Lecture 14

Gidon Rosalki

2025-12-07

1 DNA Assembly

Let us suppose that we have a solution of a bacterium, and we want to analyse its DNA. To do this, we extract the DNA, and sequence it (nice black box called Illumina). This produces around 20M short sequences, of around 150 base pairs. Each sequence of these is called a word, and we call this in bio a "read". Now we have a lot of sequences, but we want to perform DNA Assembly, and construct some genomes. If we assume that every section of the genome has 10 reads, then we call this 10x coverage. We will assume that there are many reads per area of genome.

Let us suppose that there are two parts of the genome that areas of the genome that are exactly equivalent. What if we have a read that crosses into one of these regions. If we have two paths that lead into one of these regions, then we do not know which one is meant for which region, making it much more difficult to construct the genome.

1.1 De-Bruijn graph

A k-mer is a k-tuple, but in biology, and over the alphabet {A, C, G, T}. Consider a read, which has 150 base pairs. We may now pass a window of length k over the read, and receive all the possible k-mers out of this sequence. From this, we may construct the De-Bruijn graph. This graph has nodes for each k-mer, and there is a link if:

1. The last $k-1$ letters of one node is the first $k-1$ of the second
2. This $k+1$ mer sequence of letters from merging the two nodes appears in the data

De-Bruijn graphs are huge. Therefore, we are going to compact it into something that is called an assembly graph.

1.1.1 Assembly graph

In an assembly graph, we may compress linear sequences into a single node, of length longer than k. Additionally, we will create smaller nodes, which are the connection nodes between longer sequences. Every read is still demonstrated in this graph, since we have removed no information, however there will also be reads in this graph that are *not* in the original genome.

So now, we want the sequences we acquire from the graph to be as close as possible to the original genome. Since there will be numerous junctions, called x structures, where we connect between 4 different options, we need a way to find which options at these junctions appear in the original genome. Let us return to the earlier problem of having sections that are repeated. In this graph we may simply create a loop, of an edge that points back to the same node.

We can now construct all "probable" paths. We could flow through this graph, by adding a source node connected to every node that has now entry edges, and similarly a sink to all the nodes that have no exits, and then find the maximum flow in this graph.

2 RNA Assembly

If we instead put RNA into the sequencing process at the beginning, then we will receive a similar assembly graph. This sequence will create a "transcriptome", which contains all the mRNA molecules. Since the transcripts of the genes are not continuous, they are not necessarily connected, we will get an individual connected component for the transcript of each gene. Since some connected components will have very strong connections, we may thus say that they appeared more than the CCs which have weaker connections.

Consider how one receives two copies of one gene, one from the mother, and one from the father, but both have undergone some slight changes in the process. If in these genes, there is a single letter that has changed between them (so a C in one, but a G in the other for example), then we will get a bubble in the graph, where it breaks into 2, one for each letter, and then recombines. Recall the x structures. If every x structure is sufficiently close such that the distance between them is smaller than a read, then we may solve them, simply by finding them in the reads, rather have them multiply the options by 4 every time. We can thus open the x structure into 2 different continuous sequences.

There are two mains types of RNA sequence assembly. There is reference based, which is what most people do, where we know things about the origin, and can use this to build our genomes. There is also reference free, where we know nothing about the origin, and still try and build genomes from this. This is very powerful, since these references can severely limit us. For example, cancer. When a genome is cancerous, it is connected to a section that comes from a different location altogether.

3 Metagenomic assembly

Consider trying to find the genomes of bacteria from rainwater. When we break them down, we get lots of DNA sections, but from all the different genomes, all of which will be very similar. As a result, all our connected components are going to be a very large mess. Hopefully, much like earlier where we solved the x junctions, there will be similar behaviour, such that we can split it into different, but very similar genomes. We will note that since these graphs are so *huge*, any computation of this sort will take a very long time.

4 Trees

Recall phylogenetic trees. The leaves indicate organisms that are currently alive, and internal nodes indicate organisms that are no longer alive, their ancestors. By taking the genomes of the currently living organisms, we may build the distance matrix, and then from this construct a phylogenetic tree.

Moving on, let us consider a gene tree. This may act like a phylogenetic tree, but for a specific gene. Here the leaves are genes that currently exist, and the internal nodes are previous versions of these genes. These can be rather different since genomes may get duplicated, resulting in duplicated sections of the tree.

Consider a human microbiome (like ones gut microbiome). Unless we are talking about people who live in the same flat, it is unlikely that they will have the same microbiome. To measure this sort of thing, phylogenetic trees are a little too small scale, so we may instead move on to species trees. Here, each leaf represents an example (an individual's personal microbiome). We create a distance matrix between each leaf, and from here can build the tree.

Some trees will have adjacent leaves belonging obviously to the same family of biomes. In this case, the most recent common ancestor will be the immediate parent node. However, sometimes these nodes do not appear quite so neat, and then the most recent ancestor is further up the tree.
What could cause this sort of messy tree? Well, there could be mutation rate between the generations. There could also be differences since the parent and child could receive these bacteria from locations that are not their shared home.

How do we build these sorts of trees? This is hard for the following reasons. Consider, we get 4 sets of 20M reads, two from the mother, and two from the child. We know the genome for e-coli, so may try and map our 20M reads onto the genome of e-coli. However, this is still difficult to find the difference, since often there may be shared genome sections between e-coli and other bacteria. However, we may use "marker genes", which are unique to e-coli.

If we have a gene, that we know is only located in a single organism, and we have some genes with variances from these, then we may say that they probably came from the same original organisms. We may also use reference free assembly, and compare between the entire genomes that we have built. Each method has its advantages, and disadvantages.

In summary to this section, phylogenetic trees exist, but we may use the same concepts in many different areas (like finding similarity between gut microbiomes, which is not just phylogeny).