Scribes

Genome sequencing and assembling

By
Shubhra Gupta

The ARACHNE WGS assembler

ARACHNE, a new computer system, for assembling genome sequence using paired-end whole-genome shotgun reads (short DNA sequences). **ARACHNE** has several key features, including an efficient and sensitive procedure for finding read overlaps, a procedure for scoring overlaps that achieves high accuracy by correcting errors before assembly, read merger based on forward-reverse links, and detection of repeat contigs (The assembled sequences) by forward-reverse link inconsistency. How ARACHNE WGS assembler works. There is a algorithm for assembler

Input data

ARACHNE analyzes paired end reads obtained by sequencing both ends of a plasmid of known insert size and assumes each base in each read has an associated quality score (say one obtained by PHRED program). A quality score q corresponds to the probability $10^{-q/10}$ that the base is incorrect (40 corresponds to 99.99% accuracy). As an Initial step, ARACHNE eliminates terminal regions whose quality is low and eliminates reads containing very little high-quality sequence. It also eliminates known vector sequences and known contaminants (e.g. Sequence from the bacterial host or cloning vector).

Overlap detection and alignment

ARACHNE creates a sorted table of each k-letter subword (k-mer) together with its source (which read) and its position within the read. The program then excludes k-mers that occur with extremely high frequency. This corresponds to highly repeated sequences and used to increase the efficiency of the overlap detection process. It then identifies all instances of read pairs that share one or more overlapping k-mer, and a 3 step process (similar to FASTA) to align the reads efficiently. First step is to merge overlapping shared k-mers. Second step is to extend the shared k-mers to alignments and the third step is to refine the alignment by dynamic programming. In this process some valid alignments may be missed and some invalid ones may result.

ARACHNE: Error Correction

ARACHNE can detect and corrects sequencing errors by generating multiple alignments among overlapping reads. The program then identifies instances where a base is over-whelmingly outvoted by bases aligned to it (taking into account the score quality) and corrects the base.
ARACHNE similarly corrects occasional insertions and deletions, which mostly appear due to sequencing errors. As the reads are corrected, corresponding changes are made to the alignments.

```
TAGCTTACACAGATTACTGC      C: 20    C: 20
TAGATAACACAGATTACTGA      C: 35    C: 35
TAG  TTACACAGAGTATTGC      T: 30    C: 0
TAGATAACAC  GATTACTGA      C: 35    C: 35
TAGATTACACAGACTACTGA      C: 40    C: 40
```

In the above example, correction error in reads is shown. In the fourth column of the alignment from right hand side, a base T of quality 30 is aligned only to bases C. The base T is changed to a base C of quality 0.

ARACHNE: Evaluation of Alignments

ARACHNE assigns a penalty score to each aligned pair of overlapping reads. Then a penalty scores are assigned by program to each discrepant base, based on the sequence quality score at the base and flanking bases on either side. Discrepancies in high quality sequences are assigned high penalty, and discrepancies in low quality sequences are penalized less heavily. The penalty scores for individual discrepancies are combined to yield an overall penalty score for the alignment. Overlaps incurring too high a penalty are discarded. Likely chimeric (Reads that contain genomic sequence from two disparate locations are termed **chimeric**) reads are also detected and discarded.

ARACHNE: paired pairs

Identification of paired pairs

ARACHNE identified paired reads as those reads, which are known to relate with respect to orientation and distance. ARACHNE searches for instances of two plasmids of similar insert size with sequence overlap occurring at both ends (Together called **paired pairs**). Building complexes of such pairs can be extended by these instances (figure 1.1).

```
_          ***********                    *********
_       **********                     **********
_  *********                        ********
```

Collections of paired pairs are merged together into **contigs**.

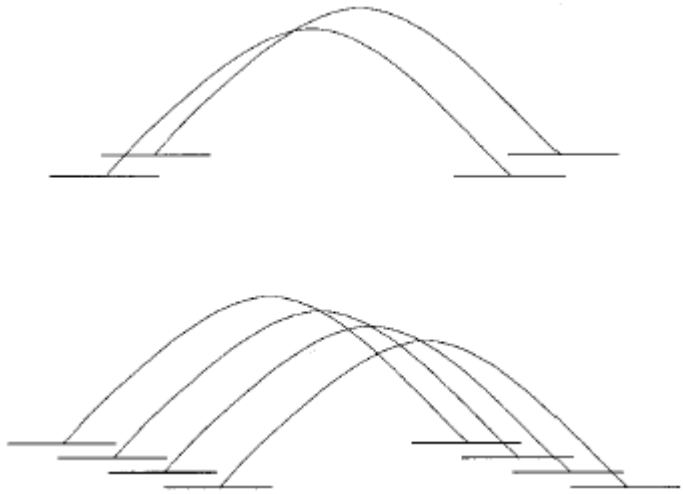

Figure 1.1

Above figure is a paired pair of overlaps. The top two reads are end sequences from one insert, and the bottom two reads are end sequences from another. Initially, the top two pairs of reads are merged. Then the third pair of reads (from the top) is merged in, based on having an overlap with one of the top two left reads, an overlap with one of the top two right reads. Similarly the merging of pair in bottom figure is shown.

ARACHNE: Contig Assembly

When repeats are absent the correct assembly can be easily obtained by merging all the overlapping reads. In the presence of repeats, false overlaps may arise between reads derived from different copies of a repeat (figure 1.2).

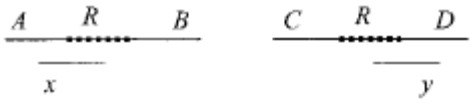

Figure 1.2

ARACHNE identifies potential repeat boundaries and avoids assembling contigs across such boundaries. Potential repeat boundaries are marked by program whenever a read r can be extended by x and y, but x and y do not overlap. Merging of overlapping read pairs that do not cross a marked repeat boundary (figure 1.3, 1.4).

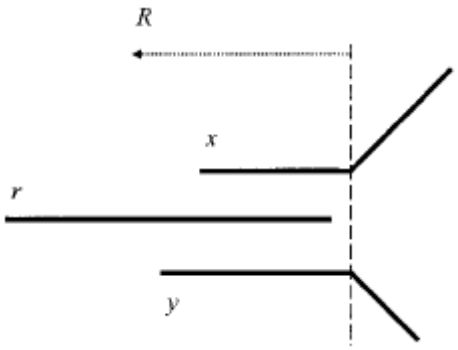

Figure 1.3

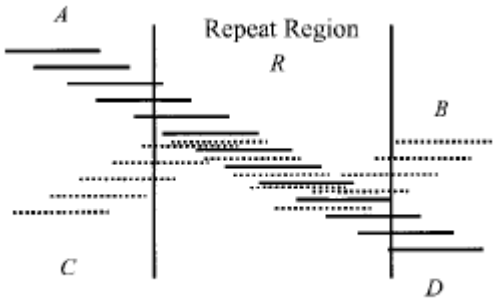

Figure 1.4

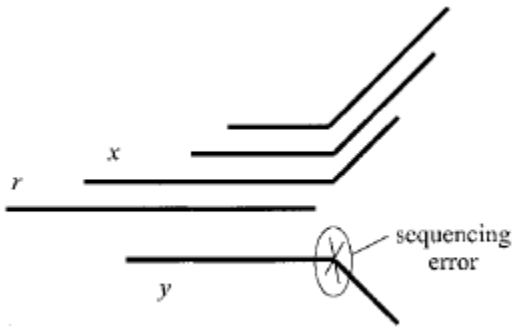

Figure 1.5

ARACHNE: Repeat Contigs and Supercontigs

Detection of repeat contigs

Detection of repeat contigs can be identified in two ways. First way of identification has an unusually high depth of coverage. Second way is to conflicting links to multiple,

distinct, non-overlapping contigs, reflecting the multiple regions that flank the repeat in the genome.

• aRb, cRd, eRf … will result in ⁻aR-, -cR-…. (Figure 1.6)

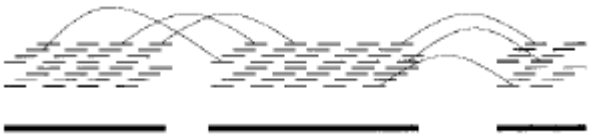

Figure 1.6

Creation of supercontigs

After marking repeat contigs the unmarked contigs (called *unitigs*) are assembled. ARACHNE use forward-reverse links from reads to order and orient unique contigs into supercontigs.

ARACHNE: Filling gaps in supercontigs

The layout is a set of contigs each of which is an ordered list of contigs with interleaved gaps. These gaps correspond to two kinds of regions. These regions marked as repeat contigs (which were omitted in supercontig construction). Such region for which there is insufficient number of shotgun reads to allow assembly. So ARACHNE fills gaps using repeat contigs. For every pair of consecutive contigs with an interleaving gap in a supercontig S, the program tries to find a path of pairwise overlapping contigs that fill the gap. Forward-reverse links from S guide the construction of the path by identifying contigs likely to fall in the gap.

Consensus derivation and postconsensus merger

The layout of overlapping reads is converted into consensus sequence with quality scores. This can be done by converting pair-wise alignments of reads into multiple alignments, and deriving the consensus base by weighed voting.