Introduction to bioinformatics, Autumn 2006, Exercise 1

22.9.2006

1. (Chapter 2, Exercise 1) The base composition of a certain microbial genome is $p_G = p_C = 0.3$ and $p_A = p_T = 0.2$. We are interested in 2-words where the letters are assumed to be independent. There are $4 \times 4 = 16$ 2-words.

   (a) Present these 16 probabilities in a table. Do your 16 numbers sum to 1.0?

   (b) Purine bases are defined by $R = \{A, G\}$ and pyrimidine bases by $Y = \{C, T\}$. Let E be the event that the first letter is a pyrimidine, and F the event that the second letter is A or C or T. Find $P(E), P(F), P(E \cup F), P(E \cap F)$ and $P(F^c)$.

   (c) Set $G = \{CA, CC\}$. Calculate $P(G|E), P(F|G \cup E), P(F \cup G|E)$.

2. (Chapter 2, Exercise 5) Verify the terms in the first row of the transition matrix P presented in Section 2.6.3. Describe how you would use the sequence of *M. genitalium* to produce this matrix.

$$P = \begin{array}{c|cccc} & A & C & G & T \\ A & 0.423 & 0.151 & 0.168 & 0.258 \\ C & 0.399 & 0.184 & 0.063 & 0.354 \\ G & 0.314 & 0.189 & 0.176 & 0.321 \\ T & 0.258 & 0.138 & 0.187 & 0.415 \end{array}$$

3. (Chapter 2, Exercise 6) Find the stationary distribution of the chain with transition matrix P in Section 2.6.3; that is, solve the equations $\pi = \pi P$ subject to the elements of π begin positive and summing to 1. Compare π to the base composition of *M.genitalium*, and comment.

4. (Chapter 2, Exercise 12 (a) – (c)) In this exercise we have two random variables X and Y which are not independent. Their joint probability distribution is given in the following table:

		Y 1	3	6	9
	2	0.11	0.05	0.20	0.08
X	3	0.20	0.02	0.00	0.10
	7	0.00	0.05	0.10	0.09

The values of X are written in the first column and the values of Y in the first row. The table is read as $P(X = 7\&Y = 6) = 0.10$, and so on.

(a) Find the marginal distribution of X and Y. (That is, $P(X = 2), P(X = 3), \ldots$)

(b) Write $Z = XY$. Find the probability distribution of Z.

(c) The **covariance** between any two random variables is defined by

$$Cov(X,Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y).$$

Show that $Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}X \times \mathbb{E}Y$.

5. (Chapter 2, Exercise 12 (d) – (f))

(a) Find $\mathbb{E}X, \mathbb{E}Y, \sigma_X^2 = VarX, \sigma_Y^2 = VarY$, and $Cov(X,Y)$ for the example in the table.

(b) The **correlation coefficient** ρ is defined by $\rho_{X,Y} = Cov(X,Y)/\sigma_X\sigma_Y$. It can be shown that $-1 \leq \rho \leq 1$, the values ± 1 arising when Y is a linear function of X. Verify this last statement.

(c) Calculate ρ for the example in the table.