Computational Biology

Lecture 2
Genetic mapping

- Single chromosome with \( n \) genes
- Single recombination point that occurs uniformly at random
- Probability of recombination between two genes at distance \( d \) is \( p = d/(n+1) \)
- Estimate \( p \) (and therefore \( d \)) by observing the frequency of different phenotypes

- Problems
  - Too many chromosomes
  - Not all genes have phenotypes that can be observed
  - We usually don’t know what we are looking for
RFLP: Restriction Fragment Length Polymorphism

- A restriction enzyme cuts the DNA molecules at every occurrence of a particular sequence, called restriction site.

- For example, HindII enzyme cuts at GTGCAC or GTTAAC

- If we apply a restriction enzyme on DNA, it is cut at every occurrence of the restriction site into a million restriction fragments, each a few thousands nucleotides long.

- Any mutation of a single nucleotide may destroy or create the site (CTGCAC or CTTAAC for HindII) and alter the length of the corresponding fragment.

- RFLP analysis is the detection of the change in the length of the restriction fragments.
Gel-Electrophoresis

- DNA is cut into fragments using an enzyme
- The cut DNA is put on a Gel material
- An electric current is applied on the Gel
- DNA is negatively charge
- DNA fragments will start moving towards the positively charged side
- Smaller fragments move faster
- After some time, we have a separation of the different fragment lengths
DNA Sample

- Some cells are obtained
- The cells are immersed in a nutritious solution on a plate and left to grow and multiply
- The cells are gathered and frozen for future use
- Liquidized DNA is obtained from these cells
Restriction Enzyme

- A restriction enzyme is used to cut the DNA into fragments

- Hind III restriction site is A∥AGCTT
Apply Enzyme

- DNA sample and Hind III are put together in a tube
- The tube is shaken by rotation for DNA and Hind III to mix
Water Bath

- The tube is put on a plate floating on water at 37°C
- It is left for 30 minutes
- This is needed for the Hind III reaction to take place
Preparing the Gel

- In the meantime, we prepare the Gel

- Agarose powder is the basic substance for making the Gel
Preparing the Gel

• The powder is mixed with water in a container
Preparing the Gel

• The container is heated (in a microwave if you want) until the powder completely dissolves in the water

• The solution becomes clear
Preparing the Gel

- The liquid Gel is poured into the inner box.

- A comb like piece is put at the edge of the inner box.

- The liquid Gel is left to cool and solidify (you can use a fridge).

- When the Gel solidifies, the comb will create wells for the DNA samples to be put.
Gel Ready

- Gel ready
- Fill the H shaped container with water
- Remove comb
Putting DNA on the Gel

- DNA samples mixed with colored solution and UV reactive solution
- DNA samples inserted into wells
- A sample DNA containing only specific fragments (called ladder) can be used for comparison

original uncet DNA
DNA cut by Hind III
ladder 1
ladder 2
Run the Gel

• Apply electric current

• DNA is negatively charged

• Fragments will migrate toward the positive charge

• Small fragments move faster
DNA Fragments Move

• The colored solution provides an indication to how much the DNA has traveled on the Gel
Viewing

- Gel can be viewed under UV light
Viewing

- Original uncut DNA sample makes a sharp band at the beginning (one big fragment)

- DNA sample cut with Hind III makes a smear (lots of fragments of all sizes)

- Ladders are used for comparison (they contain specific fragments)

- We could run it for a longer time to achieve better separation
Hybridization

• In a hybridization experiment, we try to verify whether a specific sequence known as probe binds (or 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).
RFLP Markers

- We apply a number of probes in turn on the gel
- Each probe is mixed with a radioactive material
- Each probe hybridizes with a portion of the original DNA
- After cutting, the probe will hybridize with the fragments belonging to that portion
- These fragment can now be observed due to the radioactive material
- **RFLP marker** is defined by a probe and the set of lengths (unordered) of fragments that hybridize with the probe.
- Use analysis of recombination to order RFLP markers on the chromosome

![Diagram of RFLP markers and restriction enzyme cuts](http://engr.smu.edu/~saad/)
Illustration

probe

cut DNA

smear

fragments contained in the probe

Gel
First RFLP map in 1987

- Donis-Keller et al. constructed the first RFLP map of the human genome, positioning one RFLP marker per approximately 10 million nucleotides.

- RFLP markers (probes) need to be long enough to span the whole DNA.

- 393 random probes where used to study RFLP in 21 families over 3 generations.

- Computational analysis of recombination lead to the ordering of RFLP markers on the chromosome.
RFLP and Gene Finding

• Using the ordering of RFLP markers on a chromosome, we can approximately determine the location of a gene.
  
  – How?
  – Find the difference between the RFLP markers of family members with the disease and family members not having the disease.
  – It is likely that the RFLP marker that consistently differ is on the gene responsible for the disease, since family members have more or less the same genetic characteristics.
  – But we still don’t know where and what the exact gene is.
Physical Mapping

• Genetic mapping and RFLP
  – (1) do not tell the actual distance in base pairs
  – (2) if genes (or markers) are very close, one cannot resolve their order, because the observed recombination frequencies will be zero.

• Physical mapping reflect actual distances
  – Hybridization Mapping
  – Restriction Mapping
Hybridization Mapping

- Break several copies of DNA into fragments (using different restriction enzymes).

- Obtain many copies of each fragment (cloning, incorporating a fragment into a replicating host), forming a clone library.

- Clones may overlap (cutting DNA with distinct enzymes), and we want them to (we will see why).

- *Fingerprinting* the clones: Now use DNA probes, and for every clone determine the list of probes that hybridize with the clone.

- When two clones have substantial overlap, their fingerprints will be similar.

- Reconstruct the relative order of the clones using the overlap information (this order is unknown in RFLP)
Hybridization Mapping

- For \( n \) clones, and \( m \) probes, the hybridization data consists of an \( n \times m \) matrix \( D \), such that \( d_{ij} = 1 \) if clone \( C_i \) contains probe \( p_j \).

- Let \( S \) be a string over the alphabet of probes \( p_1 \ldots p_m \). \( S \) covers a clone \( C \) if there exists a substring of \( S \) containing exactly the same set of probes as \( C \) (order and multiplicity are ignored).

- A *simple approximation* of physical mapping is the **Shortest Covering String**.
Illustration

No overlap
No information

Covering String: 
Covering string

The clone is covered by the string

This clone hybridizes with 4 probes

String of probes
## Shortest Covering String

<table>
<thead>
<tr>
<th>clones</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

A covering string: \( S = AC \ ABEG \ BCDEFG \ BCDFG \ CDFG \ ADFG \ ABDEG \ ABDG \ D \)

A shortest covering string (max overlap): \( S = C \ A \ E \ B \ G \ C \ F \ D \ A \ G \ E \ B \ A \ G \ D \)

Shortest Covering String: NP-hard Problem in general. If probes are unique, a polynomial algorithm exists.
Unique/Non-Unique Probes

• Non-unique probes: probes are short random sequences that can occur many times in the DNA. Therefore, a probe can hybridize with distant clones.

• Unique probes: probes are sufficiently long and are unlikely to occur twice in the DNA. Therefore, a probe will hybridize with close clones.

• Advantages of non-unique probes: probe generation is cheap and straight-forward.
Restriction Mapping

• Before using the list of probes in a clone as a fingerprint, biologists used the order of restriction fragments in a clone.

• Restriction map as Fingerprinting: If two clones share several consecutive fragments, they are likely to overlap.

• Restriction map of a clone: an ordered list of its restriction fragments (Hard Problem).
Double Digest

• Cut the DNA fragment with enzyme A, then enzyme B, then both

• Obtain a multiset of lengths in each case (using Gel electrophoresis)

• Using this information, construct an order of the lengths

• A: \{2,2,3\}  2 3 2
• B: \{3,4\}  3 4
• A+B: \{1,2,2,2\}  2 1 2 2

Saad Mneimneh
http://engr.smu.edu/~saad/
Partial Digestion

- Instead of obtaining lengths of restriction fragments, the DNA is digested in such a way that fragments are formed by every two cuts and the lengths of all fragments are obtained.

- The problem often might be formulated as recovering positions of points on a line when only some pairwise distances between points are known. (why?)

- Many mapping techniques lead to the following problem: \( X \) is a set of points, \( \Delta X \) is the multiset of all pairwise distances between points in \( X \): \( \Delta X = \{|x_1 - x_2| : x_1, x_2 \in X\} \), \( E \subseteq X \) is given. Reconstruct \( X \) from knowing \( E \) alone.

- Partial Digest Problem. Given \( \Delta X \), reconstruct \( X \) (\( E=\Delta X \)). Also known as the turnpike problem in computer science, construct the geography of the highway from knowing the distance between every two exits.

- No polynomial time algorithm for this problem is yet known, but in practice, efficient algorithms exist.