

Introduction to Computational Biology
Homework 1
Solution

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.

Solution:

We have two recombination points that occur independently and uniformly at random across the chromosome. We need to show that two close genes can have the same probability of recombination as two distant genes.

Assume we have n genes on the chromosome. Therefore, each recombination point can take one of $n + 1$ positions (as described in class). Denote by x_1 and x_2 the two recombination points. Two genes will recombine if exactly one recombination point occurs between the two genes (inside). Since the two recombination points are independent, this probability can be expressed as follows:

$$Pr[x_1 \text{ inside}].Pr[x_2 \text{ outside}] + Pr[x_1 \text{ outside}].Pr[x_2 \text{ inside}]$$

Assume the two genes are at a distance d from each other. Since a recombination point occurs uniformly at random across the chromosome, $Pr[x \text{ inside}] = \frac{d}{n+1}$. Therefore, the probability that two genes recombine is:

$$\frac{d}{n+1}\left(1 - \frac{d}{n+1}\right) + \left(1 - \frac{d}{n+1}\right) \cdot \frac{d}{n+1} = \frac{2}{(n+1)^2}d(n+1-d)$$

Replacing d in the above expression by $(n+1-d)$ yields the same expression. Therefore, two genes at a distance d and two genes at a distance $(n+1-d)$ have the same probability of recombination. If d is small, then $(n+1-d)$

is large, and viceversa. This proves that, under this model, two close genes can have the same probability of recombination as two distant genes.

Intuitively, two genes that are very close have a small probability of recombination because one recombination point has to be inside. But on the other hand, two distant genes also have a small probability of recombination because one recombination point has to be outside.

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?

Solution:

Property 1: Mendel's First Law: there is a 50% change for a gene to come from either chromosomes.

This property is satisfied assuming that the recombination process is equally likely to start with any of the two chromosomes. Therefore, $Pr[start X_1] = Pr[start X_2] = \frac{1}{2}$. The probability that the i^{th} gene comes from one chromosome given that we start with the other $Pr[from X_1 | start X_2] = Pr[from X_2 | start X_1] = \frac{i}{n+1}$. This is because the recombination point has to occur before the gene for this to happen. Therefore, the probability that the i^{th} gene comes from the first chromosome is:

$$\begin{aligned} &Pr[\text{from } X_1|\text{start } X_2].Pr[\text{start } X_2] + Pr[\text{from } X_1|\text{start } X_1].Pr[\text{start } X_1] = \\ &Pr[\text{from } X_1|\text{start } X_2].Pr[\text{start } X_2] + (1 - Pr[\text{from } X_2|\text{start } X_1]).Pr[\text{start } X_1] = \\ &\frac{i}{n+1}\frac{1}{2} + (1 - \frac{i}{n+1})\frac{1}{2} = \frac{1}{2} \end{aligned}$$

Property 2: The probability of recombination between two genes is higher for more distant genes.

This property is also satisfied. As we have seen in class, two genes at a distance d have a probability $\frac{d}{n+1}$ of recombination. Therefore, the probability of recombination between two genes is higher for distant genes.

Property 3: 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?

From the first property, $p_1 = p_2 = q_1 = q_2 = \frac{1}{2}$. So if two distant genes act independently, the probability of recombination between them should be $p_1 \cdot q_2 + p_2 \cdot q_1 = \frac{1}{2}$. However, from the second property above, two genes at the extremities, i.e. at distance $n - 1$, have a probability of recombination $\frac{n-1}{n+1} \rightarrow 1$. Therefore, this property is not satisfied.

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.

Solution: We will assume that the recombination process is equally likely to start with any chromosome, $Pr[\text{start } X_1] = Pr[\text{start } X_2] = \frac{1}{2}$. Under this assumption, the recombination process is very symmetric. Consider a

path (a pattern of jumps between the two chromosomes) by which a gene comes from chromosome 1. This path has a certain probability. There is a symmetric path with the same probability by which the same gene comes from chromosome 2.

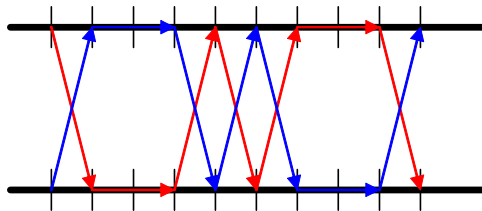


Figure 1: symmetric paths

Therefore, for every possible way of obtaining a gene from chromosome 1, there is a symmetric way of obtaining the same gene from chromosome 2. Hence, the probability that a gene comes from either chromosomes is $\frac{1}{2}$. This is independent from the jumping parameter p .

(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 .

Solution: Two genes at a distance d will recombine if there is an odd number of jumps between them. Therefore, the probability of recombination between two genes at distance d is

$$p_d = \sum_{k \text{ odd}} \binom{d}{k} p^k (1-p)^{d-k}$$

Another way to look at it is by regarding the jumping process as a Markov chain with two states X_1 and X_2 and the following transitional probability matrix

$$P = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

Then the probability of being in state X_2 after d steps given that we start at X_1 is P_{12}^d . Similarly, the probability of being in state X_1 after d steps given

that we start at X_2 is P_{21}^d . The probability of starting at X_1 is $\frac{1}{2}$ (see part (a), it is the same as the probability of a gene coming from chromosome 1). Therefore, $p_d = \frac{1}{2}P_{12}^d + \frac{1}{2}P_{21}^d = P_{12}^d$, because P is symmetric ($P_{12} = P_{21}$).

(c) Plot the expression you obtain in (b) for $1 \leq d \leq 100$ and for different values of the jumping parameter p .

Solution: We can plot P_{12}^d as function of d and p . Here are some results.

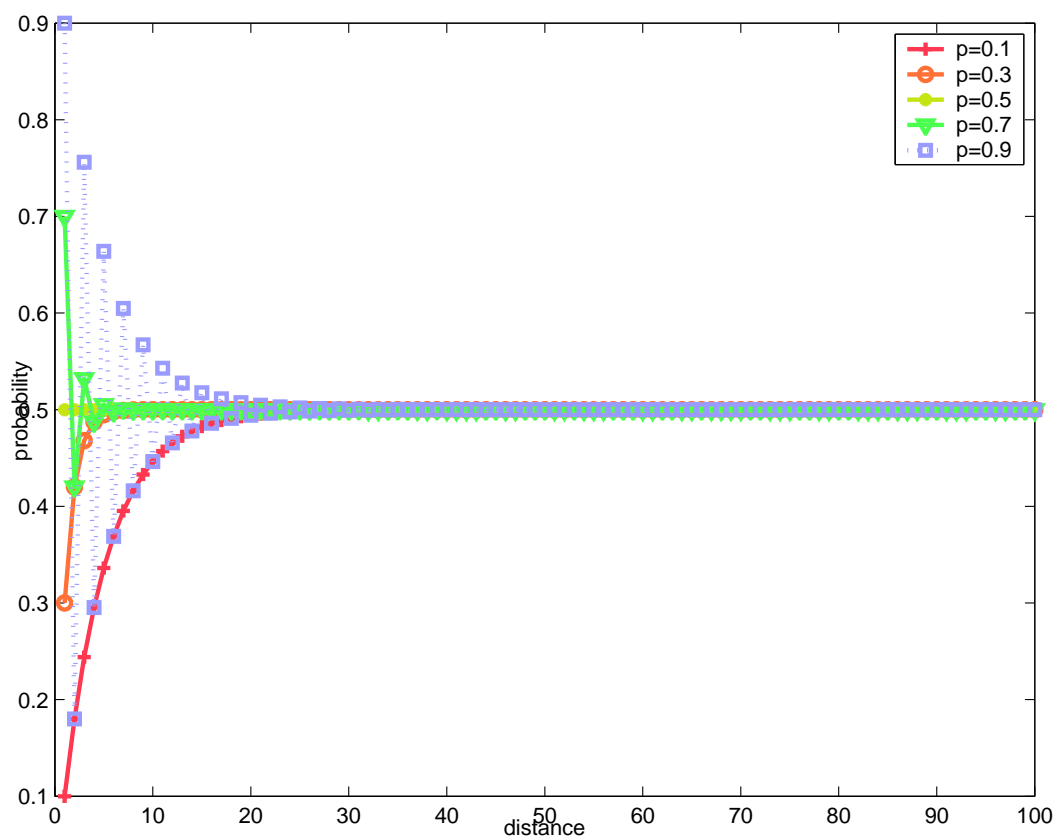


Figure 2: Probability of recombination between two genes at distance d

(d) Relying on part (c), what values of the jumping parameter p satisfy the three biological properties listed in Problem 2?

Solution: We can see that for $p < \frac{1}{2}$, all three properties will be satisfied since the first property is satisfied regardless of what p is (see part (a)), and the probability of recombination increases with d until it hits $\frac{1}{2}$ for large d where the genes act as if they are independent.

Theoretically speaking, for $p < 1$, we expect that the probability of recombination P_{12}^d will converge to $\frac{1}{2}$. This is a property of Markov chains, where P^d converges to the steady state probabilities of the states X_1 and X_2 , $\lim_{d \rightarrow \infty} P^d = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$.

The convergence will not occur when $p = 1$, because in this case we will have a periodic Markov chain with period 2 (a state can only be revisited at even intervals of steps). Intuitively, $p = 1$ means that we always jump, so the probability of recombination is 0 for even d and 1 for odd d (no convergence).

It would be interesting to explain why $p < \frac{1}{2}$ satisfies the second property (i.e. probability of recombination increases with distance), whereas $p \geq \frac{1}{2}$ does not (oscillates). The key is in the evaluation of the expression:

$$p_d = \sum_{k \text{ odd}} \binom{d}{k} p^k (1-p)^{d-k}$$

Instead of explicitly evaluating the above expression, we can express p_d in terms of p_{d-1} . Note that

$$p_d = p(1 - p_{d-1}) + (1 - p)p_{d-1}$$

Therefore, $p_d = p + (1 - 2p)p_{d-1}$. From this recurrence, and the fact that $p_1 = p$, we can compute

$$p_d = \frac{1 - (1 - 2p)^d}{2}$$

When $p < 1/2$, $(1 - 2p)$ is positive and, therefore, p_d increases to eventually $1/2$. When $p > 1/2$ (but less than 1), $(1 - 2p)$ is negative and, therefore, $(1 - 2p)^d$ alternates between positive and negative depending of whether d is even or odd respectively, but p_d will also converge to $1/2$.

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.

Solution: Each child will take one allele from the father and one allele from the mother for each of the markers A, B, and C.

	A		B		C	
	father	mother	father	mother	father	mother
1	5	1	2	7	6	6
2	1	9	3	4	8	3
3	5	1	2	7	8	6
4	5	9	3	4	6	3
5	1	9	3	7	8	3
6	5	9	2	4	6	6
7	1	1	3	4	8	6
8	5	1	2	7	6	6
9	5	9	2	4	6	3
10	1	9	3	4	8	6

(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.)

Solution: Assume that the father and mother homologous chromosomes look like the following:

father:

--1--2--6--
 --5--3--8--

mother:

--1--4--3--
 --9--7--6--

Then we count 17 recombinations for *AB*, 17 recombinations for *AC*, and 6 recombinations for *BC*. This makes the probability of recombination for *AC* 17/20, for *AB* 17/20, and for *BC* 6/20. For example, the 6 recombinations for *BC* are shown in bold (for father) and italic (for mother) in the table above.

According to the biological properties, a probability of 17/20 is not reasonable since the probability increases with distance and two distant genes have a 50% probability of recombination. Therefore, the probability for recombination cannot exceed 0.5.

The problem is our assumption about which alleles occur on which of the two homologous chromosomes. Clearly if we switch the alleles for marker *A*, then we fix the problem. So the chromosomes should look like the following:

father:

--5--2--6--
 --1--3--8--

mother:

--9--4--3--
 --1--7--6--

Every recombination that we counted for AB and AC is now not a recombination. So the number of recombinations for these two markers will be 3 for each, making the probability $3/20$, a reasonable one.

Therefore, we conclude that the order of the markers must be:

--B--A--C--

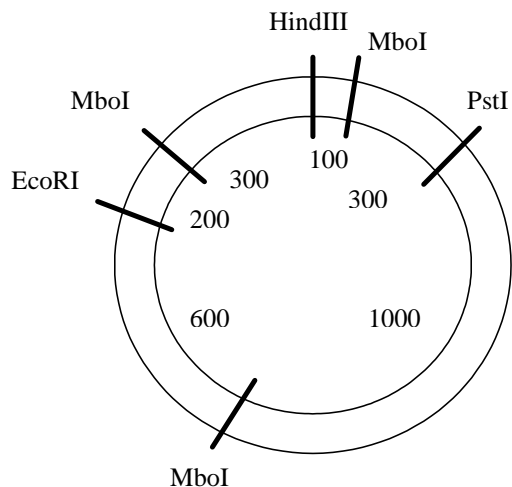
Problem 5: Physical Mapping

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.

Solution:



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.

Solution: Consider the following instance of the problem:

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

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

$C_3 : \{A, D\}$
 $C_4 : \{A, E\}$
 $C_5 : \{A, B, C, D, E\}$

The thing to note about this instance is that a shortest covering string for C_1 , C_2 , C_3 , and C_4 is also a covering string for C_5 . Therefore, we will focus on the shortest covering string for the first four probes only.

The shortest covering string must have at least one occurrence of each of B , C , D , and E .

Moreover, the shortest covering string must have two occurrences of A that are not the start or the end of the string. Proof: By contradiction. Assume that the shortest covering string has at most one A that is not the start or end of the string. Therefore, the shortest covering string has at most three A s, two of them are at the extremities of the string. We need to cover $C_1 : \{A, B\}$, $C_2 : \{A, C\}$, $C_3 : \{A, D\}$, and $C_4 : \{A, E\}$. Therefore, the middle A must cover two clones, and each A at one extremity must cover one clone. Without loss of generality, the shortest covering string looks like $AB...CAD...EA$. But then we can obtain a shorter covering string by dropping the last A and moving E to the beginning: $EAB...CAD....$

Therefore, the shortest covering string must have a length of at least 6 with at least two A s being not at the start or end of the string. As a result, any shortest covering string must have the form $-A--A-$, where the four $-$ are filled arbitrarily with B , C , D , and E , because this form has exactly length 6.

Therefore, C_5 will be covered by a substring (the whole string in this case) that contains a repetition of A .

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.

Solution: A possible shortest superstring of length 19: 0100001110110010101. The length 19 is also a lower bound on any superstring for this problem, because the best thing we can do is start with one string of 4 characters, and then add 1 character for each additional string for the remaining 15 strings.

In general, for any k , the lower bound is $k + 2^k - 1$. But the question is: Can we

always construct a superstring of length $k + 2^k - 1$ for all binary strings of length k ? The answer is yes... More on this later in the course.