

Introduction to Computational Biology
Homework 1
09/09/09

Due 09/23/08

Problem 1: Bad recombination model

Assume that we have two recombination positions (as opposed to just one as we saw in class) that occur independently and uniformly at random along the chromosome. Show that the probability of recombination between two close genes could be the same as the probability of recombination between two distant genes.

Problem 2: Properties of the one recombination position model

Recall the one recombination position model we saw in class: Only one of the $n + 1$ recombination positions occurs and it occurs uniformly at random along the chromosome. For each of the following biological properties, specify whether the one recombination position model satisfies the property.

- Mendel's First Law (there is a 50% change for a gene to come from either chromosomes)
- The probability of recombination between two genes is higher for more distant genes.
- Very distant genes act independently, i.e, the probability that a recombination occurs between two very distant genes is equal to $p_1 \cdot q_2 + p_2 \cdot q_1$, where p_i is the probability of the first gene coming from chromosome i , and q_i is defined similarly for the second gene. What is the above quantity equal to?

Problem 3: The Jumping model of recombination

Consider the following *jumping* model: At each position on the chromosome, there is a probability p (the jumping parameter) of crossing over (jumping) to the other chromosome (and hence a probability $1 - p$ of staying on the same chromosome). In other terms, this model assumes that the frequency of recombination is uniform along the chromosome (although in reality some sites are hot spots or cold spots for recombination).

- (a) What is the probability that a given gene comes from chromosome 1 and how does it depend on the jumping parameter p ? Explain your answer.
- (b) Derive an expression for the probability of recombination (or a way to compute it) between two genes at a distance d from each others as a function of d and the jumping parameter p .
- (c) Plot the expression you obtain in (b) for $1 \leq d \leq 100$ and for different values of the jumping parameter p .
- (d) Relying on part (c), what values of the jumping parameter p satisfy the three biological properties listed in Problem 2?

Problem 4: Genetic mapping

Consider using three RFLP markers A, B, and C to identify the different alleles (copies of genes) on the chromosomes. Each allele is represented by a single digit (for instance, it could be the number of fragments obtained in the RFLP marker). The alleles for the two homologous chromosomes for parents and offspring are listed in the following table:

	A	B	C
Father	1,5	2,3	6,8
Mother	1,9	4,7	3,6
Offspring 1	1,5	2,7	6,6
Offspring 2	1,9	3,4	3,8
Offspring 3	1,5	2,7	6,8
Offspring 4	5,9	3,4	3,6
Offspring 5	1,9	3,7	3,8
Offspring 6	5,9	2,4	6,6
Offspring 7	1,1	3,4	6,8
Offspring 8	1,5	2,7	6,6
Offspring 9	5,9	2,4	3,6
Offspring 10	1,9	3,4	6,8

- (a) For each of the offspring, list which alleles were inherited from the father and which from the mother.
- (b) For each pair of markers, count the number of recombinations (both maternal and paternal) that occurred between just those markers (e.g. for markers A and C, ignoring the data for marker B entirely, how many recombinations do you see?).

Use this information to build a map of the markers. (Hint: When you obtain the counts think of the three biological properties listed in Problem 2.)

Problem 5: Physical Mapping

Consider a circular DNA that is 2500 base pairs long. You wish to construct a restriction map for this DNA. You treat it with a set of restriction enzymes and you measure the resulting fragment lengths by Gel-electrophoresis to obtain the following results:

EcoRI	2500
HindIII	2500
PstI	2500
MboI	1300, 800, 400
MboI + EcoRI	1300, 600, 400, 200
MboI + HindIII	1300, 800, 300, 100
MboI + PstI	1000, 800, 400, 300
EcoRI + HindIII	2000, 500
EcoRI + PstI	1600, 900
HindIII + PstI	2100, 400

Construct a restriction map based on the above information. To break the circularity place base pair 1 at the HindIII cleavage site.

Problem 6: Shortest Covering String

Recall the shortest covering string problem we described in class. In a hybridization mapping experiment, the goal is to find a shortest string over the alphabet of probes that covers all the clones. A string S is said to cover a clone C if S has a substring that contains the exact set of probes in C (order and multiplicity are ignored).

Example:

$$C_1 : \{A, B\}, C_2 : \{A, C\}$$

The string $ABAC$ is a covering string for C_1 and C_2 . However, this string is not the shortest possible. Since the order of probes is not important, BAC , for instance, is also a covering string. In BAC the substring BA contains the probes $\{A, B\}$ of C_1 and the substring AC contains the probes $\{A, C\}$ of C_2 . In BAC both C_1 and C_2 are covered by substrings (BA and AC respectively) that do not contain probe repetitions.

Construct an example where, in the shortest covering string, one clone must be covered by a substring that contains a probe repetition.

Problem 7: Shortest Superstring

Construct a shortest superstring for all the binary strings of length 4, i.e. 0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111.