Accurate De Novo Assembly for Genomes with Repeats

Doina Bucur
d.bucur@utwente.nl

Summary

A machine-learning method obtains accurate assemblies for genomes with repeats and self-validates assemblies via **consensus**. The tool (a) assembles **variable-length raw reads**, which may unambiguously recover *interspersed repeats* in the genome, and (b) in the presence of *long, direct terminal repeats* it reports a **circular assembly**. Consensus is obtained via stochastically independent runs of the assembler.

Genomes assembled

Table 1: Genomes assembled and the size of their largest repeat cluster. Abbreviations: *direct interspersed repeat* (DIR), *direct terminal repeat* (DTR), *inverted interspersed repeat* (IIR), *inverted terminal repeat* (ITR).

Species	Acc. no.	Genome size (bp)	Repeat type (size in bp)	Main repeat
Hepatitis C virus (**HCV**)	NC004102	9646	tandem mono-DIR (51, 17, 20)	
Ciconia mitochondrial DNA (**mtDNA**)	AB026818	17347	tandem tetra-DIR (240), tandem DIR (310)	DIR
Simian virus 40 (**SV40**)	J02400	5243	tandem DIR (144)	
Frog adenovirus (**FrAdV**)	NC002501	26163	DIR (609), ITR (36)	
Human immunodeficiency virus 1 (**HIV-1**)	JQ316128	9257	DTR (142)	
Adeno-associated virus (**AAV**)	AF043303	4679	DTR (125), ITR (43) + IIR (63)	DTR
Proteus phage (**PM16**)	KF319020	41268	DTR (450)	

Synthetic read libraries: reads are randomly sampled from the reference genome: the reads have random variable lengths in a given interval, and the reference genome is covered. No artificial faults (single nucleotide variations or indels) are added in this study.

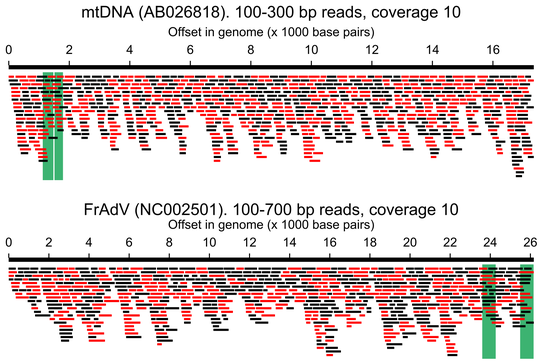


Figure 1: Synthetic read libraries. The length of the genome is shown as a contiguous black line. The repeats are marked in green. The reads are shown in two colours: black for forward reads, and red for reads which are reverse-complemented.

Each genome is assembled in trials from 5 read libraries sampled randomly with a given read-length interval from the same genome, to avoid drawing conclusions based on a sampling bias. Also, read libraries are sampled with multiple read-length intervals, to learn the extent to which read lengths improve the assembly.

Assembly algorithm

Iterative, stochastic optimization with a gradient descent [1, 2]. "Learns" the order (and sense) of the reads using a **genetic algorithm**: every iteration attempts to improve the scaffold via operators inspired by natural evolution, guided by a *fitness function*.

Candidate solution: a *segmented permutation* of raw reads. The algorithm maintains a "population" of solutions. Solutions with better fitness are selected and modified using random **mutation** and **crossover** operators.

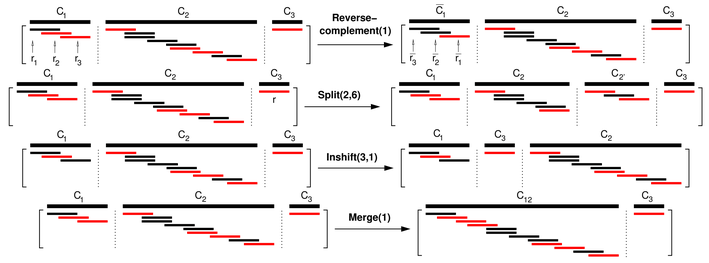


Figure 2: Mutation operators. Reverse-complemented reads are shown in red. $\overline{r_1}$ denotes the reverse-complement of r_1.

Fitness function: the number of contigs on the scaffold of the solution, plus the string length of the scaffold.

Linear or **circular** assembly modes are possible for the tool (for a genome with long DTRs, choose circular: the shortest assembly is the circular form).

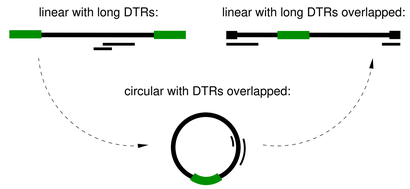
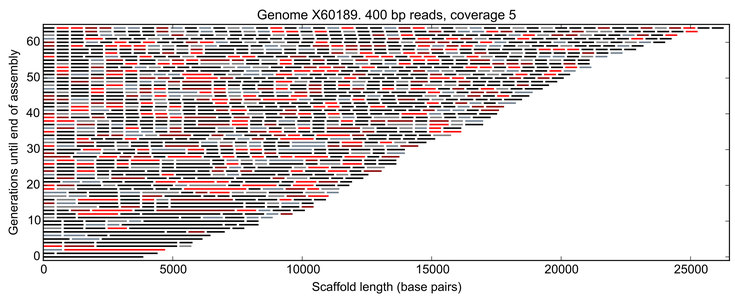


Figure 3: Equivalent linear and circular forms for genomes with DTRs. DTRs are shown in green.

Experimental settings: 100 candidate solutions per population. After selection, a sequence of operators is applied to all solutions. A run of the algorithm below.



Results

Strict **accuracy metrics**: we require that zero misassemblies occur, i.e., the resulting assembly is a single-contig, accurate, linear or circular genome. We execute 20 stochastically independent assembly runs, times 5 synthetic read libraries per genome and read-length interval.

Below, the colour of a data point gives the likelihood (% out of 100 independent assembly runs) that the correct genome is obtained. The diameter of a data point is proportional to the size of the underlying genome. The x axis measures the difference between average read length and repeat size. The y axis measures the size of the read library.

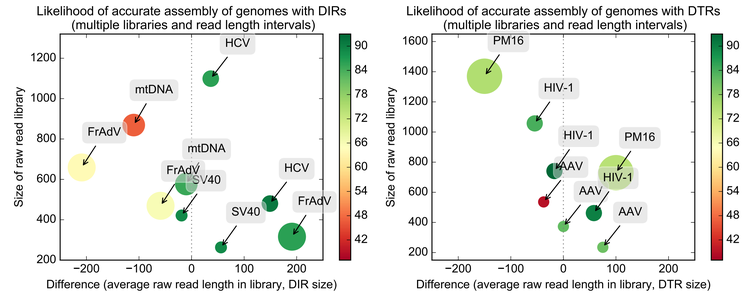


Figure 4: The likelihood of accurate assemblies for genomes with DIRs (left) and DTRs (right).

Accuracy: for linear assemblies, it varies with the average read length for genomes with DIRs (both tandem and interspersed); for circular assemblies, it is more robust. Accurate assemblies, when starting from reads of suitable length, are obtained in a large majority of the runs.

Forming consensus: this assembler is likely to converge on the correct assembly across independent runs. The user can apply this to determine the correct de-novo assembly by majority call.

References

[1] Doina Bucur. *De Novo DNA Assembly with a Genetic Algorithm Finds Accurate Genomes Even with Suboptimal Fitness*, pages 67–82. Springer International Publishing, 2017.

[2] Doina Bucur. A stochastic de novo assembly algorithm for viral-sized genomes obtains correct genomes and builds consensus. *Information Sciences*, In print, 2017.

UNIVERSITY OF TWENTE.