COMP 555 Bioalgorithms Fall 2012
Problem Set 4
Assigned: Nov 8, 2012
Due: Nov 20, 2012

1. Given a long nucleotide sequence of length n, we want to find the number of occurrences of a given l-mer of arbitrary length m.
a) What would be the appropriate data structure - Hash-table or Suffix tree ?
b) Design an efficient algorithm for this problem using the data structure chosen in (a). Note: The running time should be independent of the length of the nucleotide.
c) What is the time complexity for the query search?

2. Following are the scores of students in a class

StudentID	Score
A1	4
A2	4
A3	9
A4	-1
A5	14
A6	19
A7	17
A8	12
A9	7
AA	15

Design a hash table scheme so that you are able find the students with score = x. The size of the hash table should be at most 5.
a) Give a good hash function.
b) Show what would be stored in the hash table
c) Demonstrate how the following query would work:
Find the students with score = 4.

3. Problem 9.7 of book (palindromes of length > 1)

4. Consider the following sets of points in two dimensional space: (2.2, 3.3), (1.6, 4.5), (6.7, 3.3), (4.7, 1.2), (8.6, 7.5), (5.3, 5.6), (7.6, 6.5), (2.5, 4.6), (5.5, 4.5), (1.1, 3.8)

a) Compute distance matrix. Assume Euclidean distance is used.
b) Assume that distance between two clusters is defined as the maximal distance of any pair of their elements. Draw the cluster tree constructed by the hierarchical clustering algorithm (in page 345 of the book)

c) Assume that the distance between two clusters is defined as the average distance over all pairs. Draw the hierarchical cluster tree again. Write your observations on (b) and (c)
d) Run the K-means algorithm with k=2. Use the first two points as the initial cluster representatives.
e) Run the K-means algorithm with k=2. Use the last two points as the initial cluster representatives. Write your observations on (d) and (e)

5. Use the code given in the class notes for computing B-W-Transform. Do the following:
Repeat 500 times:

- Take any substring of size 100 from any chromosome of any organism.
- Compute the B-W-Transform
- Find the runs of characters in original substring (r1) and runs of characters in transformed substring (r2). (A run is defined as number of repeated characters. e.g run('ACTGGGAA')=5). The python function for run computation will be provided.

Plot all the 500 (r1,r2) as a scatter plot.
Submit the following as part of the assignment:
1. Describe where the substrings were taken from
2. Scatter plot
3. What do you observe in the scatter plot ? What conclusions can you make?
(No need to submit any code for this)

6. Programming assignment (Please submit code and executable):
Implement the search using Burrows Wheeler Alignment.
Input: B-W-Transformed String (as file) & Query String
Output: All the positions of the query in the original string. If the string is not found, output -1.

Specific Upload Instructions will be provided in the Announcement section of the course web page.