Mounir Abderrahmani 120201200

March 21th, 2016

ECE 551 Homework 1 (Spring 2016)

1- Consider the following 4 DNA sequences: GCAT, GCCA, GGCAT,GGCA.

Find the optimal global alignment between all pairs with the following scoring function: “match score is 3, mismatch score is -2, a linear gap model with a gap penalty of -2”.

Perform the multiple alignment of these four sequences using the Star Alignment technique. Choose the centre sequence that is most similar to the other sequences.

Optimum global alignment: GCAT, GCCA, GGCAT, GGCA

		G	**C**	**A**	**T**
	0	-2	-4	-6	-8
G	-2	3	-1	-3	-5
C	-4	1	6	4	2
C	-6	-1	4	4	2
A	-8	-3	1	7	5

G_CAT

GCCA_

Score = 5

		G	**C**	**A**	**T**
	0	-2	-4	-6	-8
G	-2	3	1	-1	-3
G	-4	1	1	-1	-3
C	-6	-1	4	2	0
A	-8	-3	2	7	5

G_CAT

GGCA_

Score = 5

		G	**G**	**C**	**A**	**T**
	0	-2	-4	-6	-8	-10
G	-2	3	1	-1	-3	-5
C	-4	1	1	4	2	0
A	-6	-1	-1	2	7	5
T	-8	-3	-3	0	5	10

GGCAT

G_CAT

Score = 10

		G	**C**	**A**	**T**
	0	-2	-4	-6	-8
G	-2	3	1	-1	-3
G	-4	1	1	-1	-3
C	-6	-1	4	2	0
A	-8	-3	2	7	5

G_CAT

GGCA_

Score = 5

Mounir Abderrahmani 120201200

		G	G	C	A	T
	0	-2	-4	-6	-8	-10
G	-2	3	1	-1	-3	-5
G	-4	1	6	4	2	0
C	-6	-1	4	9	7	5
A	-8	-3	2	7	12	10

GGCAT

GGCA_

Score = 10

		G	G	C	A
	0	-2	-4	-6	-8
G	-2	3	1	-1	-3
C	-4	1	1	4	2
C	-6	-1	-1	4	2
A	-8	-3	-3	2	7

GGCA

GCCA

Score = 7

"GGCAT" is the most similar sequence with all the others proven by summing its scores (Optimum global alignment) with the other 3 sequences

GGCAT → 10+10+5=25, GGCA → 5+7+10=22, GCAT → 10+5+5=20, GCCA → 5+7+5=17.

2- Fill in the affine gap penalty matrix of AGC with ATGCC. Match score= 1, mismatch score = -1, gap opening penalty is -4 and gap extension penalty is -1. Give the score of the best alignment and list all the alignments with this score…

		A	T	G	C	C
	M: 0	-∞	-∞	-∞	-∞	-∞
	I_x: -4	-∞	-∞	-∞	-∞	-∞
	I_y: -4	-5	-6	-7	-8	-9
A	-∞	1	-6	-7	-8	-9
	-5	-∞	-∞	-∞	-∞	-∞
	-∞	-∞	-4	-5	-6	-7
G	-∞	-6	0	-3	-6	-7
	-6	-5	-11	-12	-13	-14
	-∞	-∞	-11	-5	-6	-7
C	-∞	-7	-6	-1	-2	-5
	-7	-6	-5	-8	-11	-12
	-∞	-∞	-12	-11	-6	-7

The best alignment Scores are:

ATGCC

A__GC

Score = -4

ATGCC

AG__C

Score = -4

3- Assume that we have the following reads: ABCDEFGC EFGCDHIJ CDEFGCDH.

a) Draw a de Bruijn graph for this data set with k-mer length 3.
Edges should correspond to k-mers and nodes correspond to (k-1)-mers.

Mounir Abderrahmani 120201200

Draw one weighted edge per distinct k-mer with weight equal to the number of times the k-mer.

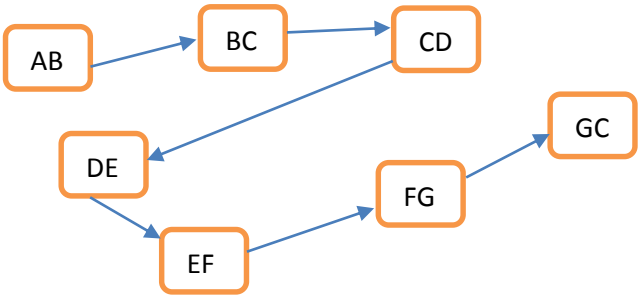
b) Determine whether the graph is Eulerian or not.
c) Give an example of a walk through the graph that traverses three nodes and spells out a 4-mer that does not appear in any of the input reads. (This is an example where de Bruijn graphs may generate sequences that do not appear in reads.)

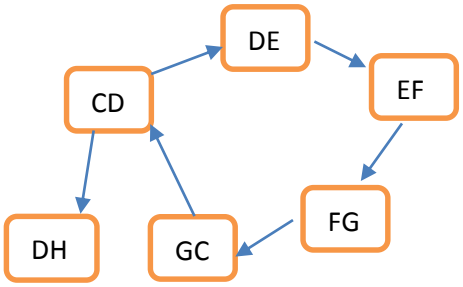
a-

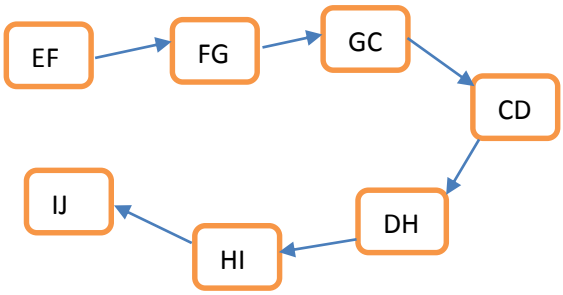
ABCDEFGC
K-mers : ABC BCD CDE DEF EFG FGC
K-1mers:AB BC BC CD CD DE DE EF EF FG FG GC

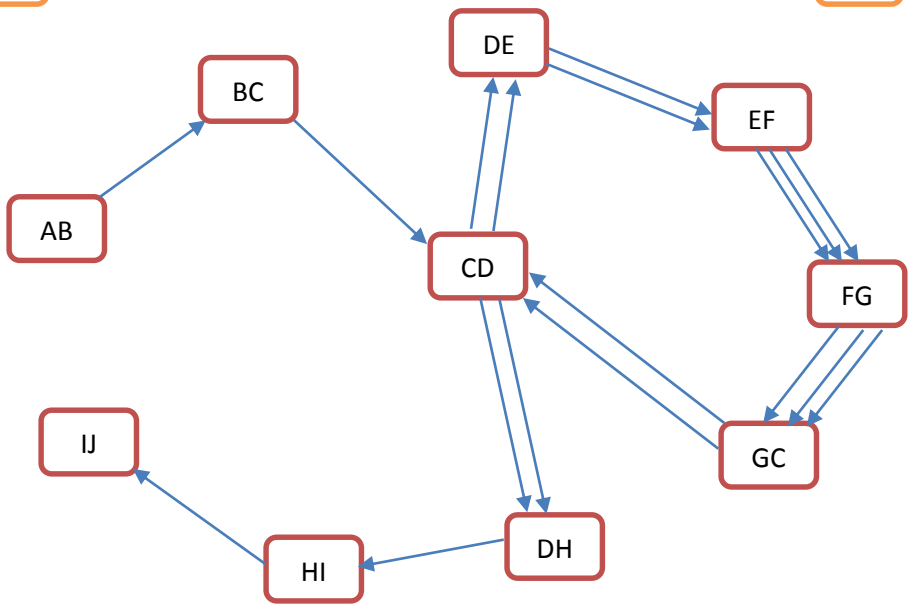
EFGCDHIJ
K-mers : EFG FGC GCD CDH DHI HIJ
K-1-mers :EF FG FG GC GC CD CD DH DH HI HI IJ

CDEFGCDH
K-mers : CDE DEF EFG FGC GCD CDH
K-1-mers : CD DE DE EF EF FG FG GC GC CD CD DH









b- Euler's theorem 2 says if a graph has more than 2 nodes of odd degree(not balanced), then it cannot have an EULER path/walk → the graph is not Eulerian.

c- "BCDH"

The mere "BCDH" doesn't appear in any of the reads above, yet can be constructed from the graph.

Mounir Abderrahmani 120201200

4. Programming

a) Write a program that takes as input a set of reads and uses the greedy fragment assembly algorithm to output a single superstring that contains all reads as substrings. You must use the graph-based (Hamiltonian path) version of the algorithm. We will assume that 1) we are assembling a single-stranded sequence and 2) that no read is a substring any other read. The reads will be read from a _le containing one read per line. To make this algorithm deterministic, you should have a specific rule for tie breaking. For two edges with the same weight choose the edge whose source node read is _rst in lexicographical order. If the source nodes are identical, then we choose the edge whose target node read is _rst in lexicographical order. You can try your algorithm with the reads in test reads:txt to make sure it works correctly (It should output the superstring the quick brown fox jumps over the lazy dog).
b) Use your greedy assemble program to assemble a small subset of the reads (ebola reads:txt) used to assemble the genome of an isolate of the Ebola virus, which caused a major epidemic in West Africa last year . Once correctly assembled, these reads form a short segment of the genome of this virus. To allow your assembler to succeed, the reads have been cleaned of errors and have been oriented so that they all come from the same strand of the genome. Once you have assembled the genomic segment, use the BLASTX web service to search the NCBI database of proteins with your assembled sequence.

4- The Algorithm used for greedy fragment assembly algorithm →
The Result/output of the algorithm on (ebola reads:txt) →

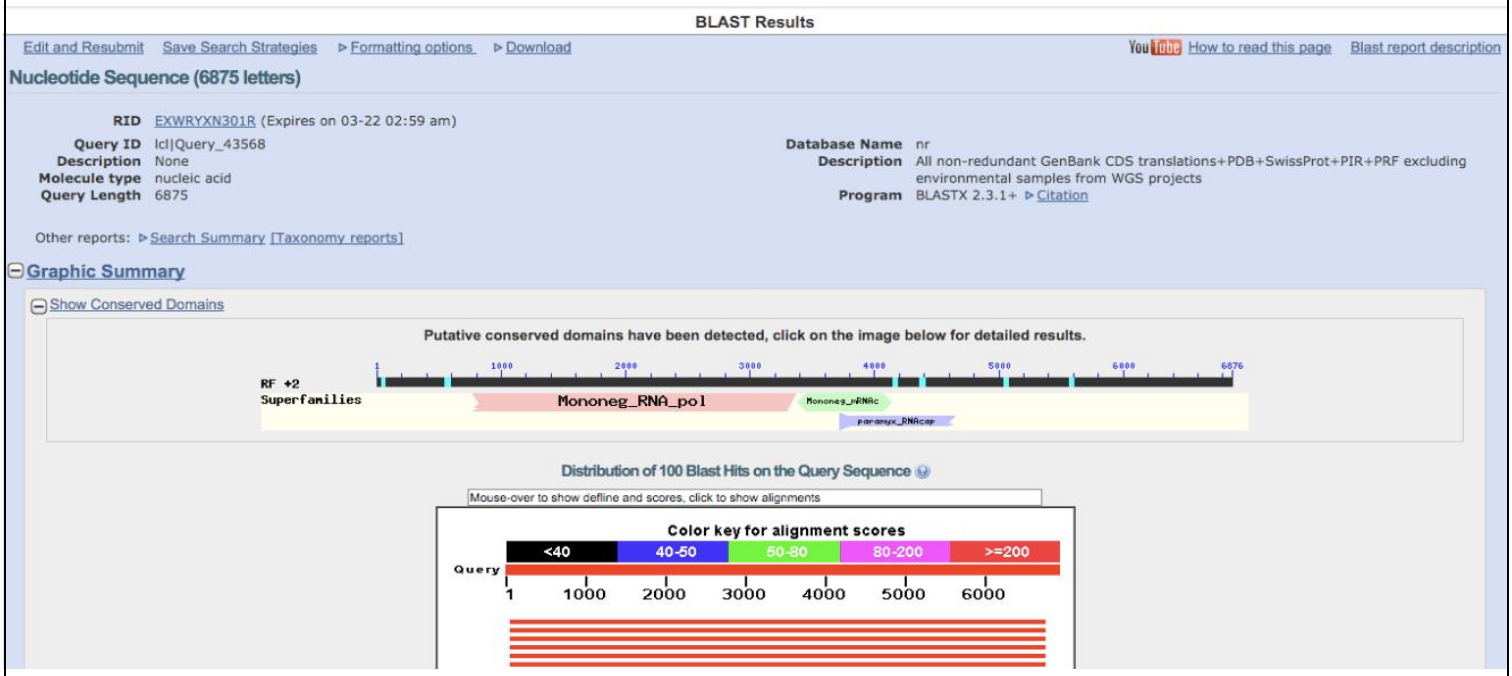
Searching for the protein in BLASTX gives the following pictures snapped from the browser.

polymerase [Zaire ebolavirus] 100% matching

Related Information

Identical Proteins-Identical proteins to AIE11805.1

Sequence ID: gb|AIE11805.1|Length: 2212Number of Matches: 1



Mounir Abderrahmani 120201200

Sequences producing significant alignments:

Select: All None Selected:1

Alignments Download GenPept Graphics

Description	Max score	Total score	Query cover	E value	Ident	Accession
polymerase [Zaire ebolavirus]	4459	4459	96%	0.0	100%	AIE11805.1
polymerase [Zaire ebolavirus]	4459	4459	96%	0.0	99%	ALX31128.1
polymerase [Zaire ebolavirus]	4459	4459	96%	0.0	99%	AKU75578.1
polymerase [Zaire ebolavirus]	4459	4459	96%	0.0	99%	AKG65777.1
polymerase [Zaire ebolavirus]	4459	4459	96%	0.0	99%	AKI83197.1
polymerase [Zaire ebolavirus]	4459	4459	96%	0.0	99%	AKI82693.1
polymerase [Zaire ebolavirus]	4459	4459	96%	0.0	99%	AKG96141.1
polymerase [Zaire ebolavirus]	4459	4459	96%	0.0	99%	AKC36541.1
polymerase [Zaire ebolavirus]	4459	4459	96%	0.0	99%	AKC36001.1

Download GenPept Graphics ▼ Next ▲ Previous ▲ Descriptions

polymerase [Zaire ebolavirus]

Sequence ID: gb|AIE11805.1| Length: 2212 Number of Matches: 1

See 662 more title(s)

Range 1: 1 to 2212 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
4459 bits(11566)	0.0	Compositional matrix adjust.	2212/2212(100%)	2212/2212(100%)	0/2212(0%)	+2

```
Query  74    MATQHTQYPDARLSSPIVLDQCDLVTRACGLYSSYSLNPQLRNCKLPKHIYRLKYDVTVT  253
             MATQHTQYPDARLSSPIVLDQCDLVTRACGLYSSYSLNPQLRNCKLPKHIYRLKYDVTVT
Sbjct  1     MATQHTQYPDARLSSPIVLDQCDLVTRACGLYSSYSLNPQLRNCKLPKHIYRLKYDVTVT  60

Query  254   KFLSDVPVATLPIDFIVPILLKALSGNGFCPVEPRCQQFLDEIIKYTMQDALFLKYYLKN  433
             KFLSDVPVATLPIDFIVPILLKALSGNGFCPVEPRCQQFLDEIIKYTMQDALFLKYYLKN
Sbjct  61    KFLSDVPVATLPIDFIVPILLKALSGNGFCPVEPRCQQFLDEIIKYTMQDALFLKYYLKN  120

Query  434   VGAQEDCVDDHFQEKILSSIQGNEFLHQMFFWYDLAIltrrgrlnrgnsrstWFVHDDLI  613
             VGAQEDCVDDHFQEKILSSIQGNEFLHQMFFWYDLAILTRRGRLNRGNSRSTWFVHDDLI
Sbjct  121   VGAQEDCVDDHFQEKILSSIQGNEFLHQMFFWYDLAILTRRGRLNRGNSRSTWFVHDDLI  180

Query  614   DILGYGDYVFWKIPISLLPLNTQGIPHAAMDWYQTSVFKEAVQGHTHIVSVSTADVLIMC  793
```

Related Information

Identical Proteins - Identical proteins to AIE11805.1