Human-guided genome assembly software

Evgeny Gerasimov,
Lomonosov Moscow State University;
Institute for Information Transmission Problems RAS
jalgard@gmail.com

Pavel Flegontov
Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava, Czech Republic;
Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic;
Institute for Information Transmission Problems RAS, Moscow, Russia
pflegontov@gmail.com

Genome assembly is arguably the most computationally intensive part of genome analysis. Nowadays it seems to be well-developed from the algorithmic point of view. But still there are many difficulties.

The main problem is a repetitive nature of genomic DNA. There are lots of repeats of different types and lengths: mobile elements, control regions (like transcription factor binding sites, promoters), duplicated genes. On the other hand, there are allelic variants in diploid genomes, which are in fact just two mutually exclusive variants of the same locus.

These features of DNA sequence make genome assembly ambiguous and sometimes even contradictory. Automated genome assemblers usually attempt to output only unambiguous (unique) regions (termed contigs), and also discover most probable arrangement of these contigs using paired reads, producing scaffolds. Scaffolding procedure is not always accurate. An obvious example is allelic polymorphism: only one variant will be present in the final assembly. But a greater problem is at the level of collapsed repeat resolution: where in the genome and how many times this sequence is actually present? Usually this is the main obstacle that prevents an automated algorithm from finding the path in the assembly graph.

Here we present a software which is specially designed for manual assembly finishing. The main goals of this software are: 1) give a researcher an instrument to look at an assembly and evaluate it; 2) close gaps in existing scaffolds, make new or correct the existing scaffolds; 3) use experimental evidence (PCR products, mate pair library sequences, assembly annotation by BLAST and so on) to perform more accurate scaffolding or to resolve situations when multiple variants are possible.

The software has GUI developed with cross-platform library QT, most actions user can perform using mouse. GS denovo assembler's (newbler's) output files are used as initial draft assembly, 454ContigGraph.txt file contains information about contig graph. Additional files (fastq and fasta) can be loaded optionally and used as hints during scaffolding procedure. In future we plan to provide more functionality especially automated procedures of resolution for simple cases in the contig graph and suggestions of most probable path.

We suggest that this software is especially useful for small genomes with complex repeats like genomes of some bacteria and many organelles. Using bacterial genome of *Paenibacillus* sp. (size 5.5 Mbp) we demonstrate that our human-guided assembly approach can significantly improve the resulting assembly.