Computational Biology
Lecture 2: Some problems in biology
Saad Mneimneh

We have seen in the last lecture the problem of genetic mapping which is to find the order of genes on the chromosome (not their exact location). In this lecture, we review the problem and look at more advanced techniques for locating genes in the genome.

Genetic Mapping

Specific genes, or specific gene mutations, are frequently the cause of diseases. In order to help us identify when a disease is associated with the presence of a particular gene, it is useful to have a map of the genome.

Genetic mapping is the process of determining the relative location of genes in a particular genome. By mapping the relative location of several genes, we can develop a map of the entire genome of a particular species e.g. humans.

Genetic mapping relies on externally observable genetic traits, or phenotypes, of a particular genome. By selectively breeding specific phenotypes together and examining the phenotypes of several generations of children, we can determine the frequency of the occurrence of particular phenotypes. After a large number of observations, we have a reasonable approximation of the probability of the occurrence of those particular phenotypes. Based on these probabilities, we can get an estimate of the distance between specific genes. As an example, let's take a genome that has n genes on a single chromosome and that recombination occurs randomly at only one point on the chromosome. Therefore, if the father's chromosome is $f_1...f_n$ and the mother's chromosome is $m_1...m_n$, a child can have either an $f_1...f_i m_{i+1}...m_n$ or an $m_1...m_i f_{i+1}...f_n$ chromosome, for some recombination position $0 \leq i \leq n$. As a result, every pair of parents can have $2(n+1)$ possible kinds of children (some might be identical) based on this single recombination position model. The probability of this recombination occurring at a particular position is $p = \frac{1}{n+1}$. The probability of two genes being separated by recombination is the probability that the recombination occurs in any position between them. This probability is expressed as $p = \frac{d}{n+1}$, where n is the number of genes in the genome and d is the distance between the two genes. Note that closer genes will have less chance of recombination (this is where the second law of Mendel is wrong, genes are not inherited independently if they are on the same chromosome). By starting with two different pure breed parents, say black and blue, we can concentrate on the states of two genes and observe the frequency of the states being different in the children (i.e. the frequency of recombination between those two genes). This will help us estimate p and therefore d. If we are able to determine the distance between all pairs of genes in our example genome, then we can use these distances to determine the exact sequence of the genes. However, real life is not as simple as our example.

In general, recombination occurs at an arbitrary number of positions in the genome. There are not always distinct phenotypes for a particular gene, or gene combination. This means that single changes cannot always be observed and in fact usually it takes multiple changes before there are any changes in phenotype. Moreover, some differences in the genotype are not reflected in any phenotype. It is highly probable that separate genes are distant and randomly distributed among multiple chromosomes. This makes our job even more difficult. On the other hand, if genes are particularly close, the order of the genes can not be determined because the two genes almost never occur separately.

Restriction Fragment Length Polymorphism RFLP

RFLP allows for the observation of genetic changes that are not expressed by differences in the phenotype. The DNA is broken into smaller fragments (called restriction fragments) by a restriction enzyme, like HindIII for instance. The enzyme cuts the DNA at a particular pattern (property of the enzyme) called restriction site. For instance, HindIII has the restriction site AAGCTT and cuts the DNA as follows: A | AGCTT. Note that the complementary strand of the DNA will have TTCGAA. This is the reverse of the pattern, so HindIII will also cut the second strand as follows: TTCGA | A.

The main idea behind using a restriction enzyme is that mutations on the DNA might create or destroy a restriction site. Hence the restriction fragment lengths obtained are going to be different. By detecting the change in the lengths of those restriction fragments, we can detect certain mutations. This is where the name comes from: Restriction Fragment Length Polymorphism. But how do we measure the fragment lengths? By a process called Gel-Electrophoresis.

Gel-Electrophoresis

The main idea behind Gel-electrophoresis is as follows. After we cut the DNA into restriction fragments, the cut DNA is put on a gel material (agarose for instance). An electric current is applied through the Gel. The DNA is negatively charged, so it will start moving on the gel towards the positively charged end. Smaller fragments move faster on the gel, so after some time we have a separation of the different lengths on the gel. It is very difficult to read all fragment length on the gel at once, because there is a large number of fragments with many possible lengths, so we almost obtain a contiguous smear on the gel.

Hybridization: In a hybridization experiment we try to verify whether a specific sequence known as *probe* binds (hybridizes) with a DNA fragment. If the binding occurs, this means that the DNA fragment contains the sequence complementary to the probe sequence (or parts of it).

After the separation of the different fragments on the gel, we apply a number of a selected probes. Each probe is mixed with a radioactive material and applied in turn on the gel. Each probe was designed to hybridize with a certain portion (sequence) of the original DNA. Therefore, after the cutting, each probe will hybridize with the different fragments that belong to (came from) the corresponding portion of the DNA. Since the probe is radioactive, we will see bands on the gel that correspond to the positions of the different fragments hybridizing with the probe. The length of these fragments could be obtained by looking at how much they traveled on the gel. We associate these lengths with that particular probe. Each combination of a probe and its' hybridized fragments lengths is called a RFLP marker.

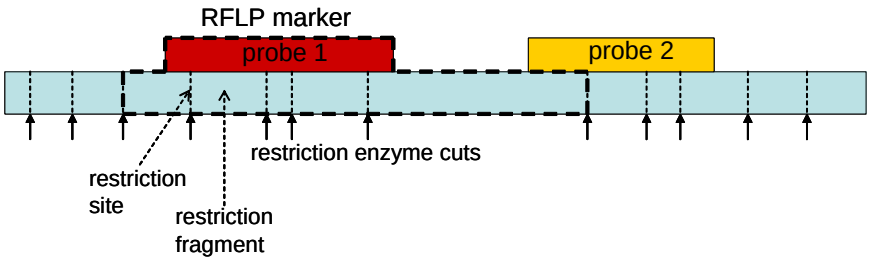

Figure 1: RFLP marker

If the following are true:

- the probe set is large
- each probe is long enough
- the probes are randomly selected to hybridize with different portions of the DNA

then the probes will span most of the DNA and the RFLP markers will capture most of the restriction fragment lengths obtained by the cutting enzyme.

Now look at these RFLP markers as an alternative to genes. Like genes, each RFLP marker has a position on the DNA (where its probe hybridizes) and has a state (the set of lengths). The gained advantage is that any change in the state of an RFLP marker can be detected experimentally (as opposed to the phenotypes). So a recombination analysis can be done on the RFLP marker to map them on the chromosome. Of course this does not tell us much about the genes themselves, but it can help locating genes. For instance, if a specific RFLP was found to be associated with a certain disease, then we know that this RFLP marker intersects with the gene responsible for that disease. Since we have a map of the RFLP markers, we have an approximate location for the gene. This is how the cystic fibrosis gene was associated with a particular marker on chromosome 7.

Physical Mapping

Genetic mapping and RFLP do not tell the actual distances. Furthermore, if genes (or RFLP markers) are very close, one cannot resolve their order because the observed recombination frequency will be zero. Physical mapping reflects actual distances. It deals with DNA fragments of known lengths and attempts to order them.

Hybridization Mapping

Several copies of the DNA are cut into fragments using different restriction enzymes. Then multiple copies of each fragment are obtained (by cloning for instance). Each copy is called a clone, thus forming a clone library. Clones may overlap because the original copies of the DNA were cut with different enzymes. The problem of hybridization mapping is to rely on this overlap information to reconstruct the order of the clones (why are we doing this? are we destructing the DNA just to reconstruct it again? Yes, because we cannot deal with very long DNA molecules. It is easier to obtain information about small fragments. Then the problem becomes to reconstruct their order of course). But how do we obtain the overlap information? We follow the concept of *fingerprinting*. We will give each clone a fingerprint, then if two clones have similar fingerprints, we will have a reason to believe that they overlap. The hybridization mapping will use a set of probes (short ones this time) and for each clone obtain the set of probes that hybridize with it. If two clones have similar sets of probes, this means that they probably overlap on the original DNA. Of course, short probes

will tend to hybridize with the original DNA at many positions. Therefore, the set of probes is not an exact measure of overlap. In fact, two distant fragments could hybridize with the same probe if the complementary sequence of the probe occurs multiple times on the DNA. Nevertheless, it is a starting point to figure out a possible order of the fragments.

Here's a simple theoretical abstraction to the problem.

Shortest Covering String: For n clones and m probes, the hybridization data consists of an $n \times m$ matrix D where $d_{ij} = 1$ if clone C_i contains probe p_j. A string S over the alphabet of probes is said to cover a clone C if there exists a substring of S containing exactly the same set of probes as C (order and multiplicity ignored). Find the shortest covering string.

In general, the shortest covering string problem is NP-hard, but if the probes are unique (i.e. each probe hybridizes only once with the original DNA), then it admits a polynomial time algorithm. Of course, practically speaking, producing unique probes is not easy.

Restriction Mapping

This problem is similar to the above, except that a different fingerprinting technique is used. Here the fingerprint is the ordered set of length obtained when cutting a clone with a specific enzyme. For each clone, it is easy to obtain the set of lengths produced by the cut (using gel-electrophoresis) but it is much harder to determine their order. In practice, the clone is cut with multiple enzymes, for instance, once with enzyme A, once with enzyme B, and once with both simultaneously. We have to obtain a physical map of the fragments lengths based on the information obtained in all of the three cases. This is called the double digest problem.

Double Digest: Given the multiset of lengths L_A obtained by enzyme A and the multiset of lengths L_B obtained by enzyme B and the multiset of lengths L_{A+B} obtained by both A+B, find an ordering of all sets that is consistent.

As an example, if $L_A = \{2, 2, 3\}$ and $L_B = \{3, 4\}$ and $L_{A+B} = \{1, 2, 2, 2\}$, then a consistent ordering will be 2, 3, 2 for L_A and 3, 4 for L_B and 2, 1, 2, 2 for L_{A+B} (try it).

The double digest problem in NP-complete. Moreover, the number of solutions grows exponentially with the size of the problem.

Here's another variation called partial digest.

Partial Digest: Given the set of lengths of all fragments that can possibly be obtained by any two cuts using a single enzyme, determine the position of all restriction sites. (It is called partial digest because not all cuts are produced simultaneously, otherwise, we could not obtain a fragment for any two possible cuts). This is equivalent to the turnpike problem in computer science, where for a given highway with n exits, we have information about the distance between any pair of exits, and we are asked to reconstruct the positions of exits on the highway. We can substitute the highway for the real line and a distance between exits by a distance between two points on the line. Then the question is to reconstruct the n points on the line given the C_2^n distances between every pair of points. It is an open question whether this can be done efficiently, and it is not known whether the problem is NP-complete or not.

References

Pevzner P., Computational Molecular Biology, Chapter 1.

Setubal J., Meidanis, J., Introduction to Molecular Biology, Chapter 1.