英語版　English version

受験番号 Examinee number					

東京大学 大学院新領域創成科学研究科 情報生命科学専攻

Department of Computational Biology, Graduate School of Frontier Sciences
The University of Tokyo

平成 23（2011）年度

2011 School Year

大学院入学試験問題　修士・博士後期課程

Graduate School Entrance Examination Problem Booklet, Master’s and Doctoral Course

専門科目

Specialties

平成 22 年 8 月 24 日（火）
Tuesday, August 24, 2010
13：30〜15：30

注意事項 **Instructions**

１．試験開始の合図があるまで、この冊子を開いてはいけません。
Do not open this problem booklet until the start of examination is announced.

２．本冊子の総ページ数は 14 ページです。落丁、乱丁、印刷不鮮明な箇所などがあった場合には申し出ること。
This problem booklet consists of 14 pages. If you find missing, misplaced, and/or unclearly printed pages, ask the examiner.

３．解答には必ず黒色鉛筆（または黒色シャープペンシル）を使用しなさい。
Use black pencils (or mechanical pencils) to answer the problems.

４．問題は 8 題出題されます。問題 1〜8 から選択した合計4問に解答しなさい。ただし、問題 1〜8 は同配点です。
There are eight problems (Problem 1 to 8). Answer four problems out of the eight problems. Note that Problem 1 to 8 are equally weighted.

５．解答用紙は計４枚配られます。各問題に必ず１枚の解答用紙を使用しなさい。解答用紙に書ききれない場合は、裏面にわたってもよい。
You are given four answer sheets. You must use a separate answer sheet for each problem. You may continue to write your answer on the back of the answer sheet if you cannot conclude it on the front.

６．解答は日本語または英語で記入しなさい。
Answers should be written in Japanese or English.

７．解答用紙の指定された箇所に、受験番号と選択した問題番号を記入しなさい。問題冊子にも受験番号を記入しなさい。
Fill the designated blanks at the top of each answer sheet with your examinee number and the problem number you are to answer. Fill the designated blanks at the top of this page with your examinee number.

８．草稿用紙は本冊子から切り離さないこと。
The blank pages are provided for making draft. Do not detach them from this problem booklet.

９．解答に関係ない記号、符号などを記入した答案は無効とします。
An answer sheet is regarded as invalid if you write marks and/or symbols unrelated to the answer on it.

10．解答用紙・問題冊子は持ち帰ってはいけません。
Do not take the answer sheets and the problem booklet out of the examination room.

（このページは草稿用紙として使用してよい）

(Blank page for draft)

（このページは草稿用紙として使用してよい）
(Blank page for draft)

Problem 1

The following recursive equation shows the iteration used in the dynamic programming for the local alignment of biological sequence x of length m and biological sequence y of length n. The x_i shows the i-th (i = 1, …, m) character of sequence x, the y_j shows the j-th (j=1,…, n) character of sequence y, and the $s(a,b)$ shows the value of substitution matrix of character a and character b, and d>0. Answer the following questions.

$$F(i, j) = \max\begin{cases} 0 \\ F(i-1, j) - d \\ F(i, j-1) - d \\ F(i-1, j-1) + s(x_i, y_j) \end{cases} \qquad (1 \le i \le m,\ 1 \le j \le n)$$

(1) Explain briefly what value d indicates, including what kind of model is supposed for the evolution of the sequences.

(2) Explain why 0 in the first line of the max operation is required, including what kind of alignment is calculated by using this recursive equation.

(3) In order to calculate this iteration, the initial values of $F(i, j)$ when i = 0 or j = 0 are given. Show all those initial values.

(4) Show the computational complexity (in time and in space) for calculating all the values of $F(i, j)$ by this recursive equation.

(5) A local alignment that gives the maximum alignment score is obtained by a procedure called traceback after finding the maximum value of $F(i, j)$. Explain briefly about the traceback in this case.

Problem 2

Genomic sequences are subjected to random substitutions along evolutionary time. Let us define a continuous-time Markov model for the base mutation at each genomic position.

$$P(b|a,t,\lambda) = [\exp(tS)]_{b,a} \quad S = \lambda\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

Here, $t \geq 0$ denotes evolutionary time, natural numbers $a, b \in \{1,2\}$ indicate 1=purine (adenine or guanine) or 2=pyrimidine (cytosine or thymine), $P(b|a,t,\lambda)$ is the probability of observing base b at time t on condition that the base is a at time 0. $\lambda \geq 0$ represents the substitution rate, and $[A]_{b,a}$ stands for the (b,a) element of matrix A. $\exp(A)$ is the matrix exponential of A and is defined as follows,

$$\exp(A) = I + \frac{A}{1!} + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{A^k}{k!},$$

where I represents the unit matrix.
Answer the following questions.

(1) Let us define a matrix $M = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. Show $M^2 = 2M$.

(2) For a real number c, show the relation,

$$\exp(cM) = I + \frac{e^{2c} - 1}{2} M.$$

(3) Compute $P(b|a,t,\lambda)$ by using the relation $S = -2\lambda I + \lambda M$ as well as the following identity on matrix exponentials,

$$\exp(A + B) = \exp(A)\exp(B) \quad \text{if } AB = BA.$$

Suppose that the genome sequences of a virus are sampled at different time points T_1, T_2 $(T_2 - T_1 = T > 0)$. Let $n_{b,a}$ represent the observed number of the sequence positions where base a changes to base b (including the cases $a = b$). Let C denote all the evolutionary changes $\{n_{b,a} |\, a, b \in \{1,2\}\}$. The probability $P(\mathrm{C}|T,\lambda)$ that such changes C occur over time T is given by,

$$P(\mathrm{C}|T,\lambda) = \prod_{a,b \in \{1,2\}} P(b|a,T,\lambda)^{n_{b,a}}$$

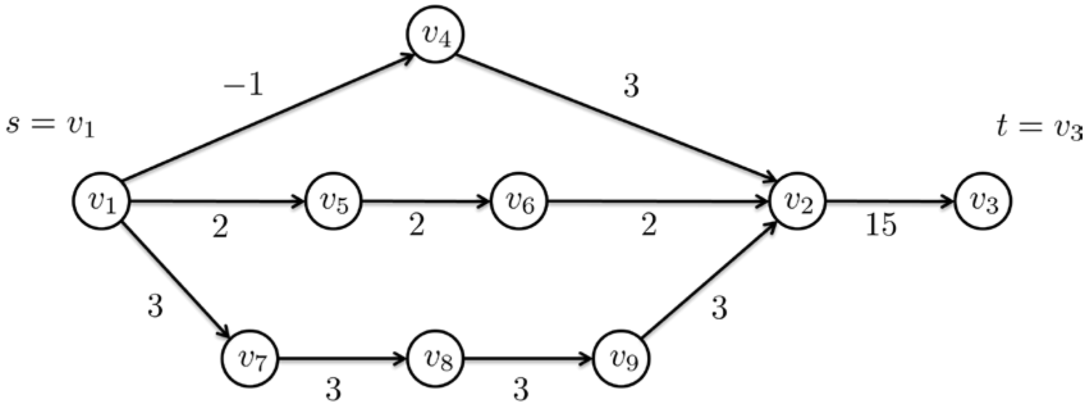
(4) The parameter λ that maximizes the log likelihood function L which is defined by $L(\mathrm{C},T,\lambda) = \log\left(P(\mathrm{C}|T,\lambda)\right)$ is called the maximum likelihood estimate of λ. Obtain the maximum likelihood estimate of λ and its existence conditions by solving the stationary equation which is defined by,

$$\frac{\partial L(\mathrm{C},T,\lambda)}{\partial \lambda} = 0.$$

Problem 3

Consider a directed acyclic graph $G=(V,E,w)$, where V is a set of vertices, E is a set of edges, $w(e)$ is an integer weight function for an edge $e\in E$. The weight of a path on G is defined as the sum of the weights of the edges on the path. Given $s\in V$ and $t\in V$, we will find a path from s to t such that the average weight, which is the weight of the path divided by the number of the edges on the path, is minimized. Such a path is called a *minimum average weight path* hereafter.

(1) In the graph shown below, find the weight of the path $v_1 \to v_4 \to v_2 \to v_3$.



(2) Find the minimum average weight path from s to t for the graph shown in (1).

(3) Assume that the weight of a minimum average weight path from s to t on G is equal to or less than a constant C. Prove the proposition below.
"Let G' be the weighted graph (V,E,w'), where $w'(e)=w(e)-C$. The weight of the shortest path from s to t on G' is equal to or less than zero."
Recall that the shortest path is a path with the minimum weight.

(4) Design an algorithm that searches the shortest path from s to t in polynomial time in the sizes of E and V.

(5) Design an algorithm that finds a minimum average weight path in polynomial time in $\log(\max_{e\in E}|w(e)|)$ and the sizes of E and V.

Problem 4

Let P be a string of characters. In P, a consecutive substring is a prefix if it starts from the first character of P, and it is a suffix if it ends at the last character of P. A substring shorter than P is called a "prefix-suffix match" if it is both a prefix and a suffix. For example, when P = CGCGCG, CGC is a prefix, and GCG is a suffix. CGCG and CG are prefix-suffix matches. The empty substring is denoted by ε and is treated as a prefix and a suffix. To design an algorithm for listing all prefix-suffix matches, answer the following questions.

(1) For each case of P = CGCGCGCG and P = AAAAA, list all prefix-suffix matches.

(2) Let $\delta(P)$ denote the longest prefix-suffix match of P. Define $\delta(\varepsilon)=\varepsilon$. Let $\delta^k(P)$ denote the result of k ($\geq$1) applications of δ to P. For each case of P = CGCGCGCG and P = CGCGCGCGC, answer $\delta(P)$ and $\delta^2(P)$.

(3) Prove that $\delta^k(P)$ ($k \geq 1$) is a prefix-suffix match of P. On the other hand, prove that any prefix-suffix match of P is equal to $\delta^m(P)$ for some m ($\geq$ 1).

(4) Design an algorithm for listing all prefix-suffix matches of P in time linear to the length of P.

Problem 5

The following nucleotide sequence, which shows only one DNA strand in the direction of 5'→3', is a part of the whole genome sequence (reference sequence) of a certain organism. Answer the following questions about this sequence.

GGATCCCCCATATAGGCCCGGGAATTCCGGAAAAATTTATTTGCTAAGCAGATGGTTGG

AAAATGATTTTCCCCAAGGGGGTCATGATCGATCATTAATTATATAATGGACCCCCGTT

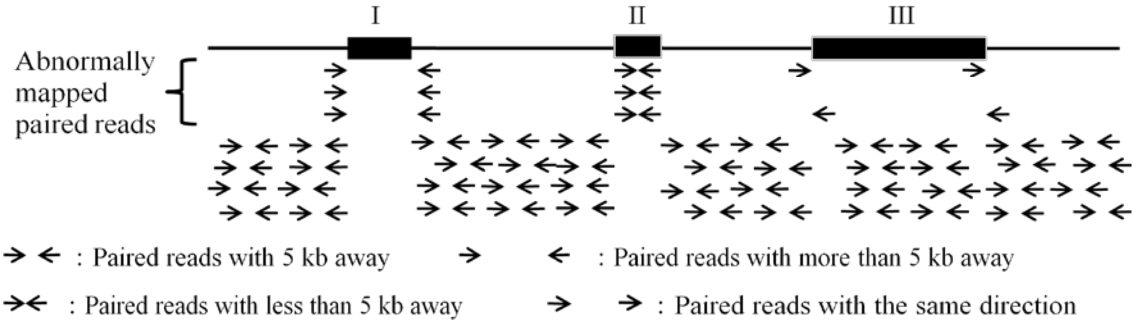
TGACTGGAATTTCCCCGATTACAGATGACTGATGCATGATGGGGCCCTTGACCAAAAGG

CCGCGCGCGCATTTAGGCCCATGGTTTTCCCATGACATGGGGATGAGATATACCCCGGG

AAGGGATAGTATCGATTAGCAAATTTACCAAGGCCATGACATTGCATCTGAATGCCAAA

AAATTGGGAGAGACTCTCTCTAGGCCATTGATACCAATTTTGCGTTTTTAGCAGGATCC

(1) Whole genome shotgun sequencing of individual X of this organism was done, and numbers of paired reads of about 500 bases were obtained from both ends of template DNA of which the size is about 5 kb. Next, the paired reads were mapped onto the reference sequence. As shown in the figure below, the result showed the presence of three regions where paired reads were abnormally mapped with longer (I) and shorter distance (II) than the expected 5 kb and with the same direction (III). What structural difference is there in the region I, II and III in individual X genome compared to the reference sequence? Choose one for I, II and III from the following four terms (*), respectively.


I
II
III
Abnormally mapped paired reads
→ ← : Paired reads with 5 kb away
→ ← : Paired reads with more than 5 kb away
→← : Paired reads with less than 5 kb away
→ → : Paired reads with the same direction

(*) Insertion, Deletion, Duplication, Inversion

(2) The result also showed that individual X has base A at the position of base G underlined in the reference sequence. What is such a base difference found between different individuals in the same organism generally called?

(3) PCR primers were designed from the sequences shown with double underlines in the reference sequence to amplify the 236 bp region containing base G underlined. Choose one correct primer set from the following a ~ f. All sequences in a ~ f are given in the direction of 5'→3'.

a: AAAATGATTTTCCCCAAGGG and CATTGCATCTGAATGCCAAA

b: AAAATGATTTTCCCCAAGGG and AAACCGTAAGTCTACGTTAC

c: CCCTTGGGGAAAATCATTTT and CATTGCATCTGAATGCCAAA

d: AAAATGATTTTCCCCAAGGG and TTTGGCATTCAGATGCAATG

e: CCCTTGGGGAAAATCATTTT and TTTGGCATTCAGATGCAATG

f: CCCTTGGGGAAAATCATTTT and AAACCGTAAGTCTACGTTAC

(4) PCR was done with above PCR primers to amplify the 236 bp region. Which is the numerical value closest to the number of molecules amplified after 30 cycles in the following a ~ f? In the PCR, it is assumed that 100 molecules of 236 bp DNA were generated in the first cycle and all amplified molecules were used as templates from the second cycle. In the calculation, 2^{10} can be approximated with 1000.

a: 5 x 10^{7}, b: 5 x 10^{8}, c: 5 x 10^{9}, d: 5 x 10^{10}, e: 5 x 10^{11}, f: 5 x 10^{12}

(5) In the following a ~ f, which is the numerical value closest to the weight (ng) of 236 bp DNA yielded in the above-mentioned PCR experiment? In this calculation, the molecular weight of 1 bp of DNA is assumed to be 600 and Avogadro's number to be 6 x 10^{23}, respectively.

a: 0.001, b: 0.01, c: 0.1, d: 1, e: 10, f: 100

(6) The genome size of this organism is about 5 Mb and its mRNA molecules have no polyA at the 3' end. Choose one organism taxonomically closest to this organism from the following organisms.

Nematoda, Drosophila, Arabidopsis, Medaka, Tetrahymena, Yeast, Aspergillus, E. coli

Problem 6

Our understanding of the chromatin remodeling is rapidly expanding in recent years. The chromatin remodeling plays a major role in regulation of gene expression and is essential in diverse biological processes. It is now becoming clear that the DNA methylation and histone modifications can influence chromatin structure. DNA methylation is known to be involved in cell differentiation, imprinting, X-chromosome inactivation in mammalian cells. Mammals have two types of DNA methylation enzymes, DNA methyltransferase 1 (Dnmt1) and DNA methyltransferase 3 (Dnmt3a, Dnmt3b). These two types of enzymes have different functions.

Answer the following questions related to the underlined terms.

(1) Which base is most frequently methylated in mammalian genome? And, write the dinucleotide sequence that contains this methylated base most frequently (Write the sequence from 5’).

(2) Acetylation is one of types of the histone modification. List other three types of the histone modification.

(3) Explain the functions of DNA methyltransferase1 and DNA methyltransferase 3. (Write the answer in a few lines for each.)

Problem 7

(1) Explain two types of cytoskeletons using all the terms listed below. (It is allowed to use the same term several times.)

myosin - microtubules - kinesin - centrosome - actin - dynein -
microfilaments - flagellum - tubulin - muscle - plus end - minus end

(2) The drug taxol binds tightly to microtubules and stabilizes them to help the assembly of microtubules. Another drug colchicine prevents microtubule formation. The two drugs with the opposite effects are both used as anticancer drugs. Explain why they are highly toxic to cancer cells despite their opposite actions. (Write the answer in a few lines.)

(3) Concerning two types of signal transduction systems, classify the following terms (1 〜 15) into (A) those that are more closely associated with the signal transduction mediated by the G-protein-coupled receptors (GPCR), and (B) those that are more closely associated with the signal transduction mediated by the receptor tyrosin kinase (RTK). Answer the combination of term numbers and classifications like 20-A, 21-B…

1. adaptor protein - 2. cyclic AMP - 3. Ca^{2+} - 4. MAP kinase -
5. endoplasmic reticulum - 6. seven-pass transmembrane protein - 7. RAS -
8. adenylyl cyclase - 9. small GTP-binding protein - 10. heterotrimeric G protein -
11. α subunit - 12. phospholypase C - 13. inositol trisphosphate (IP_3) -
14. protein kinase A (PKA) - 15. second messenger

Problem 8

(1) For each protein listed in A, choose the most related molecule from B.
A. hemoglobin, carboxypeptidase, lysozyme, rhodopsin
B. retinal, heme, zinc ion, glycan

(2) For each protein listed in C, choose the most appropriate structural feature from D.
C. neuraminidase, collagen, CAP protein, α-keratin
D. dimer, triple-stranded helix, coiled coil, tetramer

(3) Explain enzyme allostery using all the following words.
regulatory molecule, active site, conformational change, substrate

（このページは草稿用紙として使用してよい）

(Blank page for draft)

（このページは草稿用紙として使用してよい）
(Blank page for draft)