DNA Walk

Jon-Michael Deldin

Dept. of Computer Science
University of Montana
jon-michael.deldin@mso.umt.edu

September 7, 2011

Outline

Topic

Goals

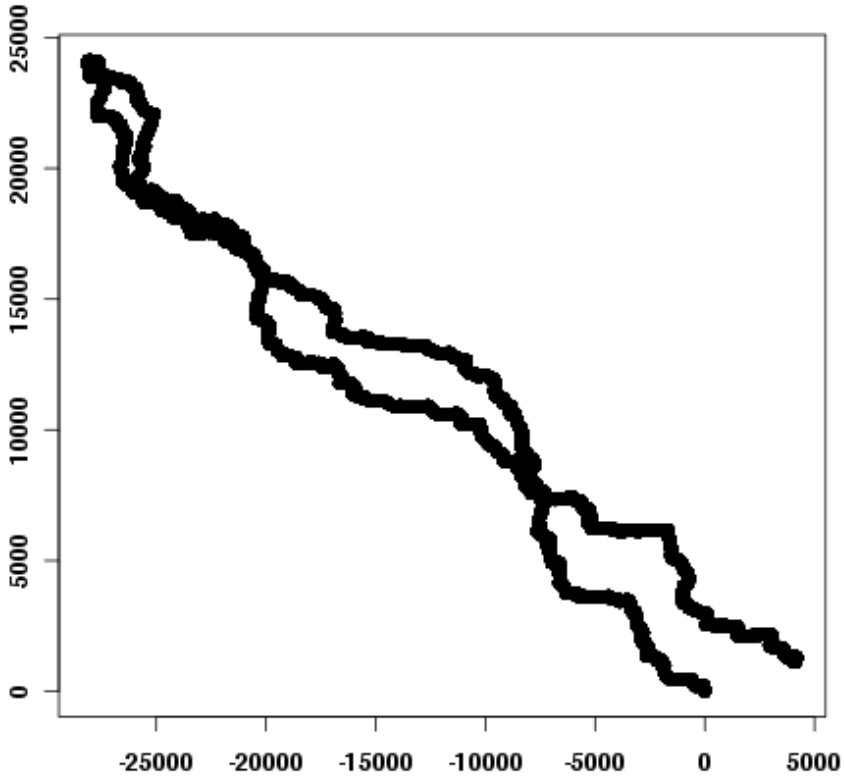
- show a simple visualization technique for compositional bias

Goals

- show a simple visualization technique for compositional bias
- improve Perl skills

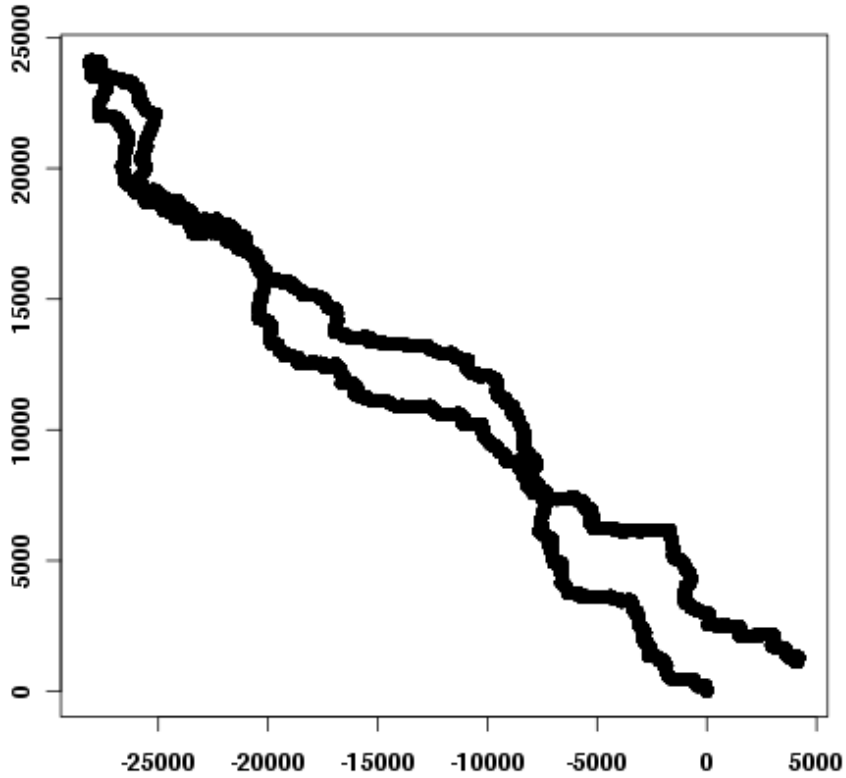
Topic

What is a DNA Walk?



- Read "Genomic landscapes" by Jean R. Lobry for background

What is a DNA Walk?



- Read "Genomic landscapes" by Jean R. Lobry for background
- For every nucleotide, adjust an X or a Y coordinate based on a "compass"

Compass

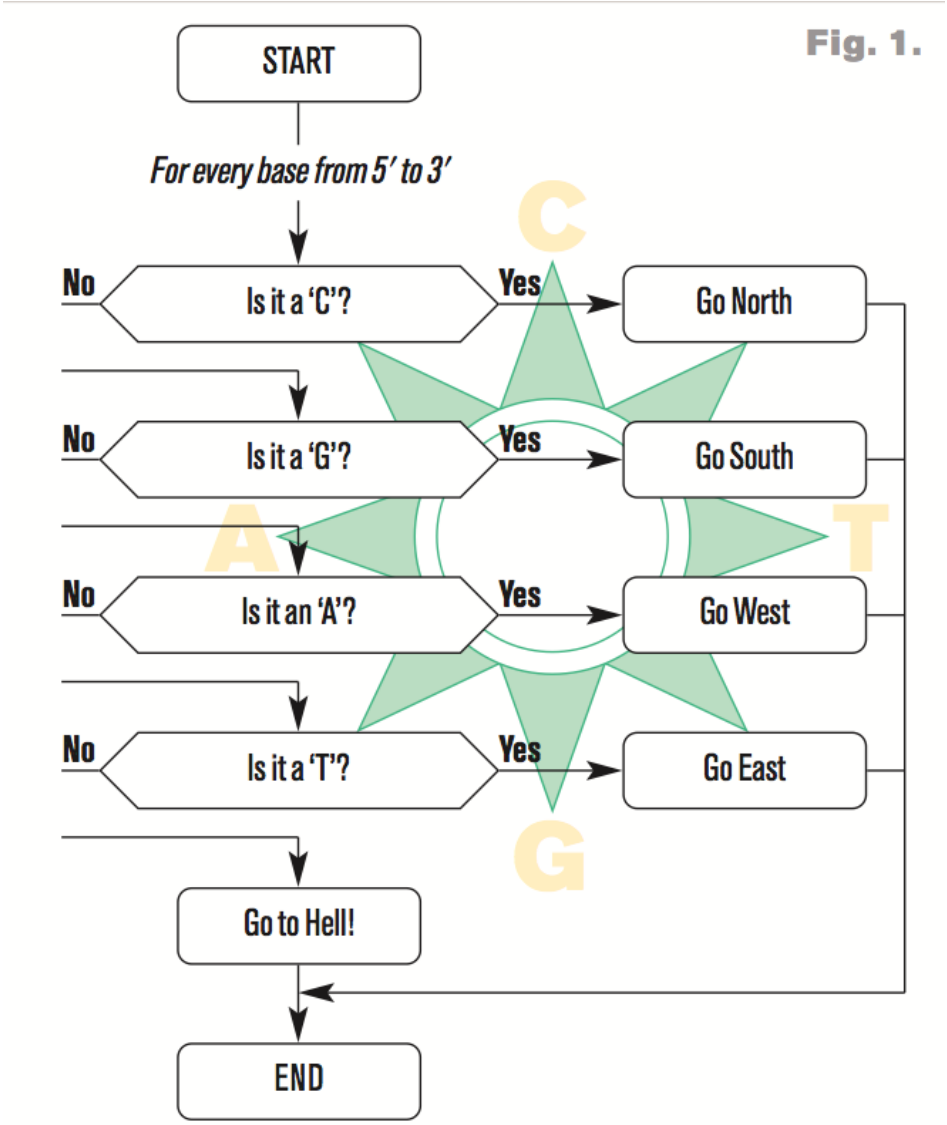


Fig. 1.
START
For every base from 5′ to 3′
No
Is it a 'C'?
Yes
Go North
No
Is it a 'G'?
Yes
Go South
No
Is it an 'A'?
Yes
Go West
No
Is it a 'T'?
Yes
Go East
Go to Hell!
END
C
A
T
G

Topic

Overview

- "walking" *Borrelia burgdorferi's* genome

Big picture

1. Get the sequence file into a format Perl can use

Big picture

1. Get the sequence file into a format Perl can use
2. "Walk" the sequence to determine the XY coordinates

Big picture

1. Get the sequence file into a format Perl can use
2. "Walk" the sequence to determine the XY coordinates
3. Create a CSV file with the coordinates

Big picture

1. Get the sequence file into a format Perl can use
2. "Walk" the sequence to determine the XY coordinates
3. Create a CSV file with the coordinates
4. Plot in R or Excel

Input, Output, Process

There are a few different parts to this project.

1. Preprocessing (getting the sequence file)

Input, Output, Process

There are a few different parts to this project.

1. Preprocessing (getting the sequence file)

   input fasta filename

Input, Output, Process

There are a few different parts to this project.

1. Preprocessing (getting the sequence file)

   input fasta filename

   output array of nucleotides

Input, Output, Process

There are a few different parts to this project.

1. Preprocessing (getting the sequence file)

   input fasta filename

   output array of nucleotides

2. Walking the genome

Input, Output, Process

There are a few different parts to this project.

1. Preprocessing (getting the sequence file)
   - input fasta filename
   - output array of nucleotides
2. Walking the genome
   - input array of nucleotides

Input, Output, Process

There are a few different parts to this project.

1. Preprocessing (getting the sequence file)
   - input: fasta filename
   - output: array of nucleotides
2. Walking the genome
   - input: array of nucleotides
   - output: array of X coordinates, array of Y coordinates

Input, Output, Process

There are a few different parts to this project.

1. Preprocessing (getting the sequence file)
   - input: fasta filename
   - output: array of nucleotides
2. Walking the genome
   - input: array of nucleotides
   - output: array of X coordinates, array of Y coordinates
3. Creating a CSV file with the coordinates

Input, Output, Process

There are a few different parts to this project.

1. Preprocessing (getting the sequence file)
   - input: fasta filename
   - output: array of nucleotides
2. Walking the genome
   - input: array of nucleotides
   - output: array of X coordinates, array of Y coordinates
3. Creating a CSV file with the coordinates
   - input: array of X coordinates, array of Y coordinates

Input, Output, Process

There are a few different parts to this project.

1. Preprocessing (getting the sequence file)
   - input: fasta filename
   - output: array of nucleotides
2. Walking the genome
   - input: array of nucleotides
   - output: array of X coordinates, array of Y coordinates
3. Creating a CSV file with the coordinates
   - input: array of X coordinates, array of Y coordinates
   - output: CSV file where each row = `X, Y`

Input, Output, Process

There are a few different parts to this project.

1. Preprocessing (getting the sequence file)
   - input: fasta filename
   - output: array of nucleotides
2. Walking the genome
   - input: array of nucleotides
   - output: array of X coordinates, array of Y coordinates
3. Creating a CSV file with the coordinates
   - input: array of X coordinates, array of Y coordinates
   - output: CSV file where each row = `X, Y`
4. Plotting in R or Excel (for completeness)

Input, Output, Process

There are a few different parts to this project.

1. Preprocessing (getting the sequence file)
   - input: fasta filename
   - output: array of nucleotides
2. Walking the genome
   - input: array of nucleotides
   - output: array of X coordinates, array of Y coordinates
3. Creating a CSV file with the coordinates
   - input: array of X coordinates, array of Y coordinates
   - output: CSV file where each row = `X, Y`
4. Plotting in R or Excel (for completeness)
   - input: DNA walk CSV

Input, Output, Process

There are a few different parts to this project.

1. Preprocessing (getting the sequence file)
   - input: fasta filename
   - output: array of nucleotides
2. Walking the genome
   - input: array of nucleotides
   - output: array of X coordinates, array of Y coordinates
3. Creating a CSV file with the coordinates
   - input: array of X coordinates, array of Y coordinates
   - output: CSV file where each row = `X, Y`
4. Plotting in R or Excel (for completeness)
   - input: DNA walk CSV
   - output: PNG or PDF of the walk

Steps

1. Get the sequence file into a format Perl can use (**preprocessing**)

Steps

1. Get the sequence file into a format Perl can use (**preprocessing**)
    1. open the sequence file

Steps

1. Get the sequence file into a format Perl can use (**preprocessing**)
    1. open the sequence file
    2. read the sequence into an array

Steps

1. Get the sequence file into a format Perl can use (**preprocessing**)
    1. open the sequence file
    2. read the sequence into an array
    3. close the file

Steps

1. Get the sequence file into a format Perl can use (**preprocessing**)
    1. open the sequence file
    2. read the sequence into an array
    3. close the file
    4. remove the header line and newline characters

Steps

1. Get the sequence file into a format Perl can use (**preprocessing**)
   1. open the sequence file
   2. read the sequence into an array
   3. close the file
   4. remove the header line and newline characters
   5. create an array of nucleotides, e.g., `('A', 'C', 'G', ...)`

Steps

1. Get the sequence file into a format Perl can use (**preprocessing**)
   1. open the sequence file
   2. read the sequence into an array
   3. close the file
   4. remove the header line and newline characters
   5. create an array of nucleotides, e.g., `('A', 'C', 'G', ...)`
2. "Walk" the sequence to determine the XY coordinates

Steps

1. Get the sequence file into a format Perl can use (**preprocessing**)
    1. open the sequence file
    2. read the sequence into an array
    3. close the file
    4. remove the header line and newline characters
    5. create an array of nucleotides, e.g., `('A', 'C', 'G', ...)`
2. "Walk" the sequence to determine the XY coordinates
    1. create `@x` and `@y` arrays to hold your coordinates

Steps

1. Get the sequence file into a format Perl can use (**preprocessing**)
   1. open the sequence file
   2. read the sequence into an array
   3. close the file
   4. remove the header line and newline characters
   5. create an array of nucleotides, e.g., `('A', 'C', 'G', ...)`
2. "Walk" the sequence to determine the XY coordinates
   1. create `@x` and `@y` arrays to hold your coordinates
   2. initialize `$x[0]` and `$y[0]` to 0 (the origin)

Steps

1. Get the sequence file into a format Perl can use (**preprocessing**)
   1. open the sequence file
   2. read the sequence into an array
   3. close the file
   4. remove the header line and newline characters
   5. create an array of nucleotides, e.g., `('A', 'C', 'G', ...)`
2. "Walk" the sequence to determine the XY coordinates
   1. create `@x` and `@y` arrays to hold your coordinates
   2. initialize `$x[0]` and `$y[0]` to 0 (the origin)
   3. for every nucleotide, assign a coordinate based on the compass

Steps

1. Get the sequence file into a format Perl can use (**preprocessing**)
   1. open the sequence file
   2. read the sequence into an array
   3. close the file
   4. remove the header line and newline characters
   5. create an array of nucleotides, e.g., `('A', 'C', 'G', ...)`
2. "Walk" the sequence to determine the XY coordinates
   1. create `@x` and `@y` arrays to hold your coordinates
   2. initialize `$x[0]` and `$y[0]` to 0 (the origin)
   3. for every nucleotide, assign a coordinate based on the compass
3. Create a CSV file with the coordinates

Steps

1. Get the sequence file into a format Perl can use (**preprocessing**)
   1. open the sequence file
   2. read the sequence into an array
   3. close the file
   4. remove the header line and newline characters
   5. create an array of nucleotides, e.g., `('A', 'C', 'G', ...)`
2. "Walk" the sequence to determine the XY coordinates
   1. create `@x` and `@y` arrays to hold your coordinates
   2. initialize `$x[0]` and `$y[0]` to 0 (the origin)
   3. for every nucleotide, assign a coordinate based on the compass
3. Create a CSV file with the coordinates
   - print the coordinates to a CSV

Steps

1. Get the sequence file into a format Perl can use (**preprocessing**)
   1. open the sequence file
   2. read the sequence into an array
   3. close the file
   4. remove the header line and newline characters
   5. create an array of nucleotides, e.g., `('A', 'C', 'G', ...)`
2. "Walk" the sequence to determine the XY coordinates
   1. create `@x` and `@y` arrays to hold your coordinates
   2. initialize `$x[0]` and `$y[0]` to 0 (the origin)
   3. for every nucleotide, assign a coordinate based on the compass
3. Create a CSV file with the coordinates
   - print the coordinates to a CSV
4. Plot in R or Excel