A fast hybrid short read fragment assembly algorithm

Bertil Schmidt, Ranjan Sinha, Bryan Beresford-Smith and Simon J. Puglist

1School of Computer Engineering, Nanyang Technological University, Singapore 639798, 2Department of Computer Science and Software Engineering, University of Melbourne, 3National ICT Australia, Victoria Research Lab, Victoria 3010 and 4School of Computer Science and IT, RMIT University, Melbourne 3001, Australia

Received on February 8, 2009; revised on June 7, 2009; accepted on June 12, 2009

Advance Access publication June 17, 2009
Associate Editor: Alfonso Valencia

ABSTRACT

Summary: The shorter and vastly more numerous reads produced by second-generation sequencing technologies require new tools that can assemble massive numbers of reads in reasonable time. Existing short-read assembly tools can be classified into two categories: greedy extension-based and graph-based. While the graph-based approaches are generally superior in terms of assembly quality, the computer resources required for building and storing a huge graph are very high. In this article, we present Taipan, an assembly algorithm which can be viewed as a hybrid of these two approaches. Taipan uses greedy extensions for contig construction but at each step realizes enough of the corresponding read graph to make better decisions as to how assembly should continue. We show that this approach can achieve an assembly quality at least as good as the graph-based approaches used in the popular Edena.

Availability and Implementation: Source code in C running on Linux is freely available at http://taipan.sourceforge.net

Contact: asbschmidt@ntu.edu.sg

1 INTRODUCTION

Recently, several second-generation DNA sequencing technologies have been introduced. Compared to traditional Sanger shotgun techniques these technologies can generate a larger amount of read data at lower cost (Mardis 2008; Pop and Salzberg, 2008). Examples of already available second-generation technology include products from Illumina, 454 Life Sciences and Applied Biosystems. However, the length of produced reads is significantly shorter than for the classical Sanger technique. For example, the Illumina Genome Analyzer has been introduced. Compared to traditional Sanger shotgun techniques these technologies can generate a larger amount of read data at lower cost (Mardis 2008; Pop and Salzberg, 2008). Examples of already available second-generation technology include products from Illumina, 454 Life Sciences and Applied Biosystems. However, the length of produced reads is significantly shorter than for the classical Sanger technique. For example, the Illumina Genome Analyzer can generate 1.5 Gb of sequence data in a single-end 60h run with a read length of around 36. Several new assemblers that are tailored towards short reads and high coverage have recently been introduced. SSAKE (Warren et al., 2007), VCake (Jeck et al., 2007) and SHARCXS (Dohm et al., 2007) are based on the greedy extension approach, while newer tools like Edena (Hernandez et al., 2008), Velvet (Zerbino and Birney, 2008) and Euler-SR (Chaisson and Pevzner, 2008) use more accurate graph-based approaches. Graph-based approaches can be further distinguished into the overlap graph approach (Edena) and the De Bruijn graph approach (Velvet, Euler-SR). Previous work (Hernandez et al., 2008) has already shown that graph-based approaches achieve superior assembly performance compared to greedy extensions. In this article we present Taipan, a fast hybrid short read fragment assembly algorithm. Taipan uses a combination of the greedy extension and the overlap graph method. Our performance evaluation using real Illumina datasets shows that Taipan can achieve assembly qualities comparable to the graph-based approaches within a reasonable execution time.

2 METHODS

Input to Taipan is a multi-set of reads $R$ of length $l$ each (consisting of the original reads plus their reverse complements), the minimal overlap parameter $k < l$, and a threshold $T$. $R$ is stored in a hash table. This hash table allows efficient processing of queries of the form $get\_overlapping\_reads(S, k, T)$ for a DNA string $S$ and $k \in \{1, 2, \ldots\}$ which return all reads in $R$ whose prefix $t$ matches the suffix $s$ of $S$. Taipan assembles a new contig by choosing a read from $R$ as a seed. This seed is iteratively extended in $3'$ direction by one base at a time until there are either insufficient overlapping reads or a repeat is found. Subsequently, the same algorithm is used in $5'$ direction.

The algorithm to extend contig $S$ in $3'$ direction by a single base works as follows. A set of overlapping reads is retrieved from the hash table by calling $get\_overlapping\_reads(S, k, T)$ for each $t \in \{1, \ldots\}$. Afterwards the directed overlap graph $G_w(O, E)$ is constructed, where $G_w$ is associated with a sequence $w$ if suffix $t$ of $w$ matches prefix $t$ of some $t \in \{1, \ldots\}$. $G_w(O, E)$ is then built by removing all associative edges from $G_w(O, E)$. $O = \{O_1, \ldots, O_T\}$, $E = \{E_1, \ldots, E_T\}$, the set $P$ consisting of all vertex-disjoint paths of $G_w$ is determined. $P$ is analyzed to determine the single-base extension of $S$ using threshold $T$ as follows: (a) if $P$ contains at least two paths $p_1, p_2$ with length $|p_1|, |p_2| \geq T$, then a repeat is found and extension of $S$ is terminated; (b) if $P$ contains exactly one path $p$ with length $|p| \geq T$, then $S$ is extended by the first nucleotide of the first path in $p$; (c) if $P$ contains no path $p$ with length $|p| \geq T$, then there are insufficient overlapping reads and extension of $S$ is terminated; (d) if none of the above cases apply, $G_w$ is enlarged along each path $p$ and the rules (a)–(d) are recursively applied (up to a maximum of $T$ steps) to the enlarged graph. After the assembly of a new contig $S$ all reads that exactly match this contig are removed from $R$ and the hash table. The implementation of Taipan gains efficiency by the fact that $O$ only slightly changes for subsequent extensions and by using a sorted trie data structure to represent $G_w$. Choice of seeds can affect assembly results. Taipan selects an unassembled read with highest occurrence as seed in order to minimize extensions of reads containing sequencing errors.

3 RESULTS

We have evaluated the performance of Taipan in terms of both runtime and assembly quality using two real Solexa/Illumina
datasets and compared the results to Edena and Velvet (Table 1). The datasets have been obtained experimentally by (Hernandez et al., 2008) and (Dohm et al., 2007) using the Illumina Genome Analyzer for sequencing the \textit{Staphylococcus aureus} strain MW2 (Aureus) and \textit{Helicobacter acinonychis} strain Sheeba (Helicobacter). Respective read files were downloaded from http://www.genomic.ch/edena.php and http://sharcgs.molgen.mpg.de/download.shtml. A summary of both datasets is provided in the format: genome GenBank ID, genome length, coverage, read length, number of reads:

- \textit{Staphylococcus}: NC\_003923, NC\_005011, 2.82 Mb, 48×, 35, 3.80M
- \textit{Helicobacter}: NC\_008299, NC\_008230, 1.55 Mb, 284×, 36, 12.3M

We have executed Taipan, Edena 2.1.1 (for Linux-64), and Velvet 0.7.31 for each dataset on a workstation with Quadcore Xeon E5440 2.83GHz CPU and 16 GB RAM running Linux CenOS 5.2-64 bit. Table 1 reports the assembly performance using the best tested parameter setting for each tool. The best tested parameter settings for the Aureus dataset are:

- \textit{Taipan}: k=19, T=8;
- \textit{Edena}: minimum overlap = 21 (strict)/20 (non-strict);
- \textit{Velvet}: hash size = 21, coverage cutoff = 7.

The best tested parameter settings for the Helicobacter dataset are:

- \textit{Taipan}: k=27, T=18;
- \textit{Edena}: minimum overlap = 27 (strict)/26 (non-strict);
- \textit{Velvet}: hash size = 27, coverage cutoff = 8.

All tools use a minimal contig length of 100 bp. Produced contigs are aligned to the respective circular reference genomic and plasmid sequences using Exonerate (Slater and Birney, 2005). A contig is classified as correct if it fully aligns with a base similarity of at least 98%. Otherwise, it is classified as misassembled. Calculation of N50, N80, average length, maximum length, and coverage only considers contigs classified as correct. Measured runtime is the sum of preprocessing and assembly time. For the Aureus dataset, strict mode Edena produces no misassembled contigs but has the lowest N50 value. Taipan produces 16 misassembled contigs but has the highest N50 value and the shortest runtime. Furthermore, the total size of Taipan’s misassembled contigs (2.8 kb) is significantly shorter compared to Velvet (14.5 kb) and Edena non-strict (36.6 kb).

For the Helicobacter dataset, both Taipan and Edena strict produce no misassembled contigs, whereby Taipan has a clearly higher N50 value. Edena non-strict and Velvet produce N50 values close to Taipan but generate a single long misassembled contig and two short misassembled contigs, respectively. Taipan remains faster than Edena, however, due to the usage of a large threshold parameter (T=18), Taipan is slower than Velvet.

In conclusion, Taipan is an efficient and accurate tool for de-novo genome assembly of short reads generated by high-throughput sequencing machines. The utilized algorithm shows that simple operations that do not require the whole overlap graph can achieve results similar to methods that construct the whole graph. Taipan’s results for both the medium-coverage (48×) Aureus dataset and the high-coverage Helicobacter dataset (284×) are competitive in terms of speed and quality compared to Edena as well as Velvet. We are planning to improve Taipan further by running read error correction beforehand and including paired-read data information.

Conflict of Interest: none declared.

REFERENCES


