BMC Bioinformatics

BioMed Central

Poster presentation

Open Access

Robust consensus computation

Tobias Rausch*[1,2], Anne-Katrin Emde[1,2] and Knut Reinert[2]

Address: [1]International Max Planck Research School for Computational Biology and Scientific Computing, Ihnestr. 63-73, 14195 Berlin, Germany and [2]Algorithmische Bioinformatik, Institut für Informatik, Takustr. 9, 14195 Berlin, Germany

Email: Tobias Rausch* - rausch@mi.fu-berlin.de

* Corresponding author

from Fourth International Society for Computational Biology (ISCB) Student Council Symposium
Toronto, Canada. 18 July 2008

Published: 30 October 2008

BMC Bioinformatics 2008, **9**(Suppl 10):P4 doi:10.1186/1471-2105-9-S10-P4

This abstract is available from: http://www.biomedcentral.com/1471-2105/9/S10/P4

Introduction

High-throughput sequencing technologies with short read data pose a new challenge to the current three-phase assembly methodology: Overlap-Phase, Layout-Phase, and Consensus-Phase. We describe a new consensus method that is robust in the face of high coverage, shorter reads, and genomic variation.

Methods

Given an initial layout of the reads, we generate a consensus sequence and a multi-read alignment with the following protocol: (1) Computation of all necessary (with respect to the layout) pairwise overlap alignments. (2) Extraction of all gapless alignment segments and generation of a segment-based weighted overlap graph (see Fig. 1). Conflicts between segment matches are resolved using a novel multiple segment match refinement algorithm [1]. (3) An adjustment of the edge-weights using a variant of the triplet extension pioneered in the T-Coffee package [2]. By means of the triplet extension we increase the weights of clique-edges within the overlap graph and thus, these edges are more likely to be chosen in the subsequent progressive alignment stage. (4) A progressive graph-based alignment of the reads using the heaviest common subsequence algorithm and a guide tree computed from the pairwise alignment scores. Note that our algorithm does not align single nucleotides but the segments identified in the overlap alignment phase. This ensures that columns with genetic variation (e.g., SNPs) are preserved. (5) Output of the multi-read alignment, the gapped consensus and all positioned reads with their respective deltas. The output can be visualized in Hawkeye [3] (see Fig. 2).

Results

We used a read simulator and real data from the NCBI trace archive to evaluate our consensus tool. The main parameters of the read simulator are the source sequence length, the average read length, the number of reads and the error rate per base call. In addition, multiple haplotypes can be simulated. Two further parameters, namely the number of SNPs and the number of indels, specify the genetic variation randomly introduced into these haplo-

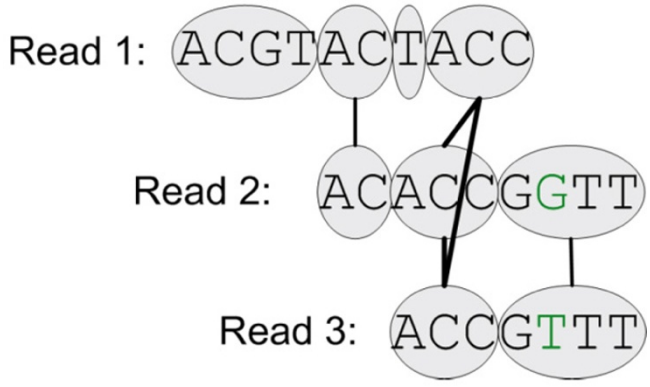

Figure 1
A segment-based alignment graph of three reads. The green colored SNP is embedded in a segment and a clique is highlighted in bold font.

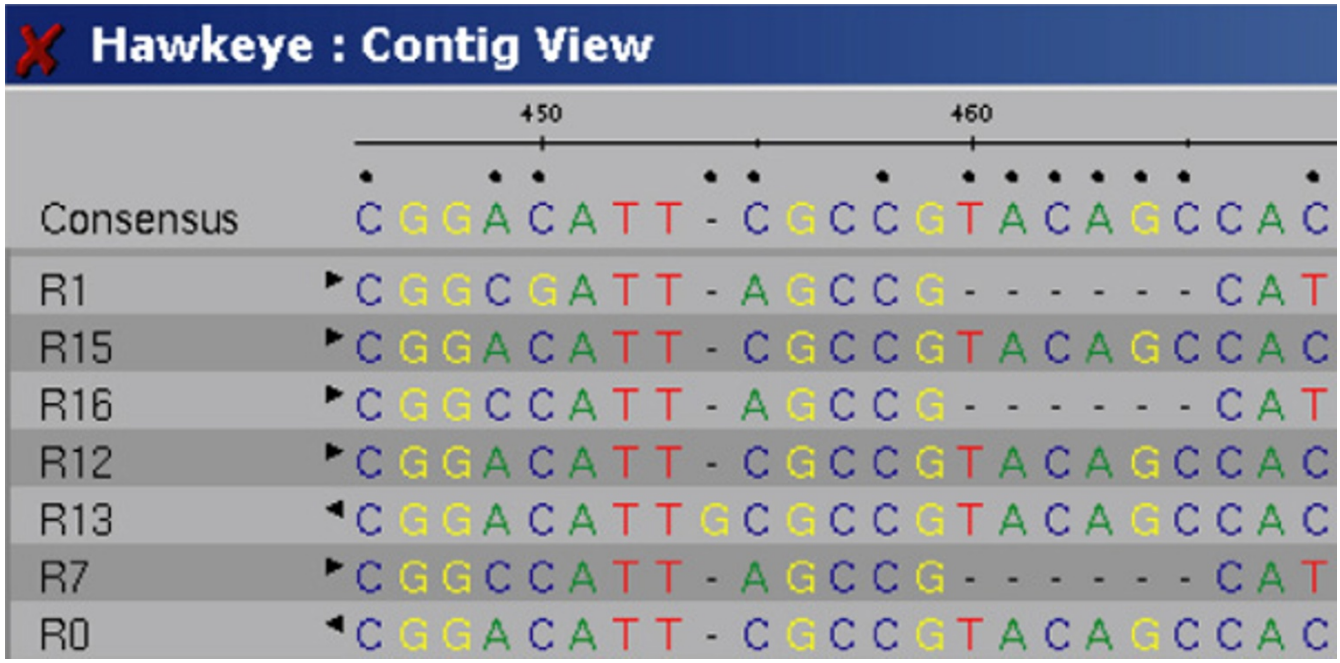

Figure 2
Section of a multi-read alignment with an indel and 2 SNP columns "A/C" and "T/C". Read names and read orientation are shown on the left, the consensus is shown in the top row.

types. We performed two experiments: (1) Given a source sequence length of 10000, we simulated reads under different settings. The read length varied from 35 to 200, the coverage from 20× to 50× and the error rate per base call from 2% to 4%. In all cases, the computed gap-free consensus matched the simulated source sequence in each position with coverage > 2. (2) Given two haplotypes each of length 10000 with 100 SNPs and 5 Indels, we simulated reads of length 200, coverage 20 and 4% error rate. We then manually inspected the multi-read alignment with Hawkeye to evaluate the consensus in case of genetic variation (see Fig. 2).

Conclusion

The results on simulated data are encouraging and preliminary results on real data show that our consensus quality is comparable to other tools. It remains to be shown that our program outperforms other tools in diffficult settings, namely high coverage and short, error-prone read data. The consensus tool is part of the SeqAn library [4] http://www.seqan.de and the read simulator is available on request: rausch@inf.fu-berlin.de.

References

1. Rausch T, Emde AK, Weese D, Döring A, Notredame C, Reinert K: **Segment-based multiple sequence alignment.** *Bioinformatics* 2008, **24(16):**i187-192.
2. Notredame C, Higgins D, Heringa J: **T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment.** *Journal of Molecular Biology* 2000, **302:**205-217.
3. Schatz M, Phillippy A, Shneiderman B, Salzberg S: **Hawkeye: an interactive visual analytics tool for genome assemblies.** *Genome Biology* 2007, **8(3):**R34.
4. Döring A, Weese D, Rausch T, Reinert K: **SeqAn – An efficient, generic C++ library for sequence analysis.** *BMC Bioinformatics* 2008, **9:**11.