Introduction to bioinformatics, Autumn 2007, Exercise 3

26.9.2007

1. (Chapter 3, Exercise 8) In a certain genome the bases appear to be iid and $p_g = 0.2$. Define the (binomial) count of the number of Gs in the first 1000 bases as $N = X_1 + X_2 + \cdots + X_{1000}$.
   (a) Give the mean and variance of N.
   (b) Approximate, using the Central Limit Theorem, $P(0 \leq N \leq 220)$ and $P(190 \leq N \leq 220)$.
   (c) Produce a histogram for 1000 replicates of N and compare the results with those of (b).

2. Describe the databases GenBank, RefSeq and UniProt found at the NCBI website at `www.ncbi.nlm.nih.gov` by answering the questions below.
   (a) What type of information is stored in each database?
   (b) How do the databases differ from each other?
   (c) In what ways are they similar?
   (d) Where is the data in the database obtained from?

3. Find the sequences with the accession numbers NM_079724 and NP_566125 from the NCBI databases. Describe both sequences by answering the following questions.
   (a) Is the sequence a dna or protein sequence?
   (b) How long is the sequence?
   (c) Which organism is it from?
   (d) What is the function of the sequence?

4. Perform global alignment of the sequences

```
s = ATTGCGAGTA
t = TCATGCGTCGA
```

with $\mu = 1$, $\delta = 2$ and uniform match score 1 by constructing the dynamic programming matrix. What is the optimal alignment and corresponding score?

5. Perform local alignment of the sequences

```
s = GGCGTAGTATACAGAGC
t = CTCTATGACTCGCAGTGA
```

with $\mu = 1$, $\delta = 2$ and uniform match score 1 by constructing the dynamic programming matrix. What is the optimal alignment and corresponding score?