STATISTICAL METHODS IN BIOINFORMATICS

University of Helsinki

Exactum, Thursday 20th of January 2005,
16.00 - 20.00

Teacher on call: Timo Koski, tel. + 046 - 70 - 237 0047
Permittted means of assistance :

- Course lecture notes and handouts.
- Ewens & Grant: Statistical Methods in Bioinformatics, Springer 2001.
- L. Råde & B. Westergren: **BETA** Mathematics Handbook for Science and Engineering or other similar handbook.
- Calculators with blank memories. Programmes or manuals for calculators are not permitted.

The solutions should be presented using clearly introduced definitions and notations. Computations should be detailed enough to be easy to follow. All numerical answers should be given with the accuracy of two decimal points.

There are 6 (six) pages in this document. The exam consists of 5 (five) assignments. Each correct solution gives 6 (six) points. Preliminarily a minimum of 12 (twelve) points is required to pass the exam with the lowest grade.

The grades from the exam will be made public at the latest on Monday the 31st of January.

Good luck !

- -

Next page →

Assignment 1. Let Z_i, $i = 1, 2, \ldots$, be independent random variables with

$$Z_i = \begin{cases} 1 & \text{with probability } 1/4 \\ 0 & \text{with probability } 3/4. \end{cases}$$

Let

$$Y_i = Z_i - \frac{1}{9}(1 - Z_i)$$

The interpretation is that Y_i is the score at the ith position of a pairwise global alignment. The score is 1 for match, $-\frac{1}{9}$ for mismatch. Z_i, $i = 1, 2, \ldots$, is model for background noise for random DNA sequences.

(a) Find the probability generating function of the score over a finite segment

$$Y_1 + Y_2 + \ldots + Y_n \qquad (3p.)$$

(b) Assume that $N \in \text{Ge}\left(\frac{1}{2}\right)$ and independent of Y_i, $i = 1, 2, \ldots$, Find the probability generating function of the score over a segment of random length

$$Y_1 + Y_2 + \ldots + Y_N \qquad (3p.)$$

Assignment 2. In genetics the number of individuals i bearing a certain genetic configuration in a population with N individuals is thought to follow a Markov chain $\{X_n\}_{n\geq 0}$ with the state space $S = \{0, 1, \ldots, N\}$ and with the transition probabilities

$$p_{i|j} = P(X_{n+1} = j | X_n = i) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}$$

In other words, $X_{n+1}|X_n = i \in \text{Bin}\left(N, \frac{i}{N}\right)$.

(a) Is this Markov chain irreducible ? $(3p.)$
Hint: Check $p_{0|0}$ and $p_{N|N}$.

(b) Compute

$$E[X_{n+1} | X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i]. \qquad (3p.)$$

Assignment 3. A continuous time Markov chain is used as a model for both nucleotide substitution. The state space is $\{A, T, C, G\}$ represented as $i \in \{1, 2, 3, 4.$

The probability transition matrix $\mathbf{P}(t)$ is elementwise given by

$$P_{ii}(t) = \frac{1}{4}\left(1 + 3e^{-\frac{4}{3}\lambda t}\right),$$

and

$$P_{ij}(t) = \frac{1}{4}\left(1 - e^{-\frac{4}{3}\lambda t}\right), \quad j \neq i.$$

Assume that there are two continuous time Markov chains X and Y with the same generator $\mathbf{Q}$ (of $\mathbf{P}(t)$) that start at time $t = 0$ with the same value,

$$X(0) = Y(0) \in \pi$$

where π is the equilibrium distribution corresponding to $\mathbf{P}(t)$, and evolve independently thereafter. Find the fraction of divergence

$$P\left(X(t) \neq Y(t)\right).$$

Assignment 4. In a paper in the journal **CABIOS**, Vol.5, (1989) R. Staden introduced and used probability generating functions, e.g., for calculating the probabilities of scores of DNA words with respect to Position Specific Scoring Matrix (PSSM) $\mathcal{W}$.

The matrix $\mathcal{W}$ has as entries the scores (e.g. frequencies) w_{ij} of nucleotide i at position j of aligned (binding) sites.

We write

$$\{1, 2, 3, 4\} = \{A, T, C, G\}$$

and

$$\mathcal{W}: \begin{array}{ccccccc} A & w_{11} & \dots & w_{1j} & \dots & w_{1n} \\ T & w_{21} & \dots & w_{2j} & \dots & w_{2n} \\ C & w_{31} & \dots & w_{3j} & \dots & w_{3n} \\ G & w_{41} & \dots & w_{4j} & \dots & w_{4n} \end{array}$$

The generating function $G_j(t)$ for column j in $\mathcal{W}$ is given by

$$G_j(t) = \sum_{i=1}^{4} p_i t^{w_{ij}}$$

where p_i is the relative frequency (probability) of nucleotide i. The probability generating function $F(t)$ w.r.t. PSSM $\mathcal{W}$ is given by

$$F(t) = \prod_{j=1}^{n} G_j(t).$$

Hence the columns regarded as independent random units.

(a) What is the interpertation of the coefficient of t^k in $F(t)$? (4*p.*)

(b) Consider $n = 5$ and a nucleotide count matrix

$$\mathcal{W}: \begin{array}{cccccc} A & 9 & 1 & 1 & 10 & 7 \\ T & 0 & 0 & 7 & 0 & 0 \\ C & 1 & 9 & 1 & 0 & 0 \\ G & 0 & 0 & 1 & 0 & 3 \end{array}$$

Find the probability of getting PSSM score = 0 from columns 1 and 2 using $F(t)$ (or the appropriate factors). (1*p.*)

(c) Suppose that you have the transition probability matrix $\mathbf{P}$ for a Markov chain on $\{A, T, C, G\}$. How would you define the probability generating function $F(t)$ w.r.t a PSSM $\mathcal{W}$ using $\mathbf{P}$? (C.f, Huang et.al. in Journal of Computational Biology 2004). (1*p.*)

Assignment 5. In a Bayesian technique of computational discovery of gene regulatory binding sites in DNA we consider the following.

Let $\mathbf{x}^{(l)}$ be a sequence $\mathbf{x}^{(l)} = x_1^{(l)} x_2^{(l)} \dots x_L^{(l)}$, where each $x_i^{(l)}$ has values in $\{1, 2, 3, 4\} = \{A, T, C, G\}$. We let

$$\mathbf{S} = \{\mathbf{x}^{(l)}\}_{l=1}^{N}$$

be a set of sequences. We assume that there may be substrings of length w that are sites of an unknown motif of length w, $w \leq L$. The location of these sites is unknown, so we introduce a missing array of indicators

$$\mathbf{a}^l = \{a_j^{(l)}\}_{j=1}^{L-w+1},$$

where $a_j^{(l)}$ is either zero or one whether or not position j in sequence l is the starting site of a motif site.

We set

$$\mathbf{A} = \left\{\mathbf{a}^l\right\}_{l=1}^{N}$$

Then $\mathbf{S}(\mathbf{A})$ is the subset of $\mathbf{S}$ that consists only the bases in the motif sites, and $\mathbf{S}(\mathbf{A}^c)$ is the complementary subset. Let $\mathbf{S}(\mathbf{A}(j))$ be the set of nucleotides in the jth position of the motif sites.

Let

$$\mathbf{N}(\mathbf{A}(j)) = (n_{1j}, \ldots, n_{4j})$$

be the set of frequency counts of the nucleotides in the jth position of the motif sites.

We set

$$\theta_{ij} = \text{Probability of nucleotide } i \text{ at site } j$$

and

$$\theta_j = (\theta_{1j}, \theta_{2j}, \theta_{3j}, \theta_{4j})$$

and

$$\mathbf{\Theta} = \{\theta_j\}_{j \in \text{motif}}$$

We let

$$\theta_j^{\mathbf{N}(\mathbf{A}(j))} = \prod_{i=1}^{4} \theta_{ij}^{n_{ij}}.$$

In addition

$$\theta_0 = (\theta_{10}, \theta_{20}, \theta_{30}, \theta_{40})$$

is the probability vector of obtaining the nucleotides at non-motif sites.

Then we get the probability of $\mathbf{S}$ as

$$p(\mathbf{S}|\mathbf{\Theta}, \theta_0, \mathbf{A}) = \theta_0^{\mathbf{N}(\mathbf{A}^c)} \times \prod_{j=1}^{w} \theta_j^{\mathbf{N}(\mathbf{A}(j))}$$

Assume now that θ_j are independent and that each θ_j has a Dirichlet prior

$$\theta_j \in \text{Dirichlet}(\beta_j),$$

where

$$\beta_j = (\beta_{1j}, \beta_{2j}, \beta_{3j}, \beta_{4j}).$$

In addition

$$\theta_0 \in \text{Dirichlet}(\beta_0),$$

where

$$\beta_0 = (\beta_{10}, \beta_{20}, \beta_{30}, \beta_{40}) .$$

(a) Give the predictive probability of $\mathbf{S}$ obtained by

$$p(\mathbf{S}|\mathbf{A}) = \int p(\mathbf{S}|\mathbf{\Theta}, \theta_0, \mathbf{A})\, p(\mathbf{\Theta}, \theta_0)\, d\mathbf{\Theta} d\theta_0$$

where

$$p(\mathbf{\Theta}, \theta_0)$$

is the joint prior, a product of the pertinent Dirichlet densities, of all the parameters. ($5p.$)

(b) How would one use $p(\mathbf{S}|\mathbf{A})$ to actually discover the binding sites ? ($1p.$)