Introduction to Computational Biology
Homework 1
2/5/2014

Due 2/12/2014

Problem 1: Biological properties of recombination
We have discussed the uniform one point recombination model in class and argued that it satisfies only the first two of the following properties:

- Mendel's first law: there is a 50% chance for a gene to come from either chromosomes.
- The probability of recombination is higher for distant genes.
- Mendel's second law: genes are asymptotically independent i.e. the probability that a recombination occurs between two genes at a large distance 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. That's a $1/2$ according to the first property.

Let's try to satisfy all three. Consider the following model: At each position $1 \dots n$ along the chromosome, there is a probability p of crossing over 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 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 as a function of d and p.

(c) What values of p satisfy the three biological properties listed above?

Problem 2: DNA coverage
Many DNA sequencing techniques rely on cutting the DNA into many overlapping fragments. This exercise should help you understand how many fragments are needed. Assume we have n fragments of length l each and a DNA sample of total length T. Assume further that the position of a fragments along the DNA is uniformly random.

(a) Show that the expected number of bases that are **not** covered by fragments is approximately $Te^{-nl/T}$. *Hint*: Let X_i be an indicator random variable that base i is not covered, i.e. $X_i = 1$ iff base i is not covered. Compute $E[X_i]$ and then use linearity of expectation.

(b) Let $n = \alpha T/l$. How big should α be?

(c) Given a fragment, what is expected number of overlapping fragments? What is the probability distribution for that number (this is needed to detect a possible repeat).

Problem 3: Shortest Covering String
Recall the shortest covering string problem we described in class. The goal is to find a shortest string over the alphabet of probes that covers all the fragments. A string S is said to cover a fragment f if S has a substring that contains the exact set of probes in f (order and multiplicity are ignored).

Example:

$f_1 : \{A, B\}$, $f_2 : \{A, C\}$

The string $ABAC$ is a covering string for f_1 and f_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 f_1 and the substring AC contains the probes $\{A, C\}$ of f_2. In BAC both f_1 and f_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 fragment must be covered by a substring that contains a probe repetition. This is not trivial!

Problem 4: 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. *Hint*: you will know it is shortest if, starting with some pattern of 4 bits, it requires only 1 bit to cover an additional pattern.

Problem 5: Circular DNA alignment
Consider two circular DNAs x and y of length m and n respectively. We are after the optimal global alignment of x and y. This can be obtained as follows: Consider a circular shift of x, $x_i...x_m x_1...x_{i-1}$ for some $1 \leq i \leq m$. Consider a circular shift of y, $y_j...y_n y_1...y_{j-1}$ for some $1 \leq j \leq n$. Find their optimal global alignment, and repeat for every possible pair of circular shifts of x and y. Finally pick the highest scoring alignment. Since there are m circular shifts of x and n circular shifts of y, the above algorithm will take $O(m^2n^2)$ time.

(a) Design an $O(mn^2)$ time algorithm that will find the optimal global alignment of two circular DNAs x and y.

(b) [optional] Can you obtain the optimal global alignments for all pairs of circular shifts (i.e. mn) of x and y in $O(mn^2 + nm^2)$?