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 24（2012）年度

2012 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 23 年 8 月 23 日（火）
Tuesday, August 23, 2011
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 12 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 12 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)

Problem 1

The following heap_sort function written in the C programming language sorts an array a of {key,value}-pairs based on the heap sort algorithm.

```
typedef  struct { int key;  int value; }  T;
void  swap(T a[],  int i,  int j) { T t = a[i];  a[i] = a[j];  a[j] = t; }
int  less(T a[],  int i,  int j) { return  a[i].key < a[j].key; }
void  sift_down(T a[],  int i,  int j) {
    int k;
    for (k = (2 * i + 1);  k < j;  k = (2 * i + 1)) {
        if ((k + 1 < j) && less(a, k, k + 1)) { ++k; }
        if (less(a, i, k)) { swap(a, i, k);  i = k;} else { break; }
    } /* end for */
} /* end sift_down */
void  heap_sort(T a[],  int n) {
    int i, j;
    for (i = (n - 1);  i >= 0;  --i) { sift_down(a, i, n); }
    for (j = (n - 1);  j > 0;  --j) { swap(a, 0, j);  sift_down(a, 0, j); }
} /* end heap_sort */
```

where n>0 represents the size of array a. Answer the following questions.

(1) Explain the heap data structure.

(2) Suppose heap_sort(a, n) is called with arguments a={{4,1},{3,2},{3,3},{2,4}} and n=4. Obtain the content of array a after the function call.

(3) Explain the roles of the sift_down function.

(4) Explain that the worst-case time complexity of the heap sort algorithm is $O(n \log n)$.

Problem 2

Let us consider the chaining method for identifying evolutionarily conserved regions by comparing two genomes efficiently. Let P be a string, and let $P[sx, ex]$ denote the consecutive substring ranging from the start position sx to the end position ex ($sx < ex$). In strings P and Q, if substrings $P[sx, ex]$ and $Q[sy, ey]$ are similar, it is called an alignment and is denoted by $A = (P[sx, ex], Q[sy, ey], score)$, where A is the name of the alignment, and $score$ is a real number that represents the similarity. The elements in the alignment, sx, ex, sy, ey, and $score$, are denoted by $A.sx$, $A.ex$, $A.sy$, $A.ey$, and $A.score$, respectively. Alignments are ordered $A < B$ if and only if $A.ex < B.sx$ and $A.ey < B.sy$. We assign an ascending order of alignments $A_1 < A_2 < \ldots < A_k$ ($k \geq 1$) to A_k and call it the chain of A_k. The score of A_k's chain is $\Sigma_{i=1,\ldots,k} A_i.score$ (denoted by $A_k.chain\_score$) and k is the length of the chain. A chain of a large score suggests an evolutionarily conserved region. Answer the following questions.

(1) List all chains of length 3 in Figure 1.

(2) In Figure 1, answer the chain with the maximum score of each alignment.

Chaining method: For each alignment A, set the chain of A to A itself initially, and insert two pairs $(A.sx, A)$ and $(A.ex, A)$ into list X. Sort X in the ascending order of the first elements of the pairs. Create another list Y for storing data of the form $(C.ey, C)$, and set Y to the empty list. Repeat the process of deleting the first element (x, B) from X and applying the following steps to (x, B) until X becomes the empty list.

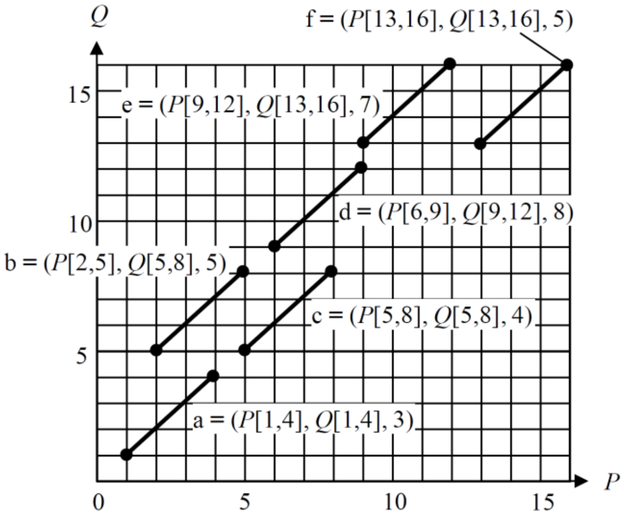

Figure 1

I. When x is the start position $B.sx$, if Y has $(C.ey, C)$ such that $C < B$, assume that $C.ey$ is the value closest to $B.sy$, append B to the end of C's chain, and let the concatenation be B's chain.

II. When x is the end position $B.ex$, if Y does not contains $(C.ey, C)$ such that $C.ey \leq B.ey$ and $C.chain\_score \geq B.chain\_score$, insert $(B.ey, B)$ to Y, sort Y in the ascending order of the first elements, and delete any $(D.ey, D)$ such that $B.ey \leq D.ey$ and $B.chain\_score > D.chain\_score$ from Y.

(3) Apply the chaining method to the alignments in Figure 1. Illustrate how the algorithm works. Show that the algorithm outputs the chain of the maximum score that ends at each alignment.

(4) After Step II, show that pairs in Y are sorted in the ascending order of the first elements and the scores of the chains of the alignments. When the number of alignments is n, give the name of a data structure and explain operations on it so that the chaining method runs in $O(n \log n)$.

Problem 3

When a sequencer processes four kinds of nucleotides in DNA sequences, read error often occurs. The sequencer that is considered in this problem skips no nucleotides and outputs no additional nucleotides, but the probabilities that nucleotides are incorrectly read are known as follows:

The probabilities of incorrect reading, A as C, C as A, G as T, T as G are equally p.

The probabilities of incorrect reading for the other combinations are equally q.

For example, A is incorrectly read as C with probability p, as G with probability q, as T with probability q. By assuming the above conditions, answer the following questions.

(1) When the original sequence is ATGGCT, calculate the probability that it is incorrectly read as ATATCT .

(2) When A, T, G, C appear at random with the equal probability (=25%), suppose two nucleotides are read as C. Calculate the probability that the two nucleotides are the same (not necessary to be C).

(3) Two totally random sequences of length three in which A, T, G, C appear at random with the equal probability (=25%) are both read as ATG. Calculate the probability that the two sequences are not exactly the same.

(4) A nucleotide, whose prior probabilities to be A, T, G, C are 20%, 20%, 30%, 30% respectively, is read as C. Calculate the probability that this result is true.

Problem 4

We consider process scheduling algorithms of Operating System (OS). We assume that we have a single Central Processing Unit (CPU) core. Process execution requests are arriving in the order and timing shown in the table below. The CPU burst time of each process is estimated accurately; the CPU burst time of a process is defined as the total time for which a CPU works in order to finish that process. We assume that the OS has a single process scheduling queue (process waiting queue).

Process	Arriving Time (ms)	CPU burst time (ms)
$\boldsymbol{P_1}$	0	135
$\boldsymbol{P_2}$	20	35
$\boldsymbol{P_3}$	100	90

(1) The First-Come, First-Served (FCFS) scheduling is a scheduling algorithm in which a process arrived earlier takes priority over the other processes that arrived later. Find the time when Process $\boldsymbol{P_2}$ ends under the FCFS scheduling.

(2) The Shortest-Remaining-Time-First (SRTF) scheduling is a scheduling algorithm in which a process with the shortest remaining CPU burst time runs first. When the CPU burst time of a new process that has just arrived is the shortest, the running process is suspended and the new process is executed instead. Under the SRTF scheduling and with the example above, calculate how much shorter the average time duration between a process arrival and its completion (we denote it by *the average waiting time* hereafter) would be than that for the FCFS scheduling.

(3) Generally speaking, the SRTF scheduling is preferred because it can minimize the average waiting time. However, considering the nature of processes being executed in the real world, the SRTF scheduling often finds difficulty in minimizing the average waiting time. Answer the reason.

(4) The Round-Robin (RR) scheduling is an extension of the FCFS scheduling algorithm; a process having been run continuously for a predefined length of time, which is called *time slice*, is preempted by the next process in the process scheduling queue and the preempted process is put into the tail of the process scheduling queue. Suppose that we use the RR scheduling with the time slice of 50 ms in the example above. Draw the timeline of the processes being executed on the CPU.

(5) A computer with many CPU cores and with cache-coherent Non-Uniform Memory Access (cc-NUMA) architecture may hold a waiting process in the process scheduling queue even if a free CPU is available and the OS can assign the waiting process to it. Explain the general architecture of cc-NUMA, and answer the reason why that strategy may finish processes faster in some situation.

Problem 5

Answer the following questions about the prokaryote.

(1) List and explain three differences with respect to the gene structure, the mRNA structure and the transcription mechanism between the prokaryote and the eukaryote.

(2) The genome size of most of bacteria ranges from one hundred and tens Kb to about ten Mb. Explain the key factor (evolutionary aspect) that causes this difference in bacterial genomes.

(3) For each word (1~3) listed in A, choose the most related word from the nine words listed in B.

A

1. Operational taxonomic unit (OTU)
2. Pribnow Box
3. Shiga toxin

B

Shine-Dalgarno sequence, 16S ribosomal RNA gene, Spore,
RNA polymerase recognition site, GC skew, Palindromic sequence,
Horizontal gene transfer, Replication origin, Symbiosis

(4) For each description (1~3) in C, choose the most related bacterium from the nine bacteria listed in D.

C

1. The whole genome of this bacterium was first sequenced in the world in 1995.
2. This soil bacterium possesses a linear chromosome that encodes many genes for the biosynthesis of secondary metabolites including antibiotics.
3. This bacterium is a member of human intestinal microbiota and is taxonomically classified into the phylum *Actinobacteria*.

D

Escherichia coli, Bacillus subtilis, Bifidobacterium adolescentis, Clostridium tetani,
Buchnera aphidicola, Haemophilus influenzae, Streptomyces griseus, Helicobacter pylori,
Staphylococcus aureus.

Problem 6

RNA is a biological macromolecule that has a number of different functions. A typical mammalian cell contains 10-30 pg total RNA, representing about 1% of the total weight of cell components. RNA can be roughly classified into two types depending on function. The first type is coding RNA or messenger RNA (mRNA) which encodes protein. The second one is noncoding RNA which does not encode protein. Noncoding RNA is also called functional RNA and is further functionally classified into several subtypes such as (a), (b), (c) and microRNA. Genes encoding microRNAs are initially transcribed by RNA polymerase II to generate precursor molecules called foldback RNAs having one or more internal hairpin structures which are formed through intrastrand base pairs. In the nucleus, foldback RNAs are cleaved by the enzyme (d) to generate small individual pre-miRNA hairpins. These pre-miRNAs are transported to the cytoplasm, where they are further cleaved by the enzyme (e) to generate microRNAs of 21-25 nucleotides. MicroRNA has a biological function (f).

Answer the following questions.

(1) Explain one method for quantifying a particular mRNA.
(2) List three functional RNAs for (a), (b), (c) and explain their functions.
(3) Answer the names of enzymes for (d) and (e), respectively.
(4) Explain the biological function of microRNA (f) in several lines.

Problem 7

Answer the following questions concerning DNA damage and repair.

(1) For each DNA damage listed in I, choose the most related base from II, and choose the most typical result of the damage from III.

I. Deamination, Thymine dimer, Depurination
II. A, T, C
III. Sickle-cell anemia, GC base pair deletion, Change of G to A, AT base pair deletion, Xeroderma pigmentosum, Change of C to T

(2) Choose the three enzymes most related to the DNA repair of a single base damage from IV and answer the order in which they function.

IV. Telomerase, DNA polymerase, DNA ligase, RNA polymerase, DNA photolyase, Nuclease

(3) DNA double strand break can be repaired by nonhomologous end-joining or homologous recombination. Compare these two mechanisms and expected results, and explain them briefly.

Problem 8

(1) Answer the following questions about iodine.

(A) Inhaling radioactive iodine increases the risk of thyroid cancer. Why is the thyroid particularly affected by radioactive iodine? Explain the reason in a few lines.

(B) Taking tablets of radioactive iodine is one of the popular treatments for thyroid cancer. Explain, in a few lines, why such a treatment would be effective.

(2) Chronic myelogenous leukemia (CML) is a disease that occurs in 1-2 people per 100,000 populations, accounting for about 20% of all the leukemia patients. CML patients show massive buildup of white blood cells due to abnormally elevated cell division of hematopoietic stem cells in the bone marrow. A female patient was diagnosed with CML. Karyotype analysis of the chromosomes of her cells showed that the leukemic cells were heterozygous for a reciprocal translocation involving chromosomes 9 and 22. However, such translocation was not observed in her other cells.

(A) Should the mutation caused by the translocation be best described as dominant? Or, recessive?

(B) Examination of her leukemic cells revealed the existence of peculiar proteins, which look like a fusion of two proteins that are usually observed as independent molecules. Explain the molecular mechanism for the generation of such proteins in 1-2 lines.

(C) The patient got pregnant later. What is the probability, compared with the mothers without CML, that her child will have CML? Explain in 1-2 lines.

(3) Oxygen ^{16}O (atomic weight = 16) and its isotope ^{18}O (atomic weight = 18) can be distinguished according to the slight difference in their weights. Oxygen and carbon dioxide in the normal air contain only ^{16}O.

(A) Put the mixture of normal air and oxygen gas of ^{18}O in a tight-sealed box. Put a mouse also in the box and let it breathe for 10 minutes. After that, the gas retrieved from the box was analyzed using a mass spectrometer. How would be the relative ratio of the carbon dioxide gas of ^{16}O to that of ^{18}O? Explain the reason referring to the molecular mechanisms of respiration.

(B) Put the mixture of normal air and carbon dioxide gas of ^{18}O in a tight-sealed box. Put an *Arabidopsis* plant also in the box and let it perform photosynthesis for 6 hours. After that, the gas retrieved from the box was analyzed using a mass spectrometer. How would be the relative ratio of the oxygen gas of ^{16}O to that of ^{18}O? Explain the reason referring to the molecular mechanisms of photosynthesis.

（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)