Maximum Likelihood Scaffold Assembly

Anton Akhi, Alexey Sergushichev, Fedor Tsarev

St. Petersburg National Research University of Information Technologies, Mechanics and Optics

Genome Assembly Algorithms Laboratory

Scaffolding Problem

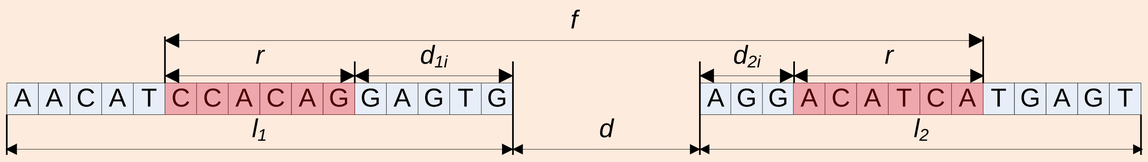
- The process of DNA assembly is usually divided into two parts: assembling contigs and assembling scaffolds from contigs
- Scaffolding is an important step of genome finishing
- Contig is a relatively long continuous fragment of DNA sequence
- Scaffold is an ordered set of oriented contigs with distance estimations between them
- Scaffolds are usually assembled from contigs using mate-paired libraries with large fragment size (usually greater than 1 Kbp)

Preliminaries

- Normal insert size distribution $N(\mu, \sigma)$ with probability density function $f_{\mu,\sigma}(x)$
- Uniform position distribution
- L – genome length
- R – number of mate-pairs in library
- Reads are aligned to contigs with Bowtie
- Proposed algorithm (GAMLET) consists of two parts: distance estimation and contig ordering and orientation

Distance Estimation

- Maximum likelihood principle



- Taking into account connecting mate-pairs:

$$P(d_{1i}, d_{2i} \mid d) = f_{\mu,\sigma}(d_{1i} + d_{2i} + d + 2r)\frac{1}{L}$$

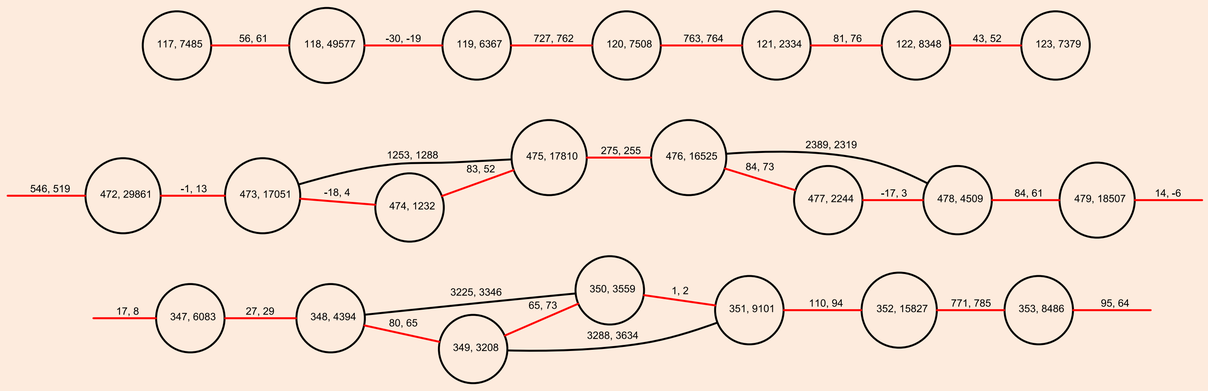
- The resulting likelihood is a product of the connecting reads probabilities and probability that all other reads do not connect contigs

$$\prod_{i=1}^{n} P(d_{1i}, d_{2i} \mid d) \times \left(1 - \sum_{s} f_{\mu,\sigma}(s)\frac{w_d(s)}{L}\right)^{R-n}$$

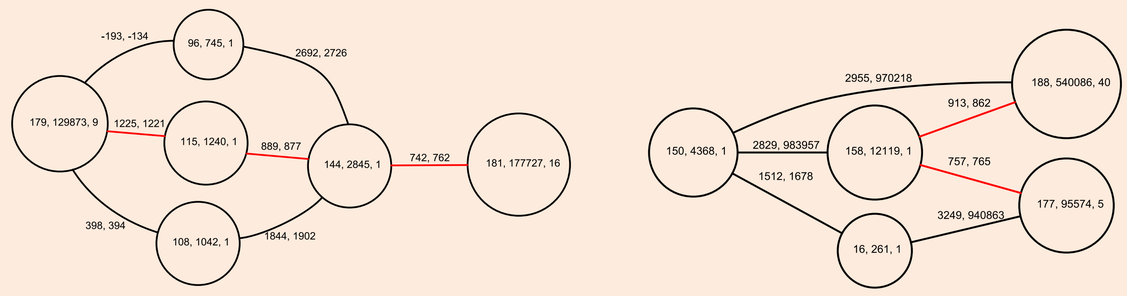
- $w_d(s)$ – number of possible ways for a pair of reads with insert size s to connect a pair of contigs
- n – number of mate-pairs connecting contigs
- Finding the most likely distance using ternary search

Ordering & Orientation

- Maximum parsimony principle
- Building graph with contigs as vertices and connections as edges
- Removing high-degree vertices and short contigs
- Building draft scaffolds along shortest edges



- Building graph with draft scaffolds as vertices
- Removing highly-covered vertices
- Searching for shortest paths between draft scaffolds
- Merging scaffolds along the shortest paths



- Contig orientation is found using mate-pair alignment orientation. Only mate-pairs connecting contigs in the same scaffold are considered.

Experiments

- *E.Coli* genome
- Set of 502 contigs: N50 = 18047, minimum length 235, maximum length 73908, average length 9126
- Three sets of 600000 mate-pairs generated by MetaSim: length 36, mean insert size 3000, standard deviation 300
- Distance estimation algorithm was compared with SOPRA and GAPEST. For each algorithm average absolute difference between estimated and correct distances was computed:

Read set	SOPRA	GAPEST	GAMLET
Set1	209±15	175±14	**149±13**
Set2	139±13	217±17	**135±13**
Set3	215±15	172±14	**153±13**

- Scaffolding algorithm was compared with OPERA

Read set	OPERA			GAMLET		
	N50	N50 (split)	errors	N50	N50 (split)	errors
Set1	**366.5k**	215.7k	11	311.7k	**311.7k**	**1**
Set2	428.6k	215.7k	11	**605.3k**	**392.0k**	**7**
Set3	465.0k	294.0k	11	**578.2k**	**322.6k**	**9**

Conclusions

- Proposed distance estimation method is more precise on average than SOPRA and GAPEST
- Proposed algorithm assembles longer scaffolds with fewer number of misassembles
- Proposed algorithm is an order of magnitude faster than OPERA: 1.5 min. vs. 15 min. on *E. Coli* dataset

Acknowledgements

Research is supported by the federal program “Research and scientific-pedagogical personnel of innovative Russia in 2009-2013” (contract 16.740.11.0495, agreement 14.B37.21.0562)

mailto:genome@mail.ifmo.ru
http://genome.ifmo.ru/en/