Genome 540 Discussion

January 8th, 2024

Clifford Rostomily

Genome Sciences
UNIVERSITY OF WASHINGTON

Agenda

- Assignment #1
- Questions
- Random Stuff

Assignment #1

Assignment Overview

- Read in two fasta files
  - Track # non-alpha characters
  - Also track base counts
- Combine the 3 sequences and store subseqs as a pointer array/vector
  - Forward of seq1, forward and reverse of seq2 = 3
- Implement and run the suffix array algorithm
  - Returns a sorted list of pointers
- Iterate through results
  - Track longest match length of each seq1 suffix to either the forward or reverse strand of seq2
  - Track the longest overall match

Fasta format

- File format for storing sequences
- Can store multiple sequences
- Sequences are preceded by a single header line denoted by a “>”

> my_sequence1

TCGATCGATGGCTTCGGATGCGCTTAG

> my_sequence2

GCTGCGCGAGAACAACGTAGCATGGC

- Can be used to store protein and nucleic acid sequences
  - The extension can tell you more about what is being stored

Extension	Meaning	Notes
fasta, fas, fa[9]	generic FASTA	Any generic FASTA file
fna	FASTA nucleic acid	Used generically to specify nucleic acids
ffn	FASTA nucleotide of gene regions	Contains coding regions for a genome
faa	FASTA amino acid	Contains amino acid sequences
mpfa	FASTA amino acids	Contains multiple protein sequences
frn	FASTA non-coding RNA	Contains non-coding RNA regions for a genome, e.g. tRNA, rRNA

https://en.wikipedia.org/wiki/FASTA_format

Fasta IUPAC ambiguity codes

Nucleic Acid Code	Meaning	Mnemonic
A	A	**A**denine
C	C	**C**ytosine
G	G	**G**uanine
T	T	**T**hymine
U	U	**U**racil
(i)	i	**i**nosine (non-standard)
R	A or G (I)	pu**R**ine
Y	C, T or U	p**Y**rimidines
K	G, T or U	bases which are **K**etones
M	A or C	bases with a**M**ino groups
S	C or G	**S**trong interaction
W	A, T or U	**W**eak interaction
B	not A (i.e. C, G, T or U)	**B** comes after A
D	not C (i.e. A, G, T or U)	**D** comes after C
H	not G (i.e., A, C, T or U)	**H** comes after G
V	neither T nor U (i.e. A, C or G)	**V** comes after U
N	A C G T U	**N**ucleic acid
-	gap of indeterminate length	

Non-alphabetic characters

- Exclude the header line
- Exclude white space (e.g. spaces)
- Include only digits from seq position numbers

Small Example

Steps:

Seq. 1: TCAA

Seq. 2: ACTG

Seq. 2: CAGT (reverse complement)

Pointer array to suffixes

TCAA
CGA
GA
A
ACTG
CTG
TG
G
CAGT
AGT
GT
T

Sorted

A
AA
AGT
AATG
CAA
CAGT
CTG
G
GT
T
TCAA
TG

Iterate through sorted list

- find the longest match to seq1
- make a histogram of seq1 longest matches

Things to consider:

- Multiple seq1 suffixes in a row
- Need to look above and below to find longest match

Match Lengths:
1: 3
2: 1

The longest match length: 2
Number of match strings: 1

Questions?

- Unless you specifically ask me not to bring it up I will try to cover common questions asked on slack during the next class discussion.

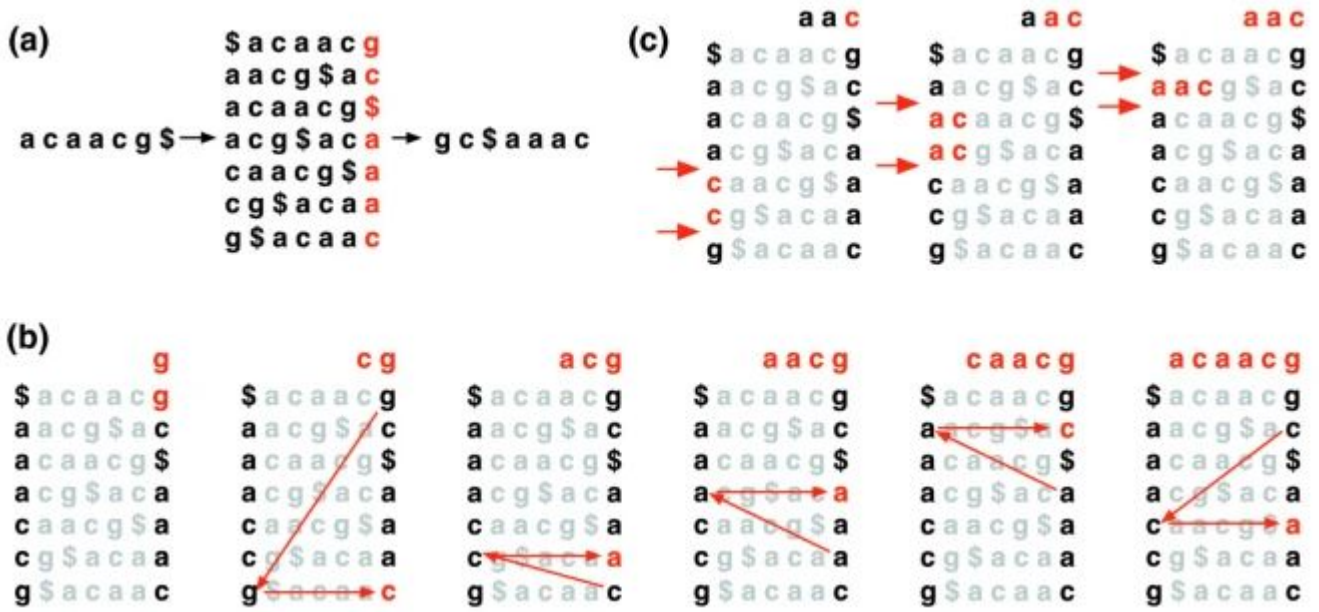
Random Stuff

The Burrows Wheeler Transform

Transformation				
1. Input	2. All rotations	3. Sort into lexical order	4. Take the last column	5. Output
^BANANA$	^BANANA$ $^BANANA A$^BANAN NA$^BANA ANA$^BAN NANA$^BA ANANA$^B BANANA$^	ANANA$^B ANA$^BAN A$^BANAN BANANA$^ NANA$^BA NA$^BANA ^BANANA$ $^BANANA	ANANA$^B ANA$^BAN A$^BANAN BANANA$^ NANA$^BA NA$^BANA ^BANANA$ $^BANANA	BNN^AA$A

- Used originally for compression
- The Bowtie aligner uses it for compression and indexing

Bowtie



Langmead, **Trapnell**, Pop, Salzberg (2009) Genome Biology