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 25（2013）年度

2013 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 24 年 8 月 21 日（火）

Tuesday, August 21, 2012

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 20 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 20 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 12 題出題されます。問題 1～12 から選択した合計4問に解答しなさい。ただし、問題 1～12 は同配点です。
   There are 12 problems (Problem 1 to 12). Answer 4 problems out of the 12 problems. Note that Problem 1 to 12 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 4 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

For each of the following problems (1)–(10), select the **single best answer** from 1–4.

(1) Choose the most **inaccurate** description on the two's complement of binary numbers.

1. It is used to express both positive and negative numbers
2. It is used to express floating-point numbers
3. It is equivalent to reversing each 0-1 bit and then adding one
4. It can express 65536 integer numbers using 16 bits

(2) In the complexity theory, the O notation gives the upper bound of functions and the Ω notation, the lower bound. Choose the most **inaccurate** expression.

1. $2 \log n \in O(\log n)$
2. $2 \log n \in \Omega(\log n)$
3. $n \in O(n^2)$
4. $n \in \Omega(n^2)$

(3) Choose the closed form of the recurrence formula $T_0=0$, $T_n=2T_{n-1} + 1$.

1. $T_n = 2^n - 1$
2. $T_n = 2(2^n - 1)$
3. $T_n = 2n + 1$
4. $T_n = (2n + 1)^n$

(4) Choose the most **accurate** description on the depth-first traversal on trees.

1. The root does not become the first node in the postorder traversal.
2. The root does not become the last node in the preorder traversal.
3. The postorder traversal and preorder traversal do not coincide.
4. The traversal ends in a linear time in terms of the tree size.

(5) Choose the most **appropriate** pairing of the tree-search method and its data structure.

1. Depth-first search and a stack
2. Depth-first search and a priority queue
3. Breadth-first search and a stack
4. Breadth-first search and a hash table

(6) Let B(m,n) be the binomial coefficient ${}_mC_n$. Find the value of B(m–1,n) + B(m,n–1) + B(m–1,n–1).

1. B(m+1,n+1)
2. B(m+1,n)
3. B(m,n+1)
4. None of the above

(7) Choose the most **inaccurate** description on the quicksort of n items.

1. A pivot is chosen to divide data into two parts.
2. The worst case time-complexity is $O(n^2)$.
3. The dividing step is repeated O(n) times.
4. Sorting completes in O(log n) time in the fastest case.

(8) Choose the most **inaccurate** description on the linked list of n items.

1. Average time-complexity of linear scan is O(log n)
2. It uses more memory space than arrays.
3. Insertion or deletion of items is efficient.
4. It is used to create hash tables.

(9) Choose the most **inaccurate** description on B-trees.

1. The path lengths from the root to all leaves are equal.
2. It is used for database searches.
3. The node degree is determined by considering page sizes of memory accesses.
4. It is an improved version of suffix trees.

(10) Choose the most **inaccurate** description on Fibonacci heaps.

1. It is an improved version of binary heaps.
2. It delays reconciliation of heap structures as much as possible.
3. It consists of trees whose sizes correspond to Fibonacci numbers.
4. It unites the roots of multiple heap structures using a list.

Problem 2

Let A, B, and C be n x n matrices. Let tr(A) denote the trace of A, the sum of the diagonal elements of A.

(1) Show tr(AB) = tr(BA).
(2) Give a counter example of tr(ABC) = tr(ACB).
(3) When C has an inverse C^{-1}, show tr($C^{-1}AC$) = tr(A).
(4) When A is a real symmetric 2 x 2 matrix, tr(A) = 1, and tr(A^2) = 1, answer the eigenvalues of A.

Problem 3

Answer the following questions about a sorting problem A, which sorts an array of size 2^n, x = x[0], x[1], …x[2^n – 1] in a descending order. Suppose that n is a positive integer, and that each element of x is a positive integer no greater than N.

(1) The following code is a part of the program that solves the sorting problem A by a merge sorting.
Explain what calculation `merge()` does.
What arguments are used for `merge_sort()` when solving the sorting problem A?

```
void merge_sort(int a[], int left, int right){
        if (right == left) return;
        int mid = (left + right - 1)/2;
        merge_sort(a, left, mid);
        merge_sort(a, mid+1, right);
        merge(a, left, right);
}
```

(2) How many times `merge()` is called when solving the sorting problem A?
Show the number of mutual comparisons between elements of x required to solve the sorting problem A is O($n2^n$).

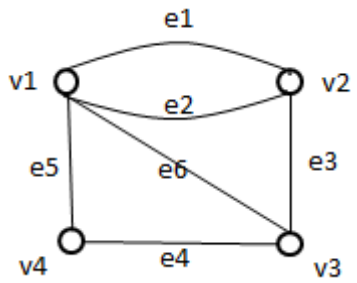
(3) The sorting problem A can be solved by a sorting method B, which assigns a memory of size N corresponding to the number of possible values of the elements of x, and counts the number of times that each value appears in x. Explain merits and demerits of the sorting method B compared with the program in (1), in terms of computation time and required memory.

(4) Explain how the sorting problem A is solved by a radix sorting, which repeats a sorting procedure that is similar to the sorting method B, including the viewpoints of the computation time and required memory.

Problem 4

Answer the following questions about an undirected and non-empty graph with no self-loops.

(1) Write incidence matrix M for the following graph G.
Incidence matrix is a matrix in which an element m_{ij} is 1 if a vertex v_i and an edge e_j are adjacent, otherwise 0.



(2) What does the sum of elements in each row of M mean?

(3) Show that the sum of the degrees in a graph G doubles the number of edges in G. Here, the degree of a vertex is the number of edges which are adjacent to the vertex.

(4) Show that the number of vertices whose degrees are odd is even.

(5) Show that if every vertex of a graph G has a degree greater than 1, G has at least one cycle.

(6) Let p be the number of vertices of a tree T, which is a connected graph without a cycle, and q be the number of edges of T. Prove that p=q+1.

Problem 5

In statistics, time series analysis deals with sequential data obtained by repeated measurements over time. In the following, we consider a simple statistical problem in which a data set is represented by a finite array.

Let X be a random array of length N in which n elements have value of 1 and the remaining $(N-n)$ elements have value of 0 $(0 < n < N)$. Suppose that every possible array X is observed with equal probability. Let m denote the number of 1's in the first M elements of X $(0 < M < N)$. Answer the following questions.

(1) Answer the range of possible values of m for a given N, M, and n.

(2) Answer the expected value of m.

(3) Explain that the probability distribution $P(m|N,M,n)$ of m is given by the following hypergeometric distribution. Here, $\binom{K}{k} = \frac{K!}{(K-k)!k!}$ represents the binomial coefficient.

$$P(m|N,M,n) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

(4) Show that the right hand side of the above equation can be rewritten as follows. Here, we define $(K)_0 = 1$ and $(K)_k = \prod_{i=0}^{k-1}(K-i) = K(K-1)(K-2)\cdots(K-k+1$ for positive integers K and k.

$$P(m|N,M,n) = \binom{M}{m}\frac{(n)_m(N-n)_{M-m}}{(N)_M}$$

(5) Show that in the limit of $N, n \to \infty$ with fixed M, m, and $\left(\frac{n}{N}\right)$, the hypergeometric distribution $P(m|N,M,n)$ approaches the following binomial distribution $P(m|M,p)$. Answer the value of parameter p in this limit.

$$P(m|M,p) = \binom{M}{m}p^m(1-p)^{M-m}$$

Problem 6

You have a newly mapped genome and its genetic map with experimental errors, and try to design an algorithm to make a correspondence between them. You find that solving the following problems is essential for the algorithm. Read the following definitions, and answer the problems.

We call a sequence of distinct real numbers $\{d_0, d_1, \ldots, d_{n-1}\}$ *a marker sequence* of length n. A marker sequence is *consistent* if and only if $d_0 < d_1 < \cdots < d_{n-1}$ or $d_0 > d_1 > \cdots > d_{n-1}$ holds. A marker sequence is *i-inconsistent* if i is the minimum number such that the marker sequence can be consistent by removing i elements $(i < n)$ from it. This i is called *the inconsistent quantity* of the marker sequence.

(1) Find which element should be removed to make a marker sequence $\{0.3, 0.6, 1.3, 0.5, 2.2\}$ consistent.
(2) Suppose that $n \geq 6$. Given a 1-inconsistent marker sequence, prove that at least one of the following four conditions holds. [1] $d_0 < d_1 < d_2$, [2] $d_0 > d_1 > d_2$, [3] $d_{n-3} < d_{n-2} < d_{n-1}$, [4] $d_{n-3} > d_{n-2} > d_{n-1}$
(3) Suppose that $n \geq 6$. Design an algorithm that runs in $O(n)$ worst-case time and finds which element should be removed from a given 1-inconsistent marker sequence to make it consistent.
(4) You are given a marker sequence M. We denote by $F(i)$ the number of the elements in the longest ascending subsequence of sequence $\{d_0, d_1, \ldots, d_{i-1}\}$ such that the sequence ends with d_{i-1} $(1 \leq i \leq n)$. Show that $F(j)$ can be computed in $O(n)$ worst-case time when $F(1), F(2), \ldots, F(j-1)$ $(1 < j < n)$ are already computed.
(5) Design an algorithm that finds the inconsistent quantity of a marker sequence M in $O(n^2)$ worst-case time.

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

Problem 7

For each of the following problems (1)–(10), select the **single best answer** from 1–4.

(1) Which is the **correct** order of intracellular ionic concentrations in typical mammalian cells?

1 $Na^+ > K^+ > Ca^{2+} > Mg^{2+}$
2 $Na^+ > K^+ > Mg^{2+} > Ca^{2+}$
3 $K^+ > Na^+ > Ca^{2+} > Mg^{2+}$
4 $K^+ > Na^+ > Mg^{2+} > Ca^{2+}$

(2) Which is the **correct** order of number of carbon atoms in intermediates of glycolysis and TCA cycle?

1 glucose > α-ketoglutarate > oxaloacetate > pyruvate
2 glucose > oxaloacetate > α-ketoglutarate > pyruvate
3 α-ketoglutarate > glucose > oxaloacetate > pyruvate
4 α-ketoglutarate > glucose > pyruvate > oxaloacetate

(3) Which metabolic pathway is the most **inappropriate** as a mitochondrial one?

1 glycolysis
2 TCA cycle
3 urea cycle
4 β-oxidation of fatty acids

(4) Which monomeric GTP-binding protein is involved in nuclear transport of proteins?

1 Rab
2 Rac
3 Ran
4 Ras

(5) Which is the most **inappropriate** description on intracellular traffic?

1 Signal recognition particle (SRP) recognizes mitochondrial import signal.
2 Clathrin and adaptin participate in the budding of vesicles.
3 SNAREs participate in specific fusion of vesicles.
4 Cell uptakes low-density lipoprotein (LDL) via receptor-mediated endocytosis.

(6) Which is the most **inappropriate** description on signal transduction?

1 Insulin receptor has tyrosine kinase activity.
2 Rhodopsin is an ion channel-coupled receptor for photon.
3 Activation of phospholipase C leads to the production of inositol trisphosphate.
4 Protein kinase C is activated by Ca^{2+} and diacylglycerol.

(7) Which is the **correct** order of diameters of cytoskeleton?

1 actin filament ≥ intermediate filament ≥ microtubule
2 actin filament ≥ microtubule ≥ intermediate filament
3 microtubule ≥ intermediate filament ≥ actin filament
4 microtubule ≥ actin filament ≥ intermediate filament

(8) Which is the **correct** order of the phases of mitosis?

1	G1	S	M	G2
2	G1	S	G2	M
3	G1	M	S	G2
4	G1	M	G2	S

(9) Which is the most **inappropriate** description on tumor suppressor genes and apoptosis?

1 Tumor suppressor gene product Rb inhibits entry into the S phase.
2 Target genes of tumor suppressor gene product p53 include the Cdk inhibitor p21.
3 Bax encoded by a target gene of p53 induces release of cytochrome-c from mitochondria.
4 Apotosis is mediated by a nuclease termed caspase.

(10) Which is the **correct** description on epithelium and intercellular adhesion?

1 Basal lamina underlying the epithelium contains type IV collagen and lamin.
2 Tight junction is composed of cadherin.
3 Adherens junction and desmosome junction are mainly composed of claudin.
4 Hemidesmosomes participate in the attachment of epithelial cells to basal lamina.

Problem 8

(1) Draw a figure of DNA replication process around a DNA replication fork such that it clearly depicts the positional relationships of the following elements A-H. Include labels A-H in the figure to indicate each element.

A. parental DNA double helix
B. leading strand
C. lagging strand
D. Okazaki fragment
E. RNA primer
F. DNA polymerase
G. DNA helicase
H. DNA primase

(2) Suppose that there is a solution that contains the following DNA double helices (a), (b), and (c). If the solution is gradually heated, in what order do they dissociate into single DNA strands? Explain your answer.

(a) CGCATGCGAC
GCGTACGCTG

(b) CGCATGCGACCCTTTAAAAATGTCG
GCGTACGCTGGGAAATTTTTACAGC

(c) CGCGTGCGACCCCGCCGGCGGGTCG
GCGCACGCTGGGGCGGCCGCCCAGC

(3) Each nucleus in human cells contains two copies of the human genome each of which consists of about $3x10^9$ nucleotide pairs. Suppose that the speed of DNA replication at a replication fork is given by 100 nucleotides per second. What is the minimum number of replication origins that a human cell must have in order to finish DNA replication in 12 hours of S phase? Explain your answer.

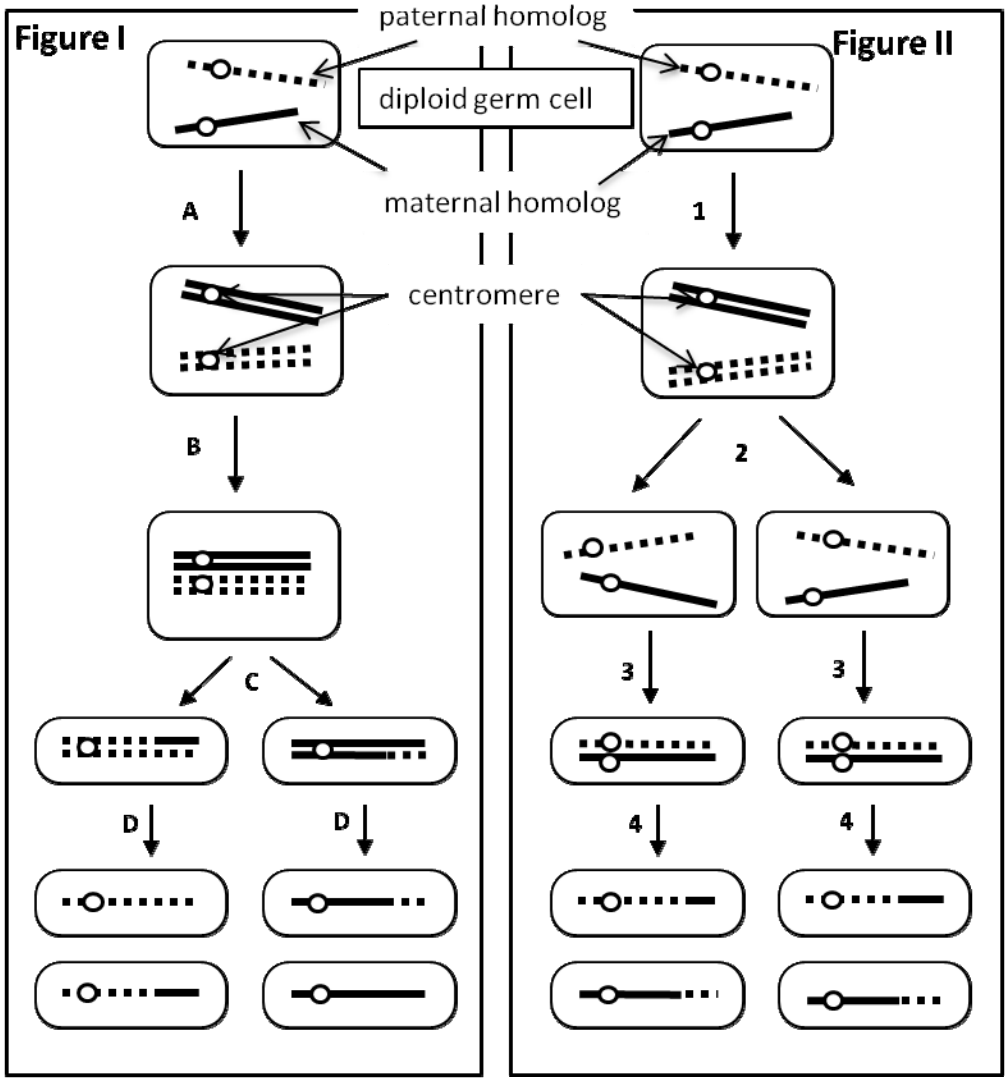
(4) Explain why telomeres are required for the replication of chromosomes in eukaryotes.

Problem 9

DNA sequences accumulate various types of changes during their course of evolution. Because of this property, one can estimate how species have evolved by comparing their DNA sequences. Such evolutionary scenarios are often represented by tree-shaped diagrams that are called (A). According to analyses, it is known that life has evolved in three domains, i.e., eukaryotes, (B), and (C).
Answer the following questions.

(1) Fill (A) in the text.
(2) Fill (B) and (C) in the text (Note: irrespective of order).
(3) For each of the species below, answer which domain of eukaryotes, (B), or (C) it belongs to.
 [I] *Bacillus subtilis*
 [II] *Drosophila melanogaster*
 [III] *Escherichia coli*
 [IV] *Haloferax volcanii*
 [V] *Methanocaldococcus jannaschii*
 [VI] *Saccharomyces cerevisiae*
(4) Assume that a substitution of one base pair (one letter) occurred on a DNA sequence during its course of evolution. How can such a change significantly affect functions of genes? For each of the categories below, answer one of such effects and its mechanism.
 [I] Substitutions in protein coding regions.
 [II] Substitutions in other regions.
(5) Changes on DNA sequences are not limited to those on one base pair. Such larger changes are known to often make it difficult to estimate (A) of biological species from DNA sequences. For each of the categories below, answer one of such changes.
 [I] Changes particularly important in evolution of prokaryotes.
 [II] Changes important in evolution of both prokaryotes and eukaryotes.

Problem 10



On the earth, meiosis in diploid organisms proceeds as illustrated in Figure I. Suppose that there exists a different form of life somewhere in the universe that has selected an alternative pathway of meiosis as illustrated in Figure II in its course of evolution. Answer the following questions by choosing an alphabet letter from Figure I and a number from Figure II for each of (1)-(6) below.

(1) Answer the stage in which homologs are being paired.
(2) Answer the cell division in which ploidy is reduced from diploid to haploid.
(3) Answer the cell division in which sister chromatids separate.
(4) Answer the stage in which breaks in DNA occur.
(5) Answer the cell division in which homologs separate.
(6) Answer the stage in which chromosome replication occurs.

Problem 11

It was found that protein P monomer binds to ligand L monomer and forms complex PL. This complex can dissociate without causing chemical reaction. In this situation, answer the following questions.

(1) Association rate between protein P and ligand L is proportional to the product of protein concentration a_P and ligand concentration a_L. Using association rate constant k_{on}, a_P and a_L, express association rate.

(2) Dissociation rate between protein P and ligand L is proportional to complex concentration a_{PL}. Using dissociation rate constant k_{off} and a_{PL}, express dissociation rate.

(3) In this case, express equilibrium constant K with k_{on} and k_{off}.

(4) Equilibrium constant K is related to the standard free-energy change $\Delta G°$ at 37°C as,

$$K=\exp[-\Delta G°/0.616]$$

where the unit of $\Delta G°$ is kcal/mol. Suppose the standard free-energy change upon the binding of protein P and ligand L is given by ΔG_1. To increase K by a factor of 100, how much should ΔG_1 change? You may use value 2.30 as the natural log of 10 (ln10).

(5) After more detailed investigation, it was found that protein P does not form a stable structure in the absence of ligand L. Protein P can bind to ligand L only when protein P transiently forms a specific structure P'. Therefore, the binding process of ligand L can be considered to consist of two steps, conformational change to structure P' and ligand binding. The standard free-energy change upon the former, ΔG_2, is positive at 37°C. Explain how the standard free-energy change ΔG_3 upon the latter is related to ΔG_1 and ΔG_2.

(6) Calculate the value of ΔG_3 when $K = 10^4$ and $\Delta G_2 = 2.0$ kcal/mol at 37°C.

Problem 12

There are various types of repetitive sequences in genomes. Answer the following questions about repetitive sequences.

(1) Among repetitive sequences of eukaryotes, there are repetitive sequences that transpose via RNA intermediate. What are the repetitive sequences via RNA intermediate generally called?

(2) The interspersed repetitive sequences mentioned in (1) are evolutionarily related to retrovirus. Among them, there is a repetitive sequence whose full-length is about 6Kb and which accounts for about 15% of the human genome. What is the proper name of this repetitive sequence? Give an example of proteins that this repetitive sequence and retrovirus have in common.

(3) The human genome contains simple sequence repeats which are formed of tandem repetitions of two to tens nucleotides such as CA. These simple repeats can be used for individual identification and linkage analysis of genetic disorders. Explain the reason.

(4) The genome analysis of a fish published in 2002 strongly suggested that a large portion of repetitive sequences might be 'junk DNA'. The genome size of the fish is about 1/8 of that of the human genome. What is the fish analyzed? What structural features of the genome and genes of the fish supported that repetitive sequence might be 'junk DNA'?

(5) Figure 1 shows the relation between the base substitution rate (horizontal axis in %) of four interspersed repetitive sequences (Rep1～Rep4) and the fraction of each repetitive sequence (vertical axis in %) in a mammalian genome. Here, the base substitution rate of a repetitive sequence element is defined as the sequence divergence from its consensus sequence. Answer the following questions, (I) and (II).

(I) Select <u>two incorrect</u> descriptions from the following five descriptions (A to E) about Figure 1.

(A) Rep1 has been explosively amplified at one time in the past.
(B) Rep4 has been actively amplified relatively recently.
(C) Rep3 is a repetitive sequence generated more recently than Rep1
(D) Most of repetitive sequences have lost the ability to transpose due to accumulation of mutations.
(E) The latest amplification of these repetitive sequences is decreasing dramatically.

(II) Assuming that the base substitution rate of this genome is 3×10^{-9} / year・base, when were

the repetitive sequences with 15% base substitution rate in Figure 1 amplified?

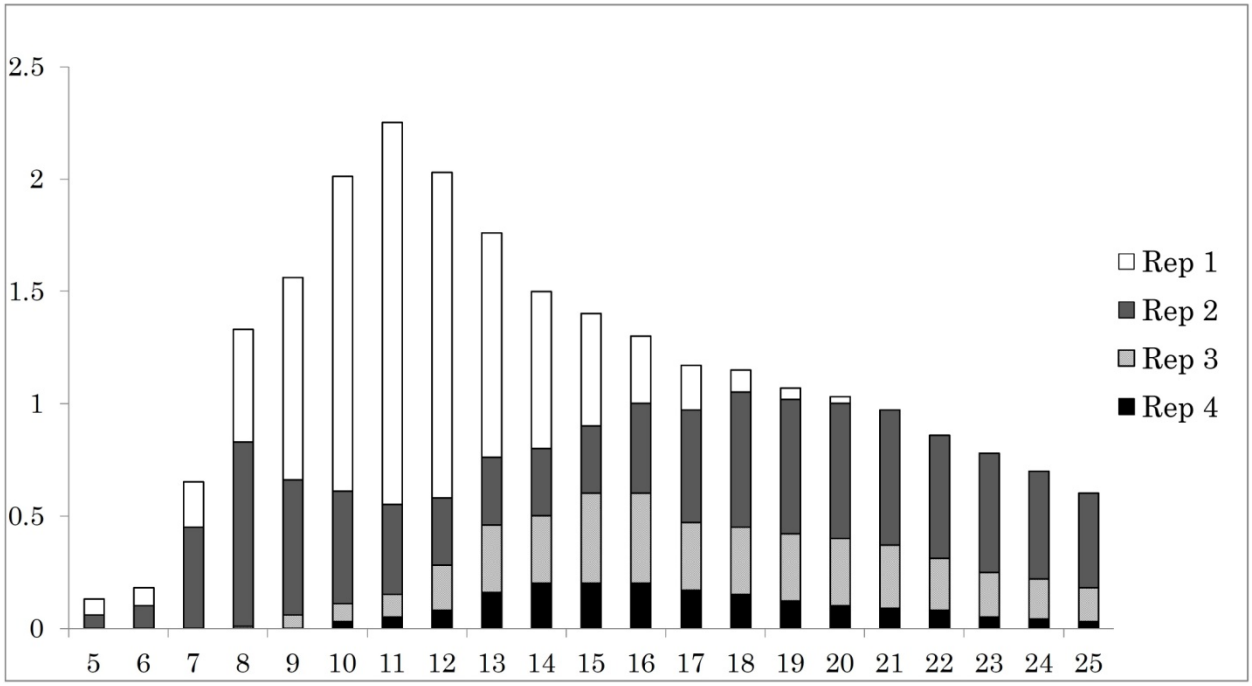


Figure 1

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