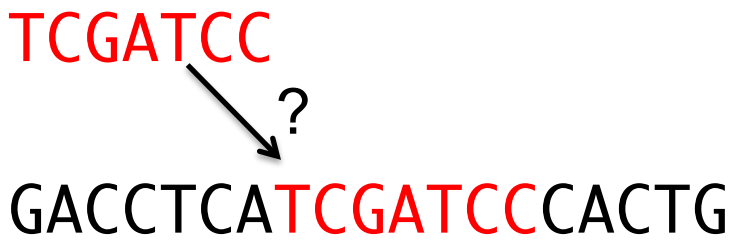
Genome:

ATAAAATGGCCAAAATTAACTAGAAGGTGAGTAGAAACTTAAATAAACTAATTACCATTGATGAGAAAAAAAATC
TGCCACTGAAAAAGGCACCCGGTCCAGAGGGTTTCATGAGCGGGAACTGTAGAAACCTTTCGAATTCAACTCTGC
CAACACCTTCCTCCTCCAGGAAGCACTCCTGGATTTCCCTCTTGCCAACAAGATTCTGGGAGGGCAGCTCCTCCA
ACATGCCCCCAACAGCTCTCTGCAGACATATCATATCATATCATATCTTCCATACCATAACTGCCATGCCATACA

How Would You Find That?

- Brute force comparison
- Smith-Waterman
- Suffix Tree
- Burrows-Wheeler Transform

Brute Force Method



Brute Force Method

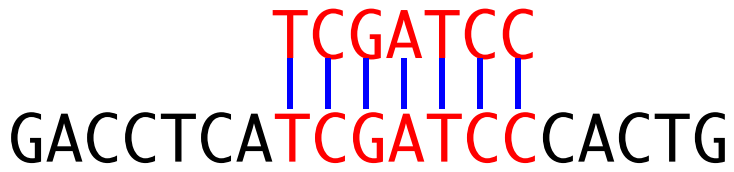
TCGATCC
x
GACCTCATCGATCCCACTG

Brute Force Method

```
 TCGATCC
 x
GACCTCATCGATCCCACTG
```

Brute Force Method

Brute Force Method



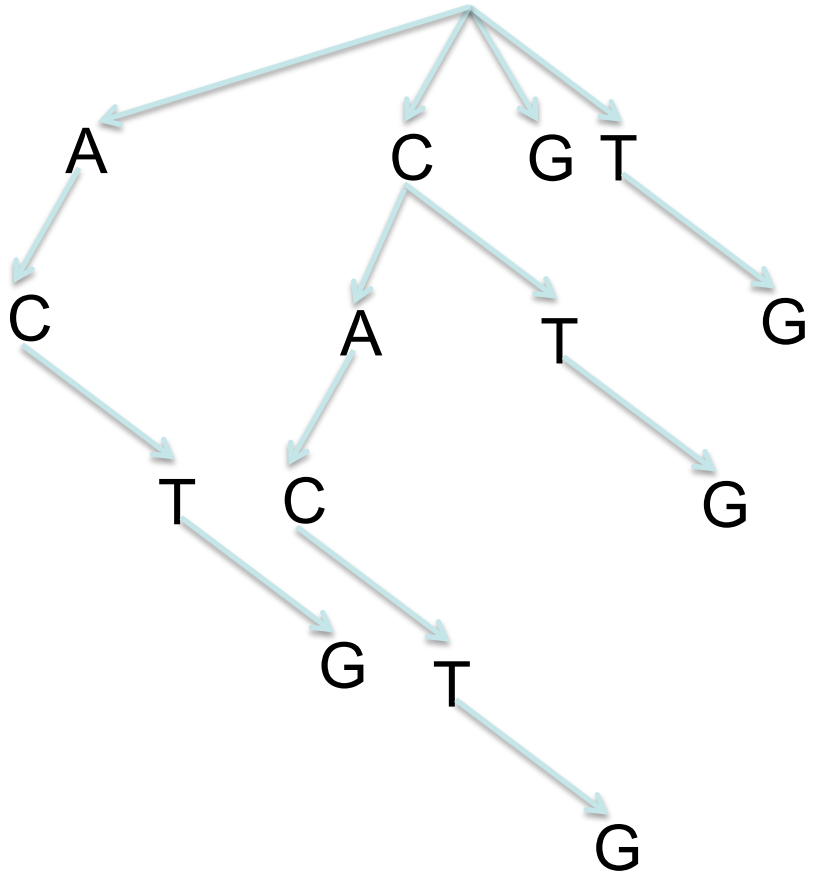
Smith-Waterman

Simplistic Scoring Scheme:
+1 match
-1 mismatch
-1 gap
(no penalty for terminal gaps)

		G	A	C	C	T	C	A	T	C	G	A	T	C	C	C	A	C	T	G
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	-1	-1	-1	-1	1	0	-1	1	0	-1	-1	1	0	-1	-1	-1	-1	1	0
C	0	-1	0	-1	0	0	2	1	0	2	1	0	0	2	1	0	-1	0	0	0
G	0	1	0	-1	-2	-1	1	1	0	1	3	2	1	1	1	0	-1	-2	-1	0
A	0	0	1	0	-1	-2	0	2	1	0	2	4	3	2	1	0	1	0	-1	-2
T	0	-1	0	0	-1	0	-1	-1	3	2	1	1	5	4	3	2	1	0	1	0
C	0	-1	-1	1	1	0	1	0	2	4	3	2	4	6	5	4	3	2	1	0
C	0	-1	-2	0	2	1	1	0	1	3	3	2	3	5	7	6	5	4	3	2

Suffix Tree

GACCTCATCGATCCCACTG
G
TG
CTG
ACTG
CACTG



Suffix Tree

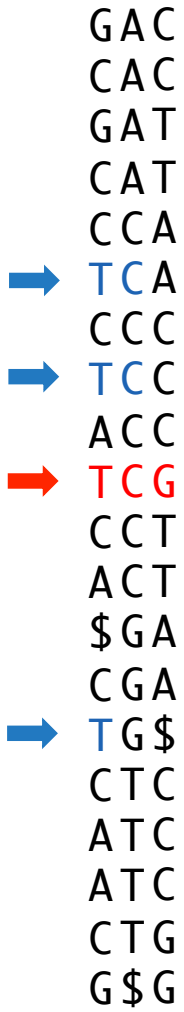
GACCTCATCGATCCCACTG

A				C								G			T			
C		T		A		C		G	T			A		$	C			G
C	T	C		C	T	A	C	T	A	C	G	C	T		A	C	G	$
T	G	C	G	T	C	C	A	C	T	A	$	C	C		T	C	A	
C	$	C	A	G	G	T	C	A	C	T		T	C		C	A	T	
A		A	T	$	A	G	T	T	C	C		C	C		G	C	C	
T		C	C		T	$	G	C	C	G		A	A		A	T	C	
C		T	C		C		$	G	A	A		T	C		T	G	C	
G		G	C		C			A	C	T		C	T		C	$	A	
A		$	A		C			T	T	C		G	G		C		C	
T			C		A			C	G	C		A	$		C		T	
C			T		C			C	$	C		T			A		G	
C			G		T			C		A		C			C		$	
C			$		G			A		C		C			T			
A					$			C		T		C			G			
C								T		G		A			$			
T								G		$		C						
G								$				T						
$												G						
												$						

Burrows-Wheeler Transform

GACCTCATCGATCCCACTG$	ACCTCATCGATCCCACTG$G
ACCTCATCGATCCCACTG$G	ACTG$GACCTCATCGATCCC
CCTCATCGATCCCACTG$GA	ATCCCACTG$GACCTCATCG
CTCATCGATCCCACTG$GAC	ATCGATCCCACTG$GACCTC
TCATCGATCCCACTG$GACC	CACTG$GACCTCATCGATCC
CATCGATCCCACTG$GACCT	CATCGATCCCACTG$GACCT
ATCGATCCCACTG$GACCTC	CCACTG$GACCTCATCGATC
TCGATCCCACTG$GACCTCA	CCCACTG$GACCTCATCGAT
CGATCCCACTG$GACCTCAT	CCTCATCGATCCCACTG$GA
GATCCCACTG$GACCTCATC	CGATCCCACTG$GACCTCAT
ATCCCACTG$GACCTCATCG	CTCATCGATCCCACTG$GAC
TCCCACTG$GACCTCATCGA	CTG$GACCTCATCGATCCCA
CCCACTG$GACCTCATCGAT	GACCTCATCGATCCCACTG$
CCACTG$GACCTCATCGATC	GATCCCACTG$GACCTCATC
CACTG$GACCTCATCGATCC	G$GACCTCATCGATCCCACT
ACTG$GACCTCATCGATCCC	TCATCGATCCCACTG$GACC
CTG$GACCTCATCGATCCCA	TCCCACTG$GACCTCATCGA
TG$GACCTCATCGATCCCAC	TCGATCCCACTG$GACCTCA
G$GACCTCATCGATCCCACT	TG$GACCTCATCGATCCCAC
$GACCTCATCGATCCCACTG	$GACCTCATCGATCCCACTG

How Do We Use This To Align?



- Start with the transform column
- My read starts with a T, so I want rows with Ts in them
- This column gives me all the single nucleotide counts
- Sort the single nucleotide counts to get the alphabetically first column
- Now these two columns give me all the dinucleotide counts
- Sort those to get the alphabetically first two columns
- Now there is only one place my read can match

Heuristic Improvements

- Seeding
  - Use a hash of exact matches to limit searches
  - Use varying hash types
- Limiting mismatches or indels
  - Constrain poor quality searches
  - Known as ‘banded’ in Smith-Waterman
- Using base quality to guide backtracking

Common Short Read Aligners

- Seed and gap-free extend
  - Eland, MAQ, Mapreads
- Seed and Smith-Waterman extend
  - Mosaik, BFAST, Novoalign
- BWA align gap-free
  - Bowtie
- BWA align with gaps
  - BWA, Bowtie2

Blasr

- Designed for long error-prone reads (PacBio)
- Combines multiple methods
- Starts by finding short exact matches using suffix or B-W
- Next locally identifies a linear chain of shorter exact matches
- Performs banded Smith-Waterman constrained by the shorter exact matches