Read Mapping

Michael Schatz

Feb 15, 2018

Lecture 6: Applied Comparative Genomics

Mission Impossible

1. ***Setup VirtualBox***
2. ***Initialize Tools***
3. ***Download Reference Genome & Reads***
4. ***Decode the secret message***
    1. *Estimate coverage, check read quality*
    2. *Check kmer distribution*
    3. *Assemble the reads with spades*
    4. *Align to reference with MUMmer*
    5. *Extract foreign sequence*
    6. *dna-encode.pl -d*

https://github.com/schatzlab/appliedgenomics2018/blob/master/assignments/assignment2/README.md

Assignment 3: Due Thursday Feb 22

Assignment 3: Genome Assembly, Phylogenetics, and the BWT

Assignment Date: Thursday, Feb. 15, 2018
Due Date: Thursday, Feb. 22, 2018 @ 11:59pm

Question 1. de Bruijn Graph construction [10 pts]

- Q1a. Draw (by hand or by code) the de Bruijn graph for the following reads using k=3 (assume all reads are from the forward strand, no sequencing errors, complete coverage of the genome)

```
ATTC
ATTG
GATT
CTTA
GATT
TATT
TGAT
TCTT
TGAT
TTAT
TTGA
TTCT
TTGA
```

- Q1b. Assume that the maximum number of occurrences of any 3-mer in the actual genome is 3 using the k-mers from Q1a. Write one possible genome sequence
- Q1c. What is the longest repeat?

Question 2. Phylogenetics Analysis [10 pts]

Your colleague is developing an experimental and computational protocol to determine the species present in food samples based on DNA sequencing. (See here for a technology working towards making this a reality!) She extracted DNA from a mixed-meat sausage a 100bp Illumina sequencing. When the data returns, she uses a short-read aligner such as Bowtie2 or BWA to align the sequencing reads. As the references, she chose several genomes of animals whose meat is commonly consumed, including chicken and pig and cow. common genomes. Next, she extracts the unmapped reads and runs a short-read assembler such as Spades on those reads. She only gets a few contigs that are longer than a few hundred base pairs.

1. Suggest two reasons there are only a few, short contigs assembled from non-mapping reads. (2)

She asks for your help in finding the origin of these "mystery meat" contigs. Fortunately you are familiar with genomic databases and offer to help her out. You use query the NCBI's database of reference genome assemblies with the longest contigs using the BLAST to alignments between your sequence and a database. One contig you examine has several high E-value alignments to scaffolds in the Macropus eugenii genome assembly. Two of the alignments are in annotated gene regions. However, the wallaby genome assembly sea

2. Based on the link above, give two indicators that this genome assembly is poor quality. (2)

Because the assembly is rough, you are suspicious that the contig has more than one alignment. It overlaps more than one annotated gene. Could there be a duplicated region or misassembly in the reference genome? Or does the tammar wallaby actually have genes t to align to both?

Homologous genes are genes with a shared evolutionary history. Homologous genes in the same genome arise from a gene duplication event long ago in evolution. Homologous genes in the same genome are called paralogs . Paralogs usually have detectable sequence ENSMEUG00000000896 and ENSMEUG00000000891 (the annotated genes within two of this contig's alignments) are paralogs. You decide to build a phylogenetic tree of these genes, as well as some sequences from other species to see whether these genes are par

Here are some protein sequences of some hits from a blastx search including the two sequences from M. eugenii. multipleFASTA Some proteins are annotated "hemoglobin epsilon" and others are annotated "hemoglobin beta" (B and E in the sequence names in the fi

3. Use the web version of MUSCLE to create a multiple sequence alignment. The tool outputs a neighbor-joining, binary phylogenetic tree. Because MUSCLE's built in tree graphic is very poor, download the data in Newick format, and open the file in visualization softw based tool such as Tree. Include an image of the tree in your report. Feel free to explore a variety of visualization options, but just make sure the leaf labels are readable and the branches have proportional length.

a. What do the leaves of the tree represent? Is the tree rooted or unrooted? (1)

b. Propose a location for the root of the tree, and justify your answer. (Mark it on the image of the tree) (1)

c. Do you think the "B" and "E" genes are paralogs? Justify your answer by referring to the tree. (2)

Here is the output from MrBayes, a Bayesian MCMC tree algorithm, run on the same protein sequences.

Table 2 | **A summary of strengths and weaknesses of different tree reconstruction methods**

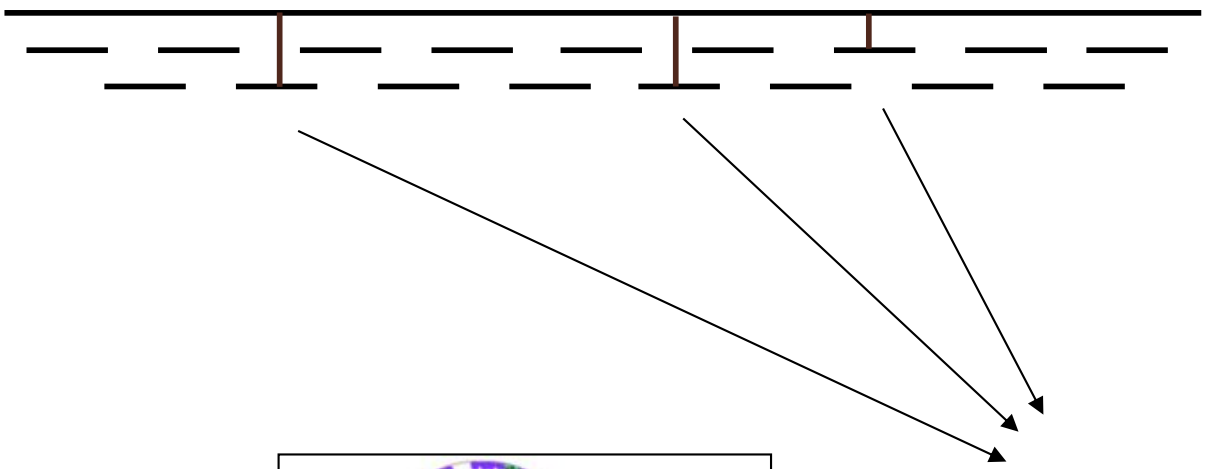
Strengths	Weaknesses
Parsimony methods	
• Simplicity and intuitive appeal • The only framework appropriate for some data (such as SINES and LINES)	• Assumptions are implicit and poorly understood • Lack of a model makes it nearly impossible to incorporate our knowledge of sequence evolution • Branch lengths are substantially underestimated when substitution rates are high • Maximum parsimony may suffer from long-branch attraction
Distance methods	
• Fast computational speed • Can be applied to any type of data as long as a genetic distance can be defined • Models for distance calculation can be chosen to fit data	• Most distance methods, such as neighbour joining, do not consider variances of distance estimates • Distance calculation is problematic when sequences are divergent and involve many alignment gaps • Negative branch lengths are not meaningful
Likelihood methods	
• Can use complex substitution models to approach biological reality • Powerful framework for estimating parameters and testing hypotheses	• Maximum likelihood iteration involves heavy computation • The topology is not a parameter so that it is difficult to apply maximum likelihood theory for its estimation. Bootstrap proportions are hard to interpret
Bayesian methods	
• Can use realistic substitution models, as in maximum likelihood • Prior probability allows the incorporation of information or expert knowledge • Posterior probabilities for trees and clades have easy interpretations	• Markov chain Monte Carlo (MCMC) involves heavy computation • In large data sets, MCMC convergence and mixing problems can be hard to identify or rectify • Uninformative prior probabilities may be difficult to specify. Multidimensional priors may have undue influence on the posterior without the investigator's knowledge • Posterior probabilities often appear too high • Model selection involves challenging computation[138,139]

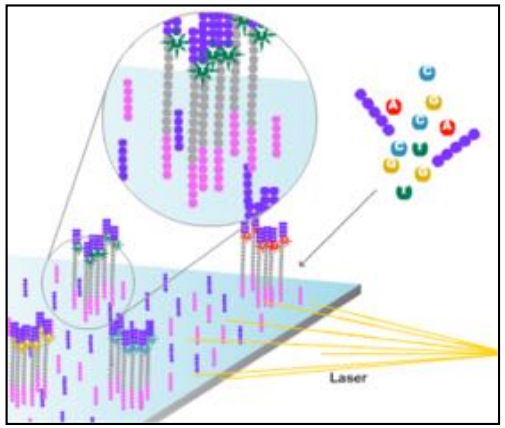
Molecular phylogenetics: principles and practice.
Yang and Rannala, Nature Reviews Genetics 2012

Part 1: Suffix Arrays

Personal Genomics

How does your genome compare to the reference?



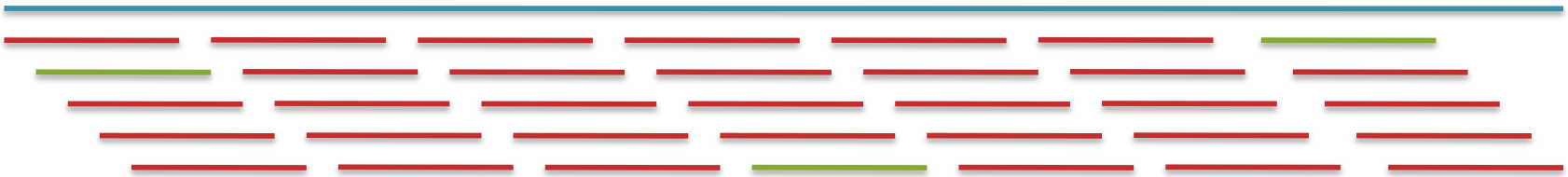


Heart Disease

Cancer

Presidential smile

Brute Force Analysis



- Brute Force:
  - At every possible offset in the genome:
    - Do all of the characters of the query match?

- Analysis
  - Simple, easy to understand
  - Genome length = n [3B]
  - Query length = m [7]
  - Comparisons: (n-m+1) * m [21B]

- Overall runtime: O(nm)

[How long would it take if we double the genome size, read length?]

[How long would it take if we double both?]

Brute Force Reflections

Why check every position?

– GATTACA can't possibly start at position 15 [WHY?]

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	…
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	…
								G	A	T	T	A	C	A	

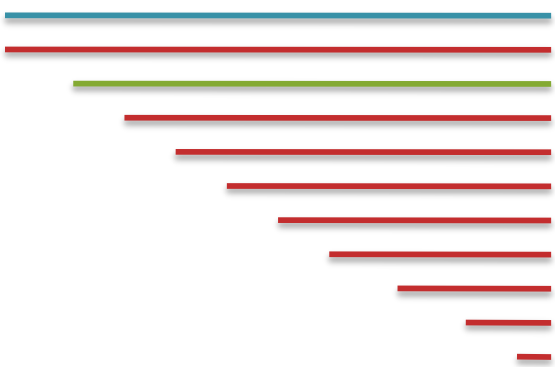
– Improve runtime to O(n + m) [3B + 7]
  - If we double both, it just takes twice as long
  - Knuth-Morris-Pratt, 1977
  - Boyer-Moyer, 1977, 1991

– For one-off scans, this is the best we can do (optimal performance)
  - We have to read every character of the genome, and every character of the query
  - For short queries, runtime is dominated by the length of the genome

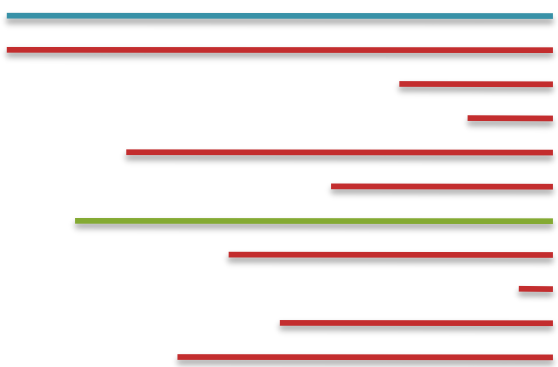
Suffix Arrays: Searching the Phone Book

- What if we need to check many queries?
  - We don't need to check every page of the phone book to find 'Schatz'
  - Sorting alphabetically lets us immediately skip 96% (25/26) of the book *without any loss in accuracy*

- Sorting the genome: Suffix Array (Manber & Myers, 1991)
  - Sort every suffix of the genome



Split into n suffixes



Sort suffixes alphabetically

[Challenge Question: How else could we split the genome?]

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15;

Lo →

#	Sequence	Pos
1	ACAGATTACC…	6
2	ACC…	13
3	AGATTACC…	8
4	ATTACAGATTACC…	3
5	ATTACC…	10
6	C…	15
7	CAGATTACC…	7
8	CC…	14
9	GATTACAGATTACC…	2
10	GATTACC…	9
11	TACAGATTACC…	5
12	TACC…	12
13	TGATTACAGATTACC…	1
14	TTACAGATTACC…	4
15	TTACC…	11

Hi →

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC

Lo →

#	Sequence	Pos
1	ACAGATTACC…	6
2	ACC…	13
3	AGATTACC…	8
4	ATTACAGATTACC…	3
5	ATTACC…	10
6	C…	15
7	CAGATTACC…	7
8	CC…	14
9	GATTACAGATTACC…	2
10	GATTACC…	9
11	TACAGATTACC…	5
12	TACC…	12
13	TGATTACAGATTACC…	1
14	TTACAGATTACC…	4
15	TTACC…	11

Hi →

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC
    => Higher: Lo = Mid + 1

Lo →

#	Sequence	Pos
1	ACAGATTACC…	6
2	ACC…	13
3	AGATTACC…	8
4	ATTACAGATTACC…	3
5	ATTACC…	10
6	C…	15
7	CAGATTACC…	7
8	CC…	14
9	GATTACAGATTACC…	2
10	GATTACC…	9
11	TACAGATTACC…	5
12	TACC…	12
13	TGATTACAGATTACC…	1
14	TTACAGATTACC…	4
15	TTACC…	11

Hi →

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC
    => Higher: Lo = Mid + 1

  - Lo = 9; Hi = 15;

	#	Sequence	Pos
	1	ACAGATTACC…	6
	2	ACC…	13
	3	AGATTACC…	8
	4	ATTACAGATTACC…	3
	5	ATTACC…	10
	6	C…	15
	7	CAGATTACC…	7
Lo →	8	CC…	14
	9	GATTACAGATTACC…	2
	10	GATTACC…	9
	11	TACAGATTACC…	5
	12	TACC…	12
	13	TGATTACAGATTACC…	1
	14	TTACAGATTACC…	4
Hi →	15	TTACC…	11

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC
    => Higher: Lo = Mid + 1

  - Lo = 9; Hi = 15; Mid = (9+15)/2 = 12
  - Middle = Suffix[12] = TACC

	#	Sequence	Pos
	1	ACAGATTACC…	6
	2	ACC…	13
	3	AGATTACC…	8
	4	ATTACAGATTACC…	3
	5	ATTACC…	10
	6	C…	15
	7	CAGATTACC…	7
	8	CC…	14
Lo →	9	GATTACAGATTACC…	2
	10	GATTACC…	9
	11	TACAGATTACC…	5
	12	TACC…	12
	13	TGATTACAGATTACC…	1
	14	TTACAGATTACC…	4
Hi →	15	TTACC…	11

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC
    => Higher: Lo = Mid + 1

  - Lo = 9; Hi = 15; Mid = (9+15)/2 = 12
  - Middle = Suffix[12] = TACC
    => Lower: Hi = Mid - 1

  - Lo = 9; Hi = 11;

	#	Sequence	Pos
	1	ACAGATTACC…	6
	2	ACC…	13
	3	AGATTACC…	8
	4	ATTACAGATTACC…	3
	5	ATTACC…	10
	6	C…	15
	7	CAGATTACC…	7
	8	CC…	14
Lo →	9	GATTACAGATTACC…	2
	10	GATTACC…	9
Hi →	11	TACAGATTACC…	5
	12	TACC…	12
	13	TGATTACAGATTACC…	1
	14	TTACAGATTACC…	4
	15	TTACC…	11

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC
    => Higher: Lo = Mid + 1

  - Lo = 9; Hi = 15; Mid = (9+15)/2 = 12
  - Middle = Suffix[12] = TACC
    => Lower: Hi = Mid - 1

  - Lo = 9; Hi = 11; Mid = (9+11)/2 = 10
  - Middle = Suffix[10] = GATTACC

	#	Sequence	Pos
	1	ACAGATTACC…	6
	2	ACC…	13
	3	AGATTACC…	8
	4	ATTACAGATTACC…	3
	5	ATTACC…	10
	6	C…	15
	7	CAGATTACC…	7
	8	CC…	14
Lo →	9	GATTACAGATTACC…	2
	10	GATTACC…	9
Hi →	11	TACAGATTACC…	5
	12	TACC…	12
	13	TGATTACAGATTACC…	1
	14	TTACAGATTACC…	4
	15	TTACC…	11

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC
    => Higher: Lo = Mid + 1

  - Lo = 9; Hi = 15; Mid = (9+15)/2 = 12
  - Middle = Suffix[12] = TACC
    => Lower: Hi = Mid - 1

  - Lo = 9; Hi = 11; Mid = (9+11)/2 = 10
  - Middle = Suffix[10] = GATTACC
    => Lower: Hi = Mid - 1

  - Lo = 9; Hi = 9;

Lo
Hi

#	Sequence	Pos
1	ACAGATTACC…	6
2	ACC…	13
3	AGATTACC…	8
4	ATTACAGATTACC…	3
5	ATTACC…	10
6	C…	15
7	CAGATTACC…	7
8	CC…	14
9	GATTACAGATTACC…	2
10	GATTACC…	9
11	TACAGATTACC…	5
12	TACC…	12
13	TGATTACAGATTACC…	1
14	TTACAGATTACC…	4
15	TTACC…	11

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC
    => Higher: Lo = Mid + 1

  - Lo = 9; Hi = 15; Mid = (9+15)/2 = 12
  - Middle = Suffix[12] = TACC
    => Lower: Hi = Mid - 1

  - Lo = 9; Hi = 11; Mid = (9+11)/2 = 10
  - Middle = Suffix[10] = GATTACC
    => Lower: Hi = Mid - 1

  - Lo = 9; Hi = 9; Mid = (9+9)/2 = 9
  - Middle = Suffix[9] = GATTACA…
    => Match at position 2!

#	Sequence	Pos
1	ACAGATTACC…	6
2	ACC…	13
3	AGATTACC…	8
4	ATTACAGATTACC…	3
5	ATTACC…	10
6	C…	15
7	CAGATTACC…	7
8	CC…	14
9	GATTACAGATTACC…	2
10	GATTACC…	9
11	TACAGATTACC…	5
12	TACC…	12
13	TGATTACAGATTACC…	1
14	TTACAGATTACC…	4
15	TTACC…	11

Lo
Hi

Binary Search Analysis

- Binary Search

  Initialize search range to entire list

  mid = (hi+lo)/2; middle = suffix[mid]

  if query matches middle: done

  else if query < middle: pick low range

  else if query > middle: pick hi range

  Repeat until done or empty range [WHEN?]

- Analysis
  - More complicated method
  - How many times do we repeat?
    - How many times can it cut the range in half?
    - Find smallest x such that: $n/(2^x) \leq 1$; $x = \lg_2(n)$ [32]

- Total Runtime: O(m lg n)
  - More complicated, but much faster!
  - Looking up a query loops 32 times instead of 3B

[How long does it take to search 6B or 24B nucleotides?]

Binary Search Analysis

- Binary Search

  Initialize search range to entire list

  mid = (hi+lo)/2; middle = suffix[mid]

  if query matches middle: done

  else if query < middle: pick low range

  else if query > middle: pick hi range

  Repeat until done or empty range [WHEN?]

- Analysis
  - More complicated method
  - How many times do we repeat?
    - How many times can it cut the range in half?
    - Find smallest x such that: $n/(2^x) \leq 1$; $x = \lg_2(n)$ [32]

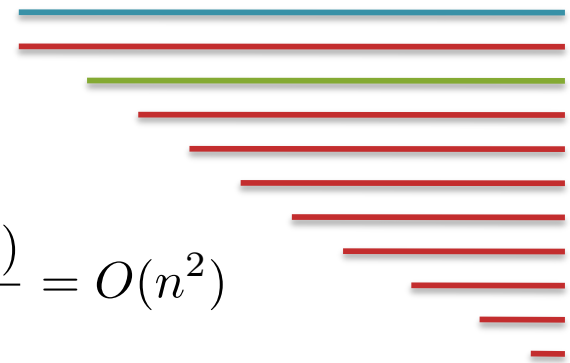
- Total Runtime: O(m lg n)
  - More complicated, but much faster!
  - Looking up a query loops 32 times instead of 3B

Can be reduced to O(m + lg n)
using an auxiliary data structure called the LCP array

Suffix Array Construction

- How can we store the suffix array?

  [How many characters are in all suffixes combined?]



$$S = 1 + 2 + 3 + \cdots + n = \sum_{i=1}^{n} i = \frac{n(n+1)}{2} = O(n^2)$$

  - Hopeless to explicitly store 4.5 billion billion characters
  - Instead use implicit representation
    - Keep 1 copy of the genome, and a list of sorted offsets
    - Storing 3 billion offsets fits on a server (12GB)

- Searching the array is very fast, but it takes time to construct
  - This time will be amortized over many, many searches
  - Run it once "overnight" and save it away for all future queries

Pos
6
13
8
3
10
15
7
14
2
9
5
12
1
4
11

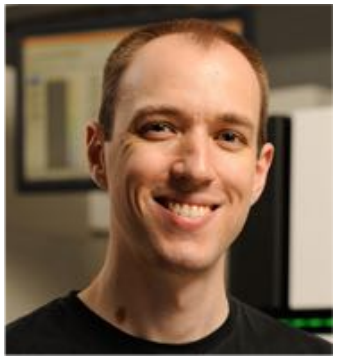
TGATTACAGATTACC

Part 2: Burrows Wheeler Transform

Algorithmic challenge

How can we combine the speed of a suffix array O(m + lg(n)) (or even O(m)) with the size of a brute force analysis (n bytes)?

What would such an index look like?

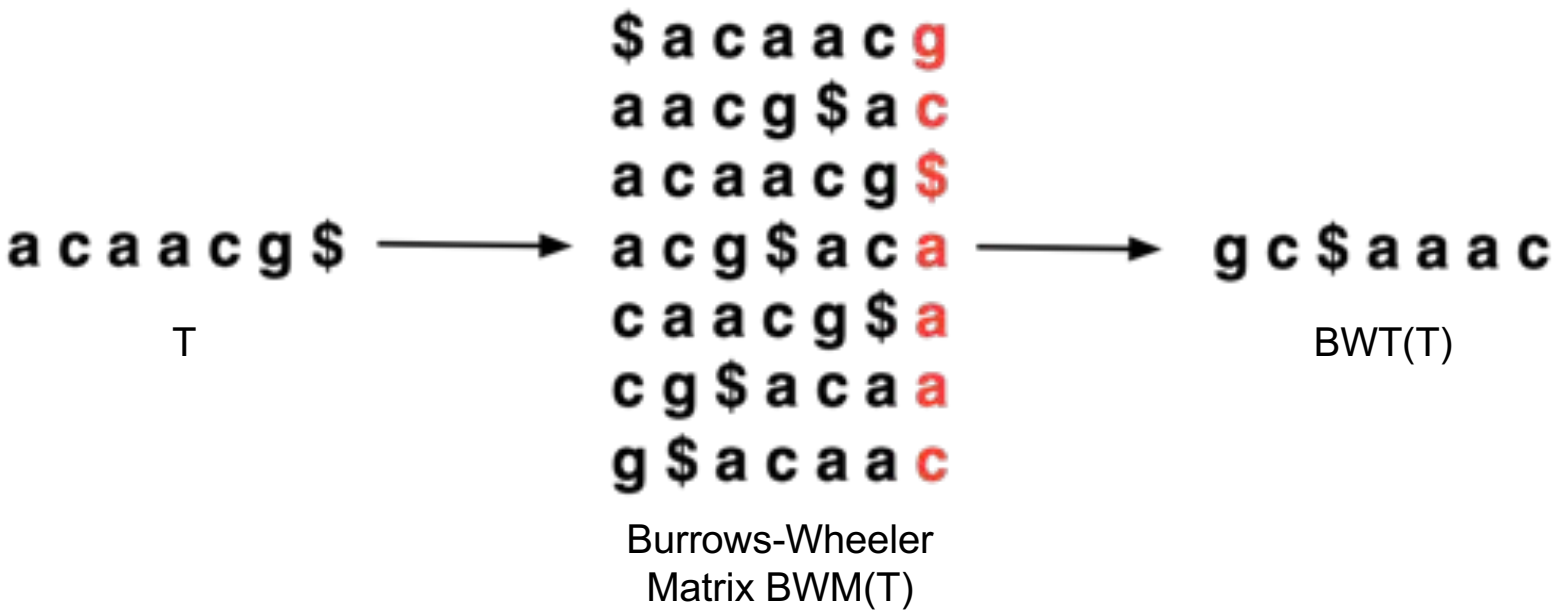


Bowtie: Ultrafast and memory efficient alignment of short DNA sequences to the human genome

Slides Courtesy of Ben Langmead

Burrows-Wheeler Transform

- Permutation of the characters in a text

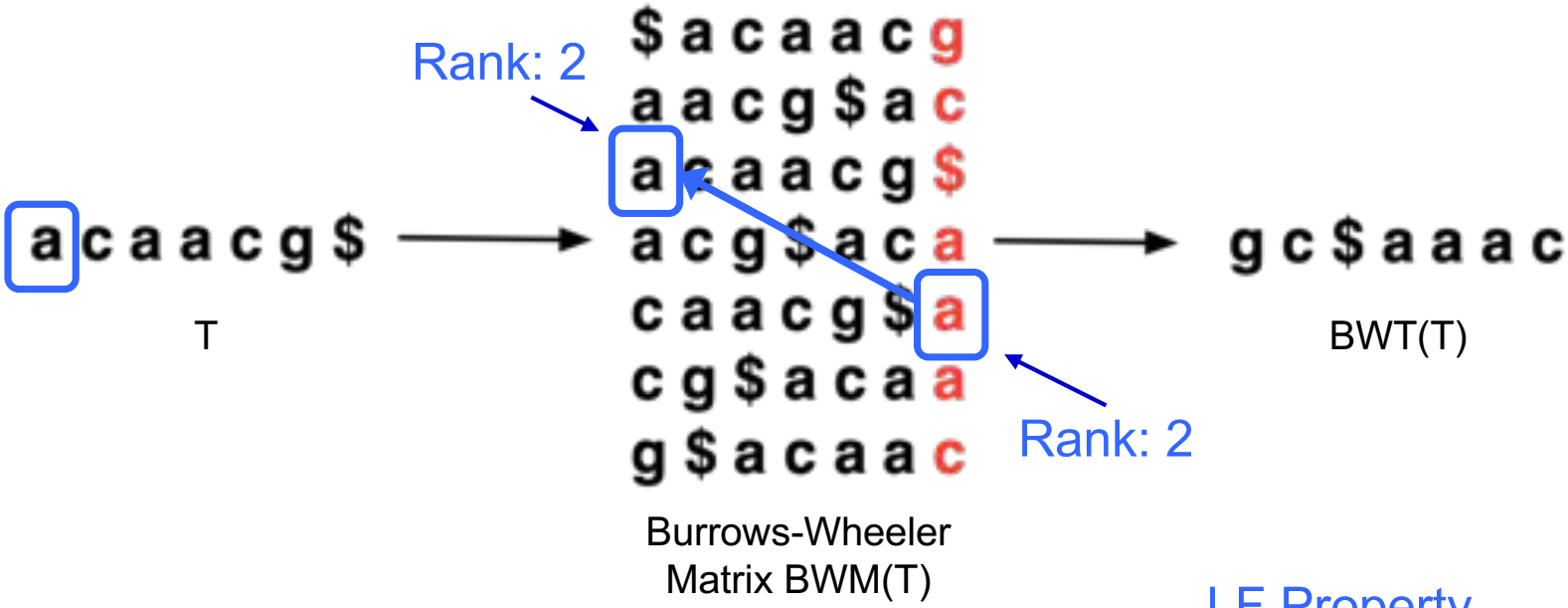


- BWT(T) is the index for T

A block sorting lossless data compression algorithm.
Burrows M, Wheeler DJ (1994) *Digital Equipment Corporation.* Technical Report 124

Burrows-Wheeler Transform

- Reversible permutation of the characters in a text



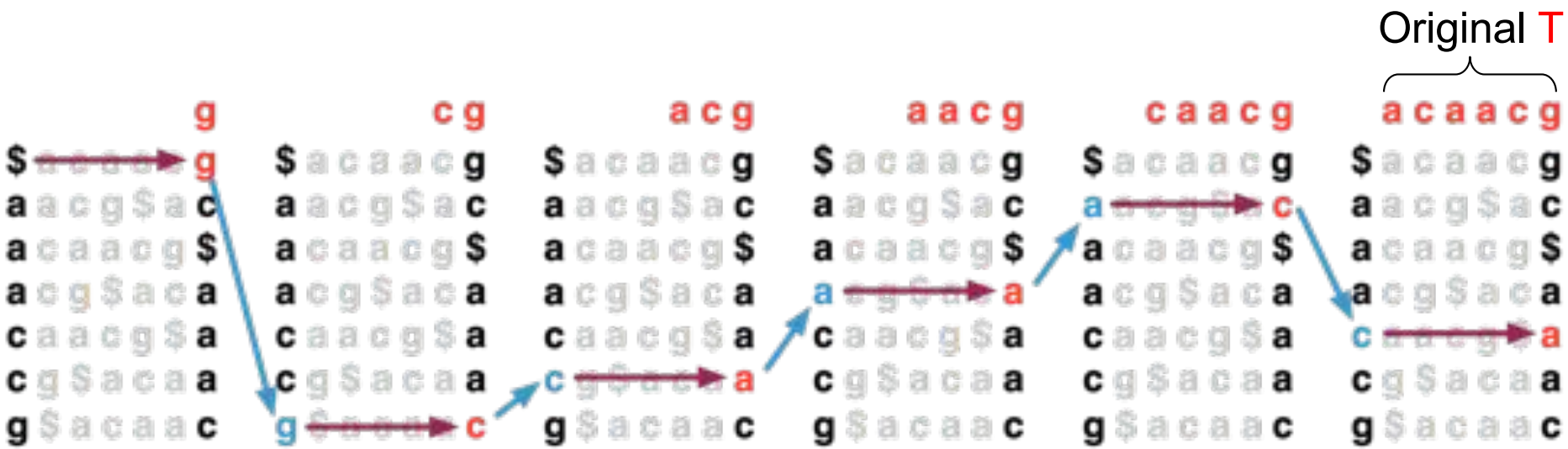
LF Property
implicitly encodes
Suffix Array

- BWT(T) is the index for T

A block sorting lossless data compression algorithm.
Burrows M, Wheeler DJ (1994) *Digital Equipment Corporation.* Technical Report 124

Burrows-Wheeler Transform

- Recreating T from BWT(T)
  - Start in the first row and apply **LF** repeatedly, accumulating predecessors along the way



[Decode this BWT string: ACTGA$TTA]


Welcome to Applied Comparative Genomics
https://github.com/schatzlab/appliedgenomics2018

Questions?