Poster H-20

A De Bruijn Subgraph Approach to Repeat Identification in Incompletely Sequenced Genomes

Authors:
José Augusto Amgarten Quitzau (*University of Bielefeld*)
Jens Stoye (*University of Bielefeld*)

Short Abstract: We present an approach for repeated regions identification that can also be applied for incompletely sequenced genomes. The approach is based on Pevzner and colleagues approach for genome assembly implemented in the Euler assembler. We conjecture that the approach may be extended to the de novo identification of repeat families.

Long Abstract:

Not all genome projects end up with the complete sequence of the target genome. Many of the already sequenced genomes are not completely known simply because the aim of their sequencing projects was not the finished sequence of the whole genome. The goal may be sometimes just to explore some specific types of sequences, like Expressed Sequence Tags (ESTs), for instance. As a result, there is a good deal of sequences in the public databases that are only parts of incompletely sequenced genomes. However, it does not mean that these sequences are useless.

Since the aim of these projects is not the complete sequence, usually they try to focus on the "most interesting" parts of the genome. With some luck it is possible to identify interesting parts of the DNA sequence, which may be amplified via PCR and studied in more detail latter. To amplify these sequences, specific PCR primers are needed. This may be a problem when the genome sequence is not completely known. In order to amplify the region of interest, the primers must be very specific; otherwise a PCR experiment would amplify the wrong region, or give no result at all. Therefore it is very important to know which parts of the sequences are unique, so that we can be sure that a primer specially designed for the region will not accidentally bind a similar sequence in another part of the genome. In this context, the identification of repeated regions in a genome may save both time and money, since valuable resources are lost when primers are designed in non-specific regions.

There are softwares able to identify repeats in completely sequenced genomes. Unfortunately, the first step of de novo repeat identification is usually the alignment of the whole genome against itself. In our case, it is not possible because we don't even have a sketch of the assembled sequence. Our input data is just a collection of sequences, probably with several copies of the same portion of the DNA molecule. We need to do at least a partial assembly before starting any kind of de novo repeat identification. This is the point where De Bruijn subgraphs become especially interesting.

The n-dimensional De Bruijn graph on an alphabet is the graph that has all the possible strings of length n over the alphabet. There is an arc going from a vertex u to a vertex v in the De Bruijn graph if by deleting the first character of u and the last character of v we have

the same string. It is not difficult to notice that strings over the same alphabet with length greater than n describe a walk on the n-dimensional De Bruijn graph. Given a set of strings over the same alphabet, we define the corresponding De Bruijn subgraph as the De Bruijn graph that contains all the corresponding walks and no extra vertex or arc.

In the year of 2001, Pavel Pevzner, Haixu Tang, and Michael Waterman presented a De Bruijn subgraph approach for genome assembly. This approach is quite simple if the genome has no repeated region larger than the dimension of the De Bruijn subgraph. Of course this is not always the case; and to solve the assembly problem for repeated regions was one of the big challenges to this approach. The good news for us is that we can use the same de Bruijn subgraph approach up to the point it becomes difficult. We actually don't need to untangle repeated region, but only identify them. While Pevzner and his colleagues has a hard to work to do after building the graph; we need only to identify and mark the parts of the graph corresponding to repeated regions.

Our approach for repeated region identification consists of two steps:

1. Partial assembly with repeated region identification.
2. Mark out repeated regions in the input sequences.

The first step is done by inserting every sequence given as input in an initially empty De Bruijn subgraph. The sequence insertions themselves are done in two steps:

1a. We use a sliding window to identify the sequence walk. Then we follow the sequence walk creating the vertices that are not in the graph. Every time the walk reaches a pre-existing vertex, we mark the vertex as an "old" one.

1b. We follow the sequence walk again, identifying any contiguous block of "old" vertices. If at the end of this step we identify no block or a single block that contains the whole sequence, we just continue the procedure; otherwise, we mark every vertex in every block as a "repeated region".

To mark out the regions corresponding to repeats in the given sequences, we use again a sliding window, marking every sequence that corresponds to a marked vertex in the graph.

This repeat identification approach was implemented in JAVA and tested with a collection of sequences of a known plant transposon family and with new sequenced sequences of the plant Beta vulgaris. The results showed not only the efficiency of the method, but also the possibility of using it for de novo repeat families identification.