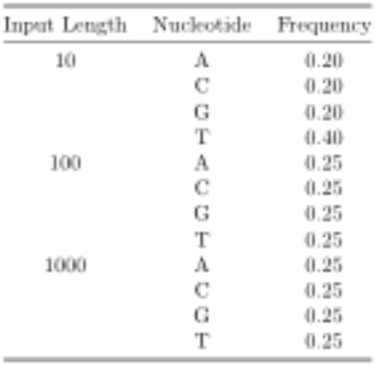
I529: Homework 1

Will Shoemaker

Section 1

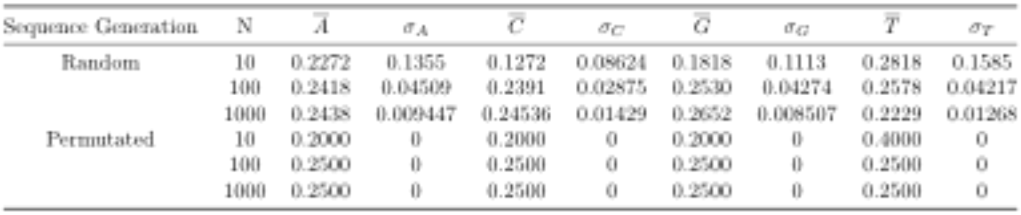
Three DNA sequences with the same nucleotide composition were genreated with nucleotide lengths of 10, 100, and 1,000 (Table1). Each sequence was run through the programs RANSEQ1 and RANSEQ2. RANSEQ1 generates random sequences with the per-base probability of a nucleotide being chosen obtained by the input DNA sequence. RANSEQ2 permutates the input sequence, generating a rearranged sequence. Both programs output their results to a user-chosen file in FASTA format. The results of both programs suggest that a permutation-based approach is innapropriate if one is interested in retaining the element of stochasticity present in the random sequence generation method. However, if one wants a deterministic approach with regards to nucleotide frequency, then a permutation-based method may be useful.

While not taken into account here, it is also important to factor in the length of the input sequence. Storing a large genome in active memory and rearranging it may require a lot of time and computational power.



Input Length	Nucleotide	Frequency
10	A	0.20
	C	0.20
	G	0.20
	T	0.40
100	A	0.25
	C	0.25
	G	0.25
	T	0.25
1000	A	0.25
	C	0.25
	G	0.25
	T	0.25

Table 1: The nucleotide frequencies of the DNA sequence used as input in the programs RANSEQ1 and RANSEQ2.



Sequence Generation	N	$\overline{A}$	σ_A	$\overline{C}$	σ_C	$\overline{G}$	σ_G	$\overline{T}$	σ_T
Random	10	0.2272	0.1355	0.1272	0.08624	0.1818	0.1113	0.2818	0.1585
	100	0.2418	0.04509	0.2391	0.02875	0.2530	0.04274	0.2578	0.04217
	1000	0.2438	0.009447	0.24536	0.01429	0.2652	0.008507	0.2229	0.01268
Permutated	10	0.2000	0	0.2000	0	0.2000	0	0.4000	0
	100	0.2500	0	0.2500	0	0.2500	0	0.2500	0
	1000	0.2500	0	0.2500	0	0.2500	0	0.2500	0

Table 2: The mean and

standard deviation of each nucleotide for random and permutated generated sequences. Each sequence generation method was run with three different input sequence lengths (N).

Section 2

1

The probability of having the genetic disease is $P(disease) = 1 * 10^{-7}$. The test is 100% sensitive and 99.99% specific, so $P(positive \mid disease) = 1$ and $P(negative \mid no\ disease) = 0.9999$, respectively.

From here we can find the probability that someone who has gotten a positive result on the test has the disease.

$$P(disease \mid positive) = \frac{P(positive \mid disease) * P(disease)}{P(positive)}$$

Then calculate the probability of a positive result using the law of total probability:

$$P(A) = \sum_n P(A \ \cap \ B_n) = \sum_n P(A \mid B_n)P(B_n)$$

$$P(positive) = (P(positive \mid disease) * P(disease)) + (P(positive \mid no\ disease) * P(no\ disease))$$

$$= (1 * (1 * 10^{-7})) + (0.9999999 * 0.0001)$$

$$= 0.0001$$

Then get the conditional probability

$$P(disease \mid positive) = \frac{1.0 * (1 * 10^{-7})}{0.0001}$$

$$= 0.001$$

There is a very low probability of having the disease if the test is positive. Excluding the possibility of multiple tests or repeating the test, I would not want to take this test, as it conveys little information regarding whether or not I'd have the disease.

However, I could take the test multiple times. Using the law of total probability, we can see what the probability of having the disease is given two positive test results.

$$P(disease \mid positive_1, positive_2) = \frac{P(positive_1 \mid disease) * P(positive_2 \mid disease) * P(disease)}{P(positive)}$$

Get the total probality of a positive result.

$$P(positive) = (P(positive \mid disease)^2 * P(disease)) + (P(positive \mid no\ disease)^2 * P(no\ disease))$$

$$= (1^2)(1 * 10^{-7}) + (0.9999999 * (0.0001)^2)$$

$$\approx 1.10 * 10^{-7}$$

Which gives us

$$P(disease \mid positive_1, positive_2) = \frac{1 * 1 * (1 * 10^{-7})}{1.10 * 10^{-7}}$$

$$\approx 0.909$$

At which point I would begin to trust the test

2

First, we construct a transition matrix. The rows represent time points t_{-1} and t. The columns represent the probability of states t and t_{+1}

	CC	**CR**	**CS**	**RC**	**RR**	**RS**	**SC**	**SR**	**SS**
CC	0.6	0.2	0.2	0	0	0	0	0	0
CR	0	0	0	0.2	0.6	0.2	0	0	0
CS	0	0	0	0	0	0	0.2	0.2	0.6
RC	0.6	0.2	0.2	0	0	0	0	0	0
RR	0	0	0	0.3	0.5	0.2	0	0	0
RS	0	0	0	0	0	0	0.2	0.2	0.6
SC	0.6	0.2	0.2	0	0	0	0	0	0
SR	0	0	0	0.2	0.6	0.2	0	0	0
SS	0	0	0	0	0	0	0.2	0.1	0.7

Using this table, given that it rains on January 1st and second and we want to know the probability of it raining on January 4th, we just need to know the transition probabilites from the 2nd to the 4th through the 3rd.

So, the probability is

$= (0.5 * 0.5) + (0.3 * 0.2) + (0.2 * 0.2)$ = 0.35$$

So there's a 35% chance of it raining on the 4th.

For looking far in the future, matrix multiplication can be used. With A being the inverse of the above matrix and x^0 being the following array

$$x^0 = \left[0, \quad 0, \quad 0, \quad 0, \quad 1, \quad 0, \quad 0, \quad 0, \quad 0\right]^{-1}$$

$A^6 x^0$ gives us the probability

Using this table and running a numpy script with the above equation, the probability of it rainin gon the 8th given that it rained on the first and second is 27.2095 %.