Report

Name: Ying Xu
E-mail: xuying1991@hotmail.com
Last four digits of student ID: 9859

Question1:

i. **Sanger Sequencing**

Sanger sequencing is a method of DNA sequencing developed by Frederick Sanger and colleagues in 1977. It's a chain-termination method that requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleosidetriphosphates (dNTPs), and modified di-deoxynucleotidetriphosphates (ddNTPs) to terminate DNA strand elongation. These chain-terminating nucleotides lack a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, causing DNA polymerase to cease extension of DNA when a modified ddNTP is incorporated. This is frequently performed using a denaturing polyacrylamide-urea gel with each of the four reactions run in one of four individual lanes (lanes A, T, G, C). The DNA bands may then be visualized by autoradiography or UV light and the DNA sequence can be directly read off the X-ray film or gel image.

ii. **Next Generation Sequencing (NGS)**

Next-generation sequencing (NGS), also known as high-throughput sequencing, is the catch-all term used to describe a number of different modern sequencing technologies including: Illumina (Solexa) sequencing, Roche 454 sequencing, on torrent: Proton / PGM sequencing, SOLiD sequencing. It is a massively parallel sequencing technology which can sequence DNA and RNA more quickly than is imaginable with Sanger sequencing.

iii. **Shotgun Sequencing**

Shotgun sequencing is a method used for sequencing long DNA strands. In shotgun sequencing, DNA is broken up randomly into numerous small segments, which are sequenced using the chain termination method to obtain reads. Multiple overlapping reads for the target DNA are obtained by performing several rounds of this fragmentation and sequencing. Computer programs then use the overlapping ends of different reads to assemble them into a continuous sequence.

iv. **Whole Genome Sequencing (WGS)**

Whole genome sequencing is the process of determining the complete DNA sequence of an organism's genome at a single time. This entails sequencing all of an organism's chromosomal DNA as well as DNA contained in the mitochondria and, for plants, in the chloroplast. Whole genome sequencing is ideal for discovery applications, such as identifying causative variants and novel genome assembly. Whole-genome sequencing can detect single nucleotide variants, insertions/deletions, copy number changes, and large structural variants.

v. **Pyrosequencing**

Pyrosequencing is a method of DNA sequencing based on the "sequencing by synthesis" principle, in which the sequencing is performed by detecting the nucleotide incorporated by a DNA polymerase. A mixture of three enzymes (DNA polymerase, ATP sulfurylase and firefly luciferase) and a nucleotide (dNTP) are added to single stranded DNA to be sequenced and the incorporation of nucleotide is

followed by measuring the light emitted. The intensity of the light determines if 0, 1 or more nucleotides have been incorporated, thus showing how many complementary nucleotides are present on the template strand. The nucleotide mixture is removed before the next nucleotide mixture is added. This process is repeated with each of the four nucleotides until the DNA sequence of the single stranded template is determined.

vi. Nanopore Sequencing

Nanopore sequencing is a third-generation approach used in the sequencing of biopolymers- specifically, polynucleotides in the form of DNA or RNA. Using nanopore sequencing, a single molecule of DNA or RNA can be sequenced without the need for PCR amplification or chemical labeling of the sample. It uses electrophoresis to transport an unknown sample through an orifice of 10–9 meters in diameter. A nanopore system always contains an electrolytic solution- when a constant electric field is applied, an electric current can be observed in the system. The magnitude of the electric current density across a nanopore surface depends on the nanopore's dimensions and the composition of DNA or RNA that is occupying the nanopore. Sequencing is made possible because, when close enough to nanopores, samples cause characteristic changes in electric current density across nanopore surfaces. Nanopore sequencing has the potential to offer relatively low-cost genotyping, high mobility for testing, and rapid processing of samples with the ability to display results in real-time.

Question2:

a) The best alignment I got is below, and best alignment score is 12

```
S1: G-ATCG-GAATC
S2: GT-TCGC--A-C
```

Please also refer to excel file hw2item2 for detailed calculation and trace back.
(Below is a screenshot of alignment in that excel)

	S1		G		A		T		C		G		G		A		A		T		C	
S2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	2	0	-1	0	-1	0	-1	0	2	0	2	0	-1	0	-1	0	-1	0	-1	0
	0	0	0	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
T	0	0	-1	2	1	2	4	2	3	2	1	2	1	2	1	2	1	2	4	2	3	2
	0	0	0	2	2	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
T	0	0	-1	2	1	2	4	4	5	4	3	4	3	4	3	4	3	4	6	4	5	4
	0	0	0	2	2	2	2	4	4	5	5	5	5	5	5	5	5	5	5	6	6	6
C	0	0	-1	2	1	2	3	4	6	5	4	5	4	5	4	5	4	5	6	6	8	6
	0	0	0	2	2	2	2	4	4	6	6	6	6	6	6	6	6	6	6	6	6	8
G	0	0	2	2	1	2	1	4	3	6	8	6	8	6	5	6	5	6	5	6	5	8
	0	0	0	2	2	2	2	4	4	6	6	8	8	8	8	8	8	8	8	8	8	8
C	0	0	-1	2	1	2	3	4	6	6	5	8	7	8	7	8	7	8	9	8	10	8
	0	0	0	2	2	2	2	4	4	6	6	8	8	8	8	8	8	8	8	9	9	10
A	0	0	-1	2	4	2	1	4	3	6	5	8	7	8	10	8	10	8	7	9	8	10
	0	0	0	2	2	4	4	4	4	6	6	8	8	8	8	10	10	10	10	10	10	10
C	0	0	-1	2	1	4	5	4	6	6	5	8	7	8	7	10	9	10	11	10	12	10
	0	0	0	2	2	4	4	5	5	6	6	8	8	8	8	10	10	10	10	11	11	12

b) The best alignment I found is below, the best score is 12

```
S1: GA-TCG-G-AATC
S2: G-TTCGC--A--C
    2002220002002=12
```

c) The scores match, both 12.
Alignment logic: in order to get highest score, align every match possible (+2), if cannot be aligned then try aligning C with T (+1), if still cannot align C and T together then insert gap (+0), avoid other mismatch (-1). Following this rule to get highest score.

d) If 6th character G in sequence 1 needs to align with a gap, then following path highlighted in orange should be the only path.

	S1		G		A		T		C		G		G		A		A		T		C	
S2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	2	0	-1	0	-1	0	-1	0	2	0	2	0	-1	0	-1	0	-1	0	-1	0
	0	0	0	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
T	0	0	-1	2	1	2	4	2	3	2	1	2	1	2	1	2	1	2	4	2	3	2
	0	0	0	2	2	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
T	0	0	-1	2	1	2	4	4	5	4	3	4	3	4	3	4	3	4	6	4	5	4
	0	0	0	2	2	2	2	4	4	5	5	5	5	5	5	5	5	5	5	6	6	6
C	0	0	-1	2	1	2	3	4	6	5	4	5	4	5	4	5	4	5	6	6	8	6
	0	0	0	2	2	2	2	4	4	6	6	6	6	6	6	6	6	6	6	6	6	8
G	0	0	2	2	1	2	1	4	3	6	8	6	8	6	5	6	5	6	5	6	5	8
	0	0	0	2	2	2	2	4	4	6	6	8	8	8	8	8	8	8	8	8	8	8
C	0	0	-1	2	1	2	3	4	6	6	5	8	7	8	7	8	7	8	9	8	10	8
	0	0	0	2	2	2	2	4	4	6	6	8	8	8	8	8	8	8	8	9	9	10
A	0	0	-1	2	4	2	1	4	3	6	5	8	7	8	10	8	10	8	7	9	8	10
	0	0	0	2	2	4	4	4	4	6	6	8	8	8	8	10	10	10	10	10	10	10
C	0	0	-1	2	1	4	5	4	6	6	5	8	7	8	7	10	9	10	11	10	12	10
	0	0	0	2	2	4	4	5	5	6	6	8	8	8	8	10	10	10	10	11	11	12

Question3:
Please refer to R script hw2item3.R and comment in it.
Trace back criteria: always go diagonal first, if there is a tie between horizontal and vertical, insert gap in shorter sequence.
My alignment score is 17, below shows result for sequence alignment, seq1 is human HEXA gene, seq2 is mouse HEXA gene:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]
alignment_seq1	"a"	"c"	"g"	"t"	"g"	"a"	"t"	"t"	"c"	"g"	"c"	"c"	"g"	"a"	"t"	"a"	"a"	"g"	"-"
alignment_seq2	"g"	"c"	"-"	"t"	"g"	"-"	"c"	"t"	"g"	"g"	"a"	"a"	"g"	"g"	"g"	"g"	"a"	"g"	"c"

	[,20]	[,21]	[,22]	[,23]	[,24]	[,25]	[,26]	[,27]	[,28]	[,29]	[,30]	[,31]	[,32]	[,33]	[,34]	[,35]	[,36]	[,37]
alignment_seq1	"t"	"c"	"a"	"c"	"g"	"g"	"g"	"g"	"g"	"c"	"g"	"c"	"c"	"g"	"c"	"t"	"c"	"a"
alignment_seq2	"t"	"g"	"g"	"c"	"c"	"g"	"g"	"t"	"g"	"g"	"g"	"c"	"c"	"-"	"a"	"t"	"g"	"g"

	[,38]	[,39]	[,40]	[,41]	[,42]	[,43]	[,44]	[,45]	[,46]	[,47]	[,48]	[,49]	[,50]	[,51]	[,52]	[,53]	[,54]	[,55]
alignment_seq1	"c"	"c"	"t"	"g"	"-"	"a"	"c"	"c"	"a"	"g"	"g"	"g"	"t"	"c"	"t"	"c"	"a"	"c"
alignment_seq2	"c"	"c"	"g"	"g"	"c"	"t"	"g"	"c"	"a"	"g"	"g"	"c"	"t"	"c"	"t"	"-"	"g"	"g"

	[,56]	[,57]	[,58]	[,59]	[,60]	[,61]	[,62]	[,63]	[,64]	[,65]	[,66]	[,67]	[,68]	[,69]	[,70]	[,71]	[,72]	[,73]
alignment_seq1	"g"	"t"	"g"	"g"	"c"	"-"	"c"	"a"	"g"	"c"	"c"	"c"	"c"	"c"	"t"	"c"	"c"	"g"
alignment_seq2	"g"	"t"	"t"	"t"	"c"	"g"	"c"	"t"	"g"	"c"	"t"	"g"	"c"	"t"	"g"	"g"	"c"	"g"

	[,74]	[,75]	[,76]	[,77]	[,78]	[,79]	[,80]	[,81]	[,82]	[,83]	[,84]	[,85]	[,86]	[,87]	[,88]	[,89]	[,90]	[,91]
alignment_seq1	"a"	"g"	"a"	"g"	"g"	"-"	"g"	"g"	"a"	"g"	"a"	"c"	"c"	"a"	"g"	"c"	"-"	"g"
alignment_seq2	"-"	"g"	"c"	"g"	"g"	"c"	"g"	"t"	"t"	"g"	"g"	"c"	"t"	"t"	"g"	"c"	"t"	"t"

	[,92]	[,93]	[,94]	[,95]	[,96]	[,97]	[,98]	[,99]	[,100]	[,101]	[,102]	[,103]	[,104]	[,105]
alignment_seq1	"g"	"g"	"c"	"c"	"a"	"t"	"g"	"a"	"c"	"a"	"a"	"g"	"c"	"t"
alignment_seq2	"g"	"g"	"c"	"c"	"a"	"c"	"g"	"g"	"c"	"a"	"c"	"t"	"g"	"t"

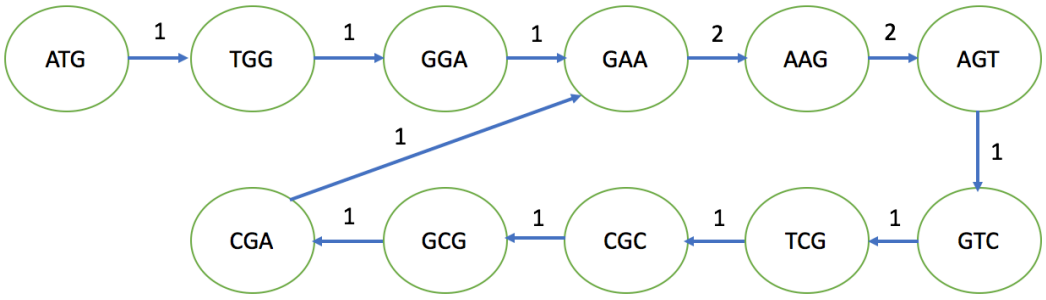
Question4:

a) ATGGAAGTCGCGAAGT

Convert to 3-mers:

ATG
TGG
GGA
GAA
AAG
AGT
GTC
TCG
CGC
GCG
CGA
GAA
AAG
AGT

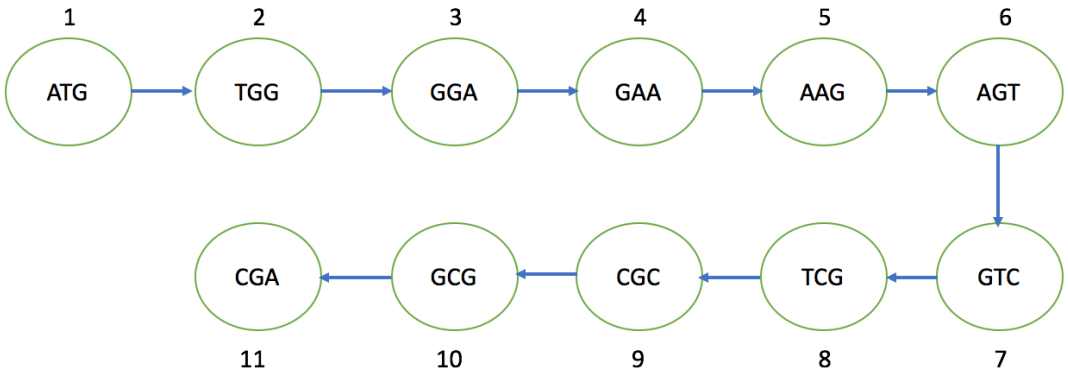
De Bruijn graph without redundancy:



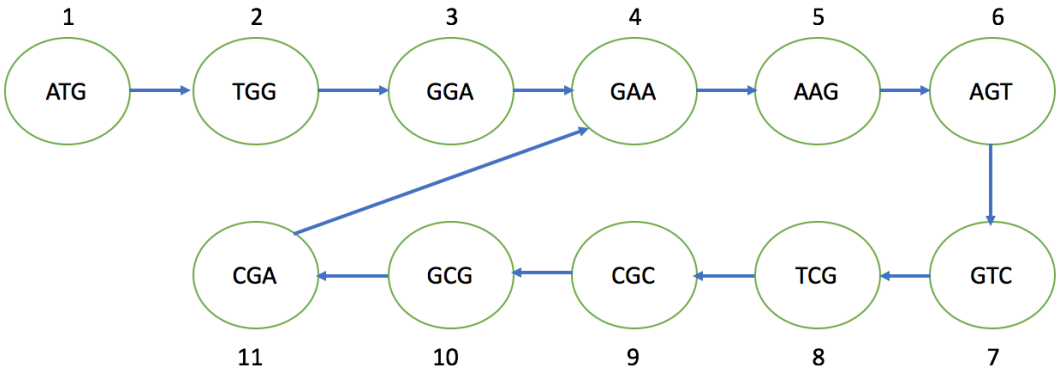
b) Hamilton Path:

Cannot traverse any vertex more than once, can travel each edge multiple times.

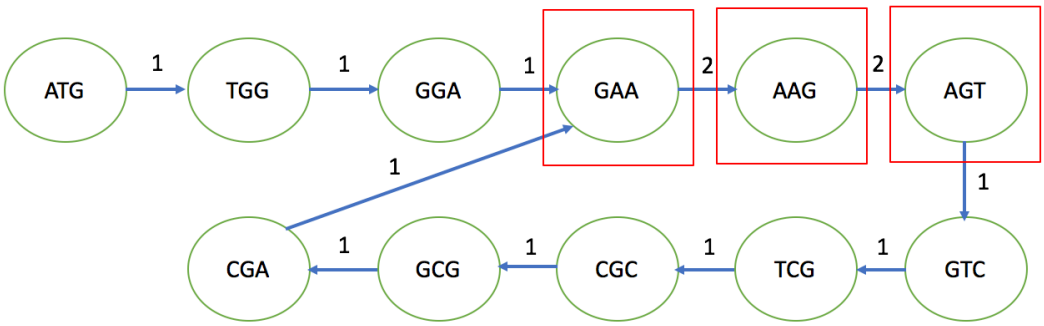
1->2->3->4->5->6->7->8->9->10->11



Euler Path:
All edges must be traversed, can travel each vertex multiple times, cannot traverse any edge more than once.
1->2->3->4->5->6->7->8->9->10->11->4



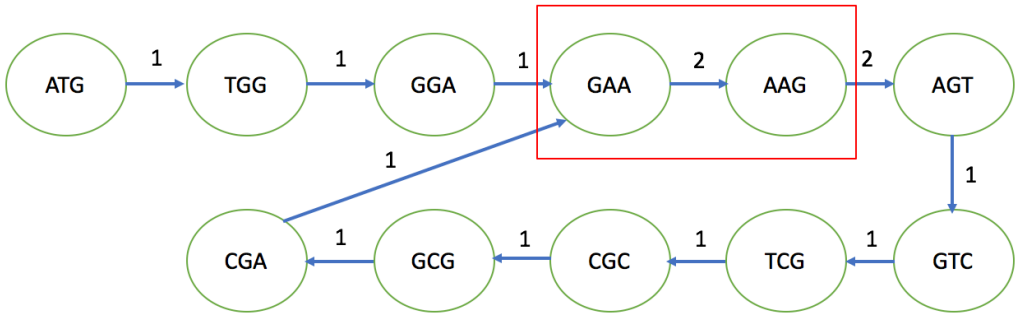
c) Repeat elements of length 3:



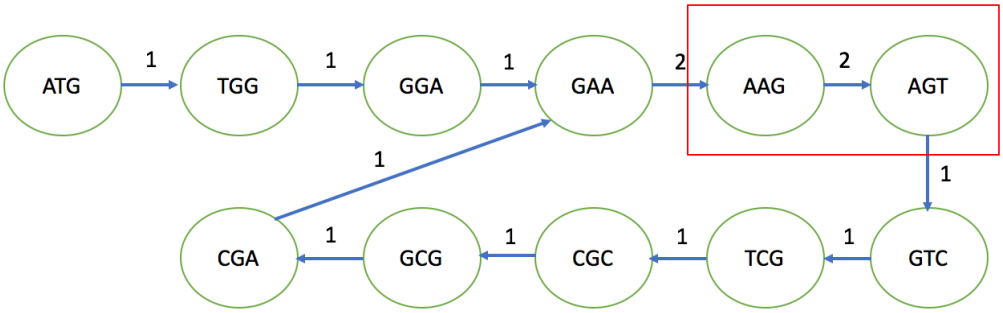
GAA distance=8
ATGGAAGTCGCGAAGT
12345678
AAG distance=8
ATGGAAGTCGCGAAGT
12345678
AGT distance=8
ATGGAAGTCGCGAAGT
12345678

d) There are 2 repeat elements length=4:

GAAG shown in below graph:



AAGT shown in below graph:



1 repeat element length=5:
GAAGT shown in below graph: