HW2 Alex zhang

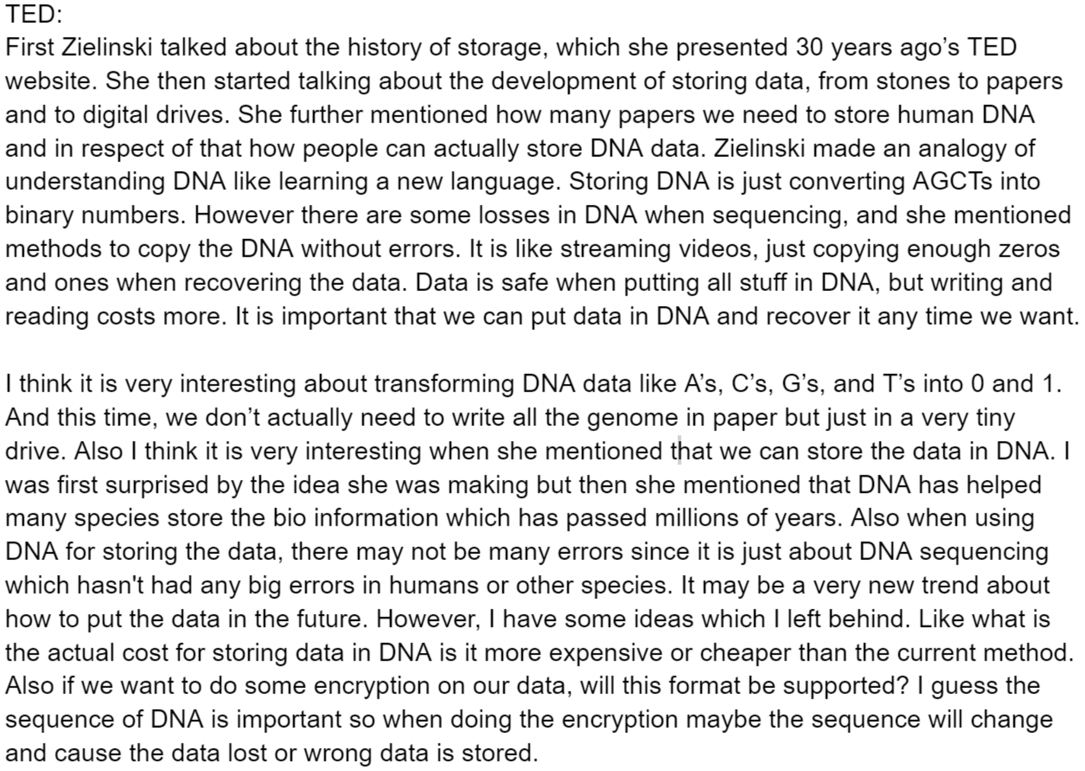
problem one:

c) I think the distribution is not really costant, the largest goes to 0.63, and lowest goes to 0.5.

d) The GC content is greater than 50%. Since this tile is human sequence, which means it is not very possible to have signiticant mutations, I think isthat GC bond is more stable compare to AT one. This is probabaly the reason we have more GC.

f) Yes, since based on graph, G C content are just slighthy more than AT content.

problem two:



TED:

First Zielinski talked about the history of storage, which she presented 30 years ago's TED website. She then started talking about the development of storing data, from stones to papers and to digital drives. She further mentioned how many papers we need to store human DNA and in respect of that how people can actually store DNA data. Zielinski made an analogy of understanding DNA like learning a new language. Storing DNA is just converting AGCTs into binary numbers. However there are some losses in DNA when sequencing, and she mentioned methods to copy the DNA without errors. It is like streaming videos, just copying enough zeros and ones when recovering the data. Data is safe when putting all stuff in DNA, but writing and reading costs more. It is important that we can put data in DNA and recover it any time we want.

I think it is very interesting about transforming DNA data like A's, C's, G's, and T's into 0 and 1. And this time, we don't actually need to write all the genome in paper but just in a very tiny drive. Also I think it is very interesting when she mentioned that we can store the data in DNA. I was first surprised by the idea she was making but then she mentioned that DNA has helped many species store the bio information which has passed millions of years. Also when using DNA for storing the data, there may not be many errors since it is just about DNA sequencing which hasn't had any big errors in humans or other species. It may be a very new trend about how to put the data in the future. However, I have some ideas which I left behind. Like what is the actual cost for storing data in DNA is it more expensive or cheaper than the current method. Also if we want to do some encryption on our data, will this format be supported? I guess the sequence of DNA is important so when doing the encryption maybe the sequence will change and cause the data lost or wrong data is stored.

problem three:

a) for the outer loop, it will be $y-x+1$ number of times.

b)

i) The maximum of comparison should be $\lfloor y/x \rfloor (x-1)+y$

ii) I think this happens when t only consists of pattern. like $t=$ "catcatcat.."
$p=$ "cat"

c)

i) The minimum number of comparisons should be $y-x+1$,

ii) it will happen when the fist letter in pattern doesn't match any letters in text, so you have to compare to the end.

d)

Based on the previous formula, the maximum should be $44+\lfloor 44/4 \rfloor \cdot (4-1) = 44+11 \cdot 3 = 77$.

The minimum should be $44-4+1=41$,

So it is closed to the minimum value.

problem Five:

c) No, I don't expect that. One reason is that Some bases are incorrectly read by the sequencer. Another reason is that there might be some mutation which cause the sequence differ from reference genome sequence

d) There are 459/1000 exact matches.

e) There are 932/1000 exact matches.

Problem Six:

	BCR	GSR	Choice
one	GTTATAG	GTTATAG2T	9 shifts
two	G	GATCGCGG C	9 shifts
three	0	0	0