A one-step method for de novo genome assembly using short paired-end sequence reads

David H. Silver and Itai Yanai

Department of Biology, Technion – Israel Institute of Technology, Haifa, Israel

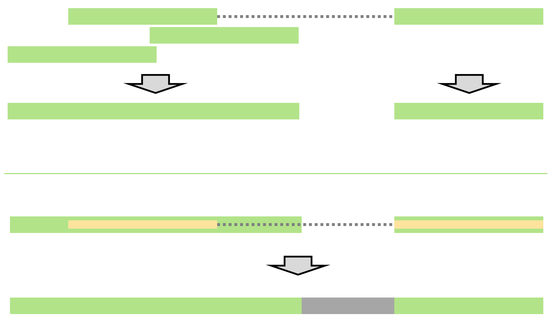
ABSTRACT

Genome sequencing is currently only possible by assembling short sequence fragments of DNA by identifying overlapping regions. To help in this computationally intensive task, large DNA regions can be sequenced from both ends, yielding "paired-end reads". Currently however, paired-end information is incorporated as an additional step following the assembly of contiguous sequences ("contigs"). The underlying reason for this has been the notion that this extra information derived from "paired-end" reads contributes mostly to distinguishing repeat regions, rather than to the main task of assembling. Here, we demonstrate that in fact the paired-end information is of tremendous aid to the assembly process itself. Further, we present a mathematical model which provides an explanatory framework for the success of our algorithm, thus closing the big gap between the theory of *de novo* assembly and the practice. Our algorithm benefits *de novo* genome assembling by a better assembly in terms of sequence length with fewer assembly errors and requiring significantly less sequence coverage.

SHORT-READS ASSEMBLY

All methods for finding consensus sequences use a two step approach in which the paired end reads are first treated as independent sets and the information about the link between is utilized only for scaffolding the contigs together. Those merged sequences are the Contigs. We notice that by far, this method is exactly as sequencing single-reads. The next step, which uses the paired-end information, is to take the contigs and look for pairs of contigs that we have both ends of the same read on two contigs, by this we can deduce the relative position of contigs and we merged them while putting 'N's to denote unknown part of the sequence.

Two-step approach:



1) Merging overlapping reads into contigs. The dotted line represent connection to the paired-end.

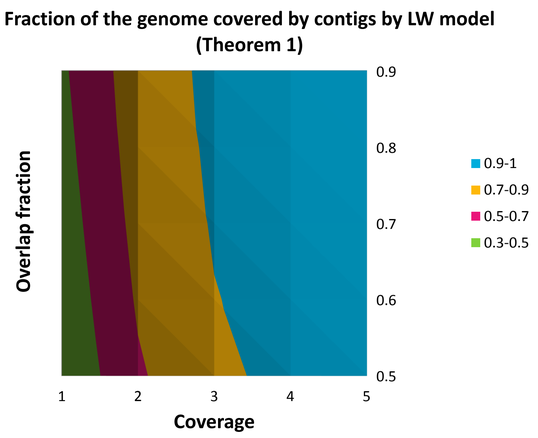
2) Connecting two contigs based on paired-end information into a scaffold.

MATHEMATICAL FRAMEWORK

Lander-Waterman Model

We define the following terms:

$$
\begin{aligned}
G &= \text{genome length} \\
L &= \text{read length} \\
N &= \text{number of inserts} \\
c &= \frac{L \cdot N}{G}, \text{ the expected coverage} \\
T &= \text{length requiered to detect overlap} \\
\theta &= \frac{T}{L}, \text{ the overlap fraction} \\
\sigma &= 1 - \theta
\end{aligned}
$$



Theorem 1. *If $\theta < \frac{1}{2}$ $(\sigma > \frac{1}{2})$, that is, the overlap we requiere for a detection is strictly less than half of the read length, then the expected fraction of the genome covered by contigs is:*

$$1 + e^{-2c\sigma}\left(\left(c\left(1-\sigma\right)-1\right)\left(2-e^{-c\sigma}\right)-c\right) - e^{-c}\left(1-e^{-c\sigma}\right)^2$$

Theorem 2. *The probability that a declared overlap of T-bp is false positive in a random sequence of length G is approximated by:*

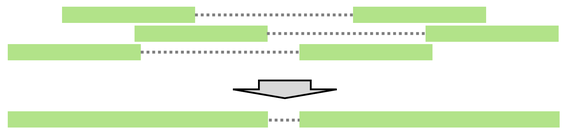
$$c \cdot (G - L + 1) \cdot 4^{-T} \approx 4^{-T}$$

The Theoretical Prediction for Assembly

From the Lander-Waterman model, we would expect that coverage of 10x would be enough to assemble the genome; given that other parameters such as read length are optimized. In practice however, much more coverage is needed. We can deduce that the main reason for that it is that when we have short-reads (36-100bp), an issue arises regarding the length of the overlapping detection T ("k-mer length") and by that also choosing θ. While on one hand we want θ to be small enough (requiring a small T) to detect overlaps, for θ has a major effect on the success of the assembly in an exponential rate (Theorem 1). On the other hand, a small overlap threshold (*T*) will result in a lot of false positives in a polynomial rate (Theorem 2). So in practice a high coverage is necessary and search empirically best overlap length.

OUR METHOD

In our method, we are incorporating paired-end information in the contiging stage. Instead of treating the paired reads as two single reads and looking for overlapping reads at each end, we are looking for pairs which overlap the pair in question at both end, which gives us the freedom to take smaller overlap threshold without risking getting misassembles.



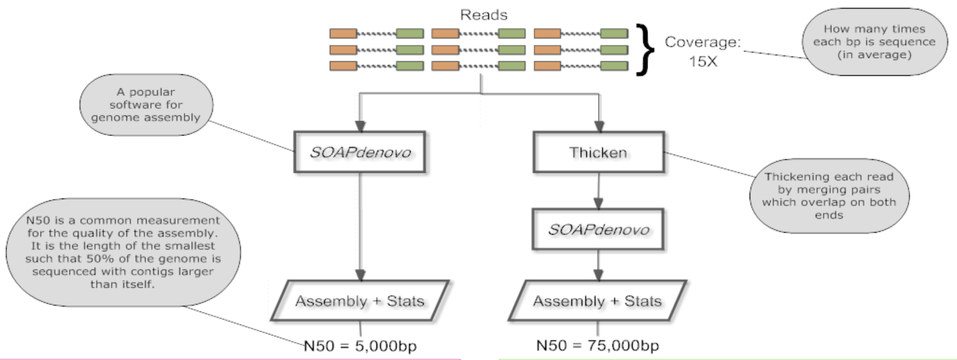
Merging reads which overlapping at both ends. The dotted line represent connection to the paired-end.

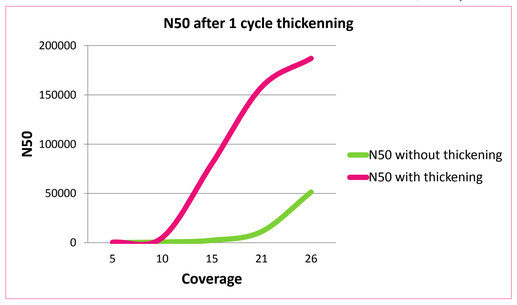
Our method allows one to have the cake and eat it too. For when we take θ to be small, that is, the overlap we require in each of the two ends is small. Yet, the robustness that follows from this threshold is not that associated with T, but rather with 2*T. This is because we get an overlap of T on both ends which is effectively similar to having an overlap of 2*T on one end. Another way to phrase it, is that with our method we can look as if reads are 2*L.

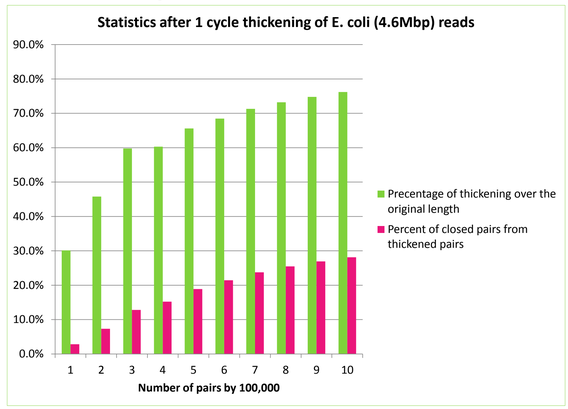
RESULTS

We can use our method either as a preprocessing step to the raw reads by thickening every read-pair by 2-ends overlapping reads and then feeding it to any assembly software, or we can thicken the reads over and over ("cycles") until exhaustion and we get long reads which are easy to assemble due to large overlap regions between term.

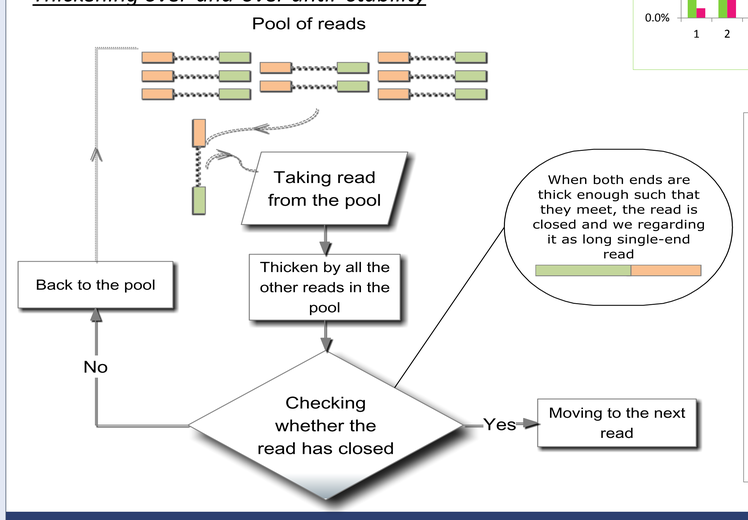
An assembly of 180kbp phage genome, with and without preprocessing (thickening)

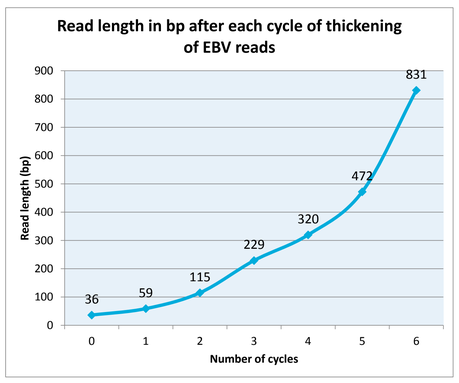






Thickening over-and-over until stability





IMPLEMENTATION

Currently our algorithm is implemented in Perl and we are using a module from ABySS software to make the overlap graph between reads for both of their ends. This module constructs the graph using Google Sparse Hash, which makes the algorithm fast enough for processing high-throughput data from large genomes.

REFERENCES

1. Genomic mapping by fingerprinting random clones: A mathematical analysis. ES Lander & MS Waterman, Genomics 1988
2. Genomic mapping by end-characterized random clones: a mathematical analysis . E Port, F Sun, D Martin & MS Waterman, Genomics 1995
3. SOAPdenovo: http://soap.genomics.org.cn/soapdenovo.html
4. ABySS: A parallel assembler for short read sequence data. JT Simpson, K Wong, SD Jackman, JE Schein, SJM Jones & I Birol, Genome Research 2009
5. Google-sparsehash http://code.google.com/p/google-sparsehash/