APPLICATION NOTE Genomics & Informatics Vol. 7(1) 46-48, March 2009

A Simple Java Sequence Alignment Editing Tool for Resolving Complex Repeat Regions

Seong-Il Ham[3,4], Kyung-Eun Lee[2,4] and Hyun-Seok Park[1,2]*

[1]Institute of Bioinformatics, Macrogen Inc., Seoul 153-023, Korea, [2]Department of Computer Science, Ewha Womans University, Seoul 120-750, Korea, [3]Department of Architectural Engineering, Yonsei University, Seoul 120-749, Korea, [4]Department of Biochemistry and Molecular Biology, College of Medicine, Seoul National University, Seoul 110-799, Korea

Abstract

Finishing is the most time-consuming step in sequencing, and many genome projects are left unfinished due to complex repeat regions. Here, we have developed BACContigEditor, a prototype shotgun sequence finishing tool. It is essentially an editor that visualizes assemblies of shotgun sequence fragment reads as gapped multiple alignments. The program offers some flexibility that is needed to rapidly resolve complex regions within a working session. The sole purpose of the release is to promote collaborative creation of extensible software for fragment assembly editors, foster collaborative development, and reduce barriers to initial tool development effort. We describe our software architecture and identify current challenges. The program is available under an Open Source license.

Availability: BACContigEditor is a public domain software. The software is a pre-alpha release and provides only basic fragment assembly editing features. You may download this program from the source (http://www.sourceforge.net/BACContigEditor/) for use on your computer. A documented API makes adding new algorithms, visualization modes, and sequence features easy.

Keywords: fragment assembly, genome sequencing, whole genome fragment assembly

Introduction

High-throughput methods for genome sequencing have yielded piles of genomic sequences. However, genome assembly is a traditionally open problem, as complicated parts of sequenced genomes tend to be left unfinished to a large extent. It means that current shotgun sequencing assembly programs (Phrap: http://www.phrap.org/, AMOS: http://www.tigr.org/software/AMOS/, Celera Assembler: http://sourceforge.net/projects/wgs-assembler, and Arachne: http://www.broad.mit.edu/wga/arachnewiki/) are not yet designed to handle long stretches of repeated DNA in the target sequence. One of the difficulties in repeat classification is that many repeats represent mosaics of subrepeats (Bailey *et al*., 2002).

Repeats cause large artificial rearrangements of contigs due to misassembled repeat regions (Fig. 1). The presence of repeated regions in the target sequence is thus the key problem in shotgun sequencing (She *et al*., 2004; Eichler *et al*., 2004). A successful strategy for solving the problem of short repeat regions has been the use of mate pairs (Myers *et al*., 2000). However, this strategy fails when nearly identical repeats are organized in tandem stretches that are longer than multiple shotgun fragment insert lengths. Still, these problems of the common assembly methods place a heavy burden on the bioinformaticians who work on the finishing stage of sequencing projects (Edwards *et al*., 1990).

Besides these traditional difficulties, the recent technological divergence of high-throughput sequencing platforms, including 454 Life Sciences (Roche) (http://www.454.com/), Solexa of the Illumina Genome Analysis System (http://www.illumina.com), and Applied Biosystems SOLiD Sequencing (http://ww.appliedbiosystems.

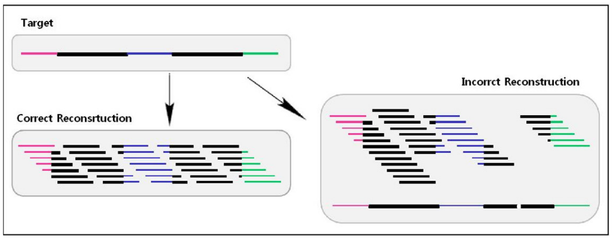


Fig. 1. An example of artificial rearrangement: the repeat copies are piled on top of each other and merged into alignments of high coverage.

*Corresponding author: E-mail neo@ewha.ac.kr
Tel +82-2-3277-2831, Fax +82-2-3277-2306

Accepted 10 February 2009

com), causes thorny issues. The fragment assembly and editing process again is becoming the most critical and challenging problem in the area of genomic research.

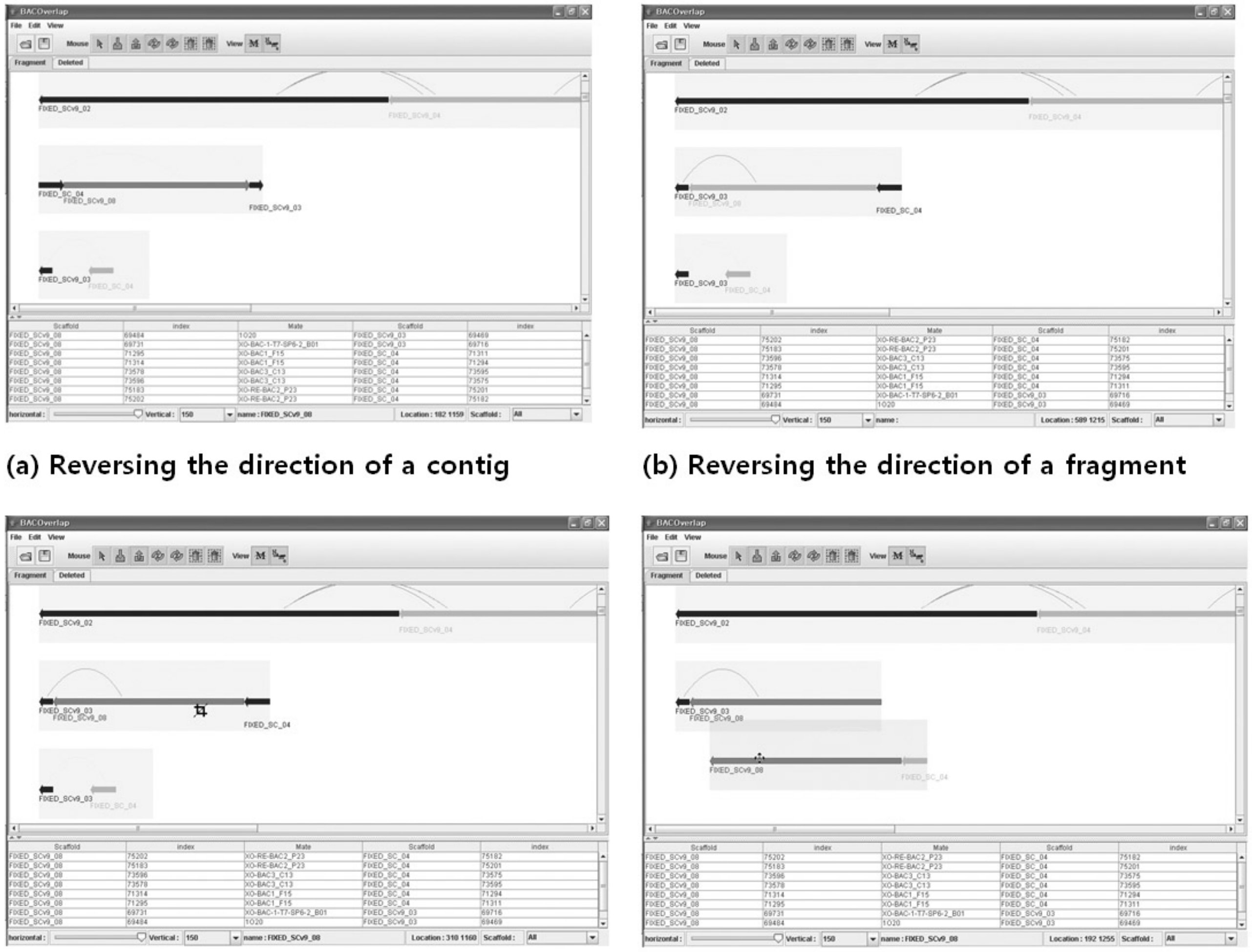
Thus, we present a pre-alpha release of BACContigEditor 0.1, an assembly editing and visualization tool that is specifically designed for manual analysis and finishing of repeated regions. The software has been used only to assemble BAC shotgun sequences.

At this stage, the purpose of the release of BACContigEditor is to enhance the current version of our software. In this way, we believe that we can collaborate with other researchers to develop a better version of our software. The source code is available from the authors under an Open Source license.

Overview of the System

BACContigEditor is a simple sequence alignment editing tool, developed in Java. It was originally developed to finish BAC shotgun sequencing projects, but the program could be easily extended to whole genome projects. It deals with visualization of problematic assembly locations, providing detailed information and allowing rapid decision-making to make necessary corrections. The user can move sequences using drag-drop, realign, cut-paste, and add-remove features. The user interface is a front end to a database, and changes are propagated to the database.

The assemblies can be produced by an assembler that produces the supported .frg-file format, translated from .ace-files from Phrap (http://www.phrap.org), and can be exported to the same format after repeat analysis and resolution. In addition to the read sequences,



Scaffold	index	Mate	Scaffold	index
FIXED_SCv9_08	69484	1020	FIXED_SCv9_03	69469
FIXED_SCv9_08	69731	XO-BAC-1-T7-SP6-2_B01	FIXED_SCv9_03	69716
FIXED_SCv9_08	71295	XO-BAC1_F15	FIXED_SC_04	71311
FIXED_SCv9_08	71314	XO-BAC1_F15	FIXED_SC_04	71294
FIXED_SCv9_08	73578	XO-BAC3_C13	FIXED_SC_04	73595
FIXED_SCv9_08	73596	XO-BAC3_C13	FIXED_SC_04	73575
FIXED_SCv9_08	75183	XO-RE-BAC2_P23	FIXED_SC_04	75201
FIXED_SCv9_08	75202	XO-RE-BAC2_P23	FIXED_SC_04	75182

horizontal : Vertical : 150 name : FIXED_SCv9_08 Location : 182 1159 Scaffold : All

(a) Reversing the direction of a contig

Scaffold	index	Mate	Scaffold	index
FIXED_SCv9_08	75202	XO-RE-BAC2_P23	FIXED_SC_04	75182
FIXED_SCv9_08	75183	XO-RE-BAC2_P23	FIXED_SC_04	75201
FIXED_SCv9_08	73596	XO-BAC3_C13	FIXED_SC_04	73575
FIXED_SCv9_08	73578	XO-BAC3_C13	FIXED_SC_04	73595
FIXED_SCv9_08	71314	XO-BAC1_F15	FIXED_SC_04	71294
FIXED_SCv9_08	71295	XO-BAC1_F15	FIXED_SC_04	71311
FIXED_SCv9_08	69731	XO-BAC-1-T7-SP6-2_B01	FIXED_SCv9_03	69716
FIXED_SCv9_08	69484	1020	FIXED_SCv9_03	69469

horizontal : Vertical : 150 name : Location : 589 1215 Scaffold : All

(b) Reversing the direction of a fragment

Scaffold	index	Mate	Scaffold	index
FIXED_SCv9_08	75202	XO-RE-BAC2_P23	FIXED_SC_04	75182
FIXED_SCv9_08	75183	XO-RE-BAC2_P23	FIXED_SC_04	75201
FIXED_SCv9_08	73596	XO-BAC3_C13	FIXED_SC_04	73575
FIXED_SCv9_08	73578	XO-BAC3_C13	FIXED_SC_04	73595
FIXED_SCv9_08	71314	XO-BAC1_F15	FIXED_SC_04	71294
FIXED_SCv9_08	71295	XO-BAC1_F15	FIXED_SC_04	71311
FIXED_SCv9_08	69731	XO-BAC-1-T7-SP6-2_B01	FIXED_SCv9_03	69716
FIXED_SCv9_08	69484	1020	FIXED_SCv9_03	69469

horizontal : Vertical : 150 name : FIXED_SCv9_08 Location : 310 1160 Scaffold : All

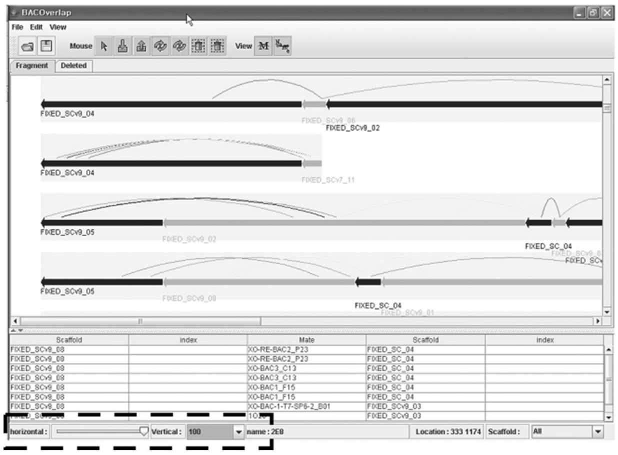
(c) Separating a contig

Scaffold	index	Mate	Scaffold	index
FIXED_SCv9_08	75202	XO-RE-BAC2_P23	FIXED_SC_04	75182
FIXED_SCv9_08	75183	XO-RE-BAC2_P23	FIXED_SC_04	75201
FIXED_SCv9_08	73596	XO-BAC3_C13	FIXED_SC_04	73575
FIXED_SCv9_08	73578	XO-BAC3_C13	FIXED_SC_04	73595
FIXED_SCv9_08	71314	XO-BAC1_F15	FIXED_SC_04	71294
FIXED_SCv9_08	71295	XO-BAC1_F15	FIXED_SC_04	71311
FIXED_SCv9_08	69731	XO-BAC-1-T7-SP6-2_B01	FIXED_SCv9_03	69716
FIXED_SCv9_08	69484	1020	FIXED_SCv9_03	69469

horizontal : Vertical : 150 name : FIXED_SCv9_08 Location : 192 1255 Scaffold : All

(d) Concatenating two contigs

Fig. 2. Contig realigning movements.



Scaffold	index	Mate	Scaffold	index
FIXED_SCv9_08		XO-RE-BAC2_P23	FIXED_SC_04	
FIXED_SCv9_08		XO-RE-BAC2_P23	FIXED_SC_04	
FIXED_SCv9_08		XO-BAC3_C13	FIXED_SC_04	
FIXED_SCv9_08		XO-BAC3_C13	FIXED_SC_04	
FIXED_SCv9_08		XO-BAC1_F15	FIXED_SC_04	
FIXED_SCv9_08		XO-BAC1_F15	FIXED_SC_04	
FIXED_SCv9_08		XO-BAC-1-T7-SP6-2_B01	FIXED_SCv9_03	
FIXED_SCv9_08		[illegible]	FIXED_SCv9_03	

horizontal : Vertical : 100 name : 2E8 Location : 333 1174 Scaffold : All

Fig. 3. Zooming features.

different features and mate pairs are visualized in the editor according to the preferences of the user. Sequence features can be present in the input assembly files or be added during the finishing process.

The finishing of repeated regions is made less complicated using BACContigEditor by providing a couple of key features: the availability of a birds-eye view of all of the reads in a contig and manual editing operations. By zooming out, the user gets an overview of the contig and its length and depth and can get a clear picture of the general properties of the region at once. By using drag-drop, it is possible to sort reads into different repeat groups. Vertical and horizontal movement of reads is allowed, and realigning also can be performed (Fig. 2).

Software Release Stage

BACContigEditor has been implemented using Java and relies on Open Source software. Recently, state-of-the-art sequence alignment editing tools (Arner *et al*., 2006) have been developed, and compared with these tools, our tool is only a pre-alpha release. Generally, it needs to be improved with a user interface; with a rigid close-up view, the user can only view a small portion of the repeat region, and much scrolling is required to get a clear understanding of the region (Fig. 3).

However, we are developing utility programs to handle various formats, based on multiple sequencing platforms. Although these utility programs are not included in the current distribution, they will be extremely useful for finishing large sequencing projects in the future. Updates to these features will be available through the project web site (http://www.sourceforge.net/BACContigEditor/) at frequent intervals.

References

Arner, E., *et al*. (2006). DNPTrapper: an assembly editing tool for finishing and analysis of complex repeat regions. *BMC Bioinformatics* 20, 155.

Bailey, J., *et al*., (2002). Human-specific duplication and mosaic transcripts: The recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.* 70, 83-100.

Edwards, A., and Caskey, C.T. (1990). Closure strategies for random DNA sequencing methods. *A Companion to Methods in Enzymology* 3, 41-47.

Eichler, E.E., Clark, R.A., and She, X. (2004). An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat Rev Genet.* 5, 345-354.

Myers, E.W., *et al*. (2000). A wholegenome assembly of Drosophila. *Science* 287, 2196-204.

She, X., *et al*. (2004). Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* 431, 927-930.