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

Seong-II Ham^{3,4}, Kyung-Eun Lee^{2,4} and Hyun-Seok Park^{1,2}*

¹Institute of Bioinformatics, Macrogen Inc., Seoul 153-023, Korea, ²Department of Computer Science, Ewha Womans University, Seoul 120-750, Korea, ³Department of Architectural Engineering, Yonsei University, Seoul 120-749, Korea, ⁴Department 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

*Corresponding author: E-mail neo@ewha.ac.kr Tel +82-2-3277-2831, Fax +82-2-3277-2306 Accepted 10 February 2009

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.

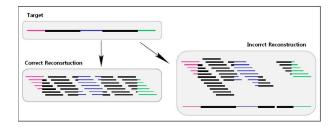


Fig. 1. An example of artificial rearrangement: the repeat copies are piled on top of each other and merged into alignments of high coverage.

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 BACContig-Editor 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.

A Simple Java Sequence Alignment Editing Tool 47

Overview of the System

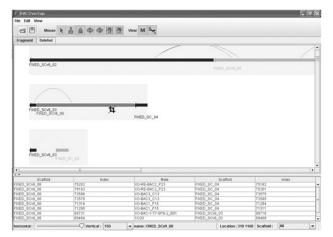
BACContigEditor is a simple sequence alignment editing tool, developed in Java. It was originally developed to finish BAC shotgun sequencing projects, but the program could be easily extended to whole genome projects. It deals with visualization of problematic assembly locations, providing detailed information 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,

Mouse I		View M Sing			
		/			1
XED_SCV9_02			FRED_BCV9_04		-
ED_SC_04					
ED_SC_04 FRED_SC	9_08	FORED_SCH9_03			
XED_SC_04 FRED_SCV	n_01	FRED_SCH_03			
		FDED_SCV9_03			
		FOED_SCV1_03			
NED_SCY0_D		FINED_SCH_03			•
NED_SCV9_03	ED_EC_04	Mate	Boarfold	inder	
NED_SCV9_03	ED_SC_04	Mate 1020	FOED_SCV9_03	69469	
NED_SCV9_03	80, 90, 94	1020 Male 1020 DO BACT-177.9F9-2.B01	FOED_SCV9_03	69469 69716	Dia Contractione
DED_SCV9_07	ED_SC_04	Mate 1020		69469	1)

(a) Reversing the direction of a contig

3



(c) Separating a contig

Fig. 2. Contig realigning movements.

	lew			10 10 10 10 10 10 10 10 10 10 10 10 10 1					
6	Mouse	* 4 4 4	> <> Bill B	Wew M Mag					
Fragment	Deleted								
				/	/				
							_	_	1
FDE	D_SCv9_02				1				
-									
FDOE	D_SCv9_03								
				FRED_SC_04					
		_							
FDE	D_8CV9_03	THED_SC_04							
	0_6Cv9_03	WED_SC_04							
	D_8Cv9_03	TKED_80_04							•
		10ED_8C_04	index	Mate		Scattold		index	•
ED_SC/4	Scaffold 00	75202	index	XO-RE-BAC2_P23	FOED_SC_0		75162	index	•
ED_SCVA	Scaffold 08	75202	index	X0-RE-BAC2_P23 X0-RE-BAC2_P23	FDED_SC_0		75201	index	•
ED_SCV9 ED_SCV9 ED_SCV9	Scaffold 00 00	75202 75183 73596	index	X0-RE-BAC2_P23 X0-RE-BAC2_P23 X0-BAC3_C13	F0ED_8C_0 F0ED_8C_0 F0ED_8C_0		75201	index	•
ED_SCV8 ED_SCV8 ED_SCV8 ED_SCV8	Scaffold 00 00 00 00	75202	index	X0-RE-BAC2_P23 X0-RE-BAC2_P23	FDED_SC_D FDED_SC_D FDED_SC_D FDED_SC_D		75201	index	•
- E0_504 E0_504 E0_504 E0_504 E0_504 E0_504	Scaffold 08 08 08 00 00 00	76202 75183 73596 73576 71514 71295	index	X0-RE-BAC2_P23 X0-RE-BAC2_P23 X0-BAC3_C13 X0-BAC3_C13 X0-BAC1_F15 X0-BAC1_F15	F0ED_9C_0 F0ED_9C_0 F0ED_9C_0 F0ED_9C_0 F0ED_9C_0 F0ED_9C_0		75201 73575 73595 71294 71311	Index	•
* *ED_SC/9 XED_SC/9 XED_SC/9 XED_SC/9 XED_SC/9 XED_SC/9 XED_SC/9	Scaffold 00 00 00 00 00 00 00 00 00 00	75202 75183 73596 73578 71314 71295 68731	index :	X0-RE-BAC2_P23 X0-RE-BAC2_P23 X0-BAC3_C13 X0-BAC3_C13 X0-BAC1_F15 X0-BAC1_F15 X0-BAC1_F15	F0ED_9C_0 F0ED_9C_0 F0ED_9C_0 F0ED_9C_0 F0ED_9C_0 F0ED_9C_0 F0ED_9C_0 F0ED_9C_0	03	75201 73575 73595 71294 71311 69716	index	•
Ļ	Scaffold 00 00 00 00 00 00 00 00 00 00	76202 75183 73596 73576 71514 71295	index	X0-RE-BAC2_P23 X0-RE-BAC2_P23 X0-BAC3_C13 X0-BAC3_C13 X0-BAC1_F15 X0-BAC1_F15	F0ED_9C_0 F0ED_9C_0 F0ED_9C_0 F0ED_9C_0 F0ED_9C_0 F0ED_9C_0	03	75201 73575 73595 71294 71311	index	•

(b) Reversing the direction of a fragment

Distribution None Market Preconscience Statistical Preconscience Statis Statis Statis </th <th>BACOverlap</th> <th></th> <th></th> <th></th> <th></th>	BACOverlap				
Tradework Defects Defects FreeD_Bood_82 FreeD_Bood_94 FreeD_Bood_93 FreeD_Bood_94 FreeD_Bood_94 FreeD_Bood_95	le Edit View				
Tradework Defects Defects FreeD_Bood_82 FreeD_Bood_94 FreeD_Bood_93 FreeD_Bood_94 FreeD_Bood_94 FreeD_Bood_95	en E Marrie h	上 余 命 命 節 節	Warner Mr. Stern		
PRED_SCH_92 PRED_		100 00 No No 190 19	i was in the		
NEED_BOX_04 O FRED_BOX_04 0 FRED_BOX_05 </th <th>ragment Deleted</th> <th></th> <th></th> <th></th> <th></th>	ragment Deleted				
NEED_BOX_04 O FRED_BOX_04 0 FRED_BOX_05 </td <td></td> <td></td> <td></td> <td></td> <td></td>					
NEED_BOX_04 O FRED_BOX_04 0 FRED_BOX_05 </td <td></td> <td></td> <td>/</td> <td></td> <td>/</td>			/		/
FIEL <th< td=""><td>FIXED_SCv9_02</td><td></td><td></td><td></td><td></td></th<>	FIXED_SCv9_02				
Factor None Factor Factor Factor					
Factorial Factorial 0 0 0					
Factorial Factorial 0 0 0					
Factorial Factorial 0 0 0					
Factorial Factorial 0 0 0					
Factorial Factorial 0 0 0		1			
Destination Destination Destination 0 <t< td=""><td>ENED SOLD DO</td><td></td><td></td><td></td><td></td></t<>	ENED SOLD DO				
PEC_BONG_BR PECO_BONG_BR PEC	FIRED_SCV9_08				
Red_Box_08 Party_acpt 50 500 (0) 500 (0) 500 (0) 50 500 (0) 7560 (0) 500 (0) 500 (0) 50 500 (0) 7560 (0) 500 (0) 500 (0) 50 500 (0) 7560 (0) 500 (0) 500 (0) 50 500 (0) 7560 (0) 500 (0) 500 (0) 50 500 (0) 7560 (0) 500 (0) 500 (0) 50 500 (0) 7571 (0) 500 (0) 7591 (0) 50 500 (0) 7571 (0) 500 (0) 7591 (0) 50 500 (0) 7571 (0) 500 (0) 7591 (0) 50 500 (0) 7571 (0) 500 (0) 7591 (0) 50 500 (0) 7571 (0) 500 (0) 7571 (0) 50 500 (0) 7571 (0) 500 (0) 7571 (0) 50 500 (0) 7571 (0) 500 (0) 9716 (0) 50 500 (0) 9811 (0) 500 (0) 9716 (0) 50 </td <td></td> <td></td> <td></td> <td></td> <td></td>					
Next Biolog, Bit Procession 1 1 1 <td></td> <td></td> <td></td> <td></td> <td></td>					
Red_Box_08 Party_acpt 50 500 (0) 500 (0) 500 (0) 50 500 (0) 7560 (0) 500 (0) 500 (0) 50 500 (0) 7560 (0) 500 (0) 500 (0) 50 500 (0) 7560 (0) 500 (0) 500 (0) 50 500 (0) 7560 (0) 500 (0) 500 (0) 50 500 (0) 7560 (0) 500 (0) 500 (0) 50 500 (0) 7571 (0) 500 (0) 7591 (0) 50 500 (0) 7571 (0) 500 (0) 7591 (0) 50 500 (0) 7571 (0) 500 (0) 7591 (0) 50 500 (0) 7571 (0) 500 (0) 7591 (0) 50 500 (0) 7571 (0) 500 (0) 7571 (0) 50 500 (0) 7571 (0) 500 (0) 7571 (0) 50 500 (0) 7571 (0) 500 (0) 9716 (0) 50 500 (0) 9811 (0) 500 (0) 9716 (0) 50 </td <td></td> <td></td> <td></td> <td></td> <td></td>					
Field (Sec.) Statistic	ENED SC				
Number Index Reals Fordbal Index 55 Ord, Bin 7562 00.5 Educ2 (F3) FOED, 50 (A) 95192 55 Ord, Bin 7562 50.5 (A) FOED, 50 (A) 95192 55 Ord, Bin 00.6 Educ2 (F3) FOED, 50 (A) 95193 50 Ord, Bin 00.6 Educ2 (F3) FOED, 50 (A) 75193 50 Ord, Bin 00.6 Educ2 (F3) FOED, 50 (A) 75193 50 Ord, Bin 75194 00.6 Educ2 (F3) FOED, 50 (A) 75193 50 Ord, Bin 75191 00.6 Educ2 (F3) FOED, 50 (A) 75193 50 Ord, Bin 75191 00.6 Educ2 (F3) FOED, 50 (A) 75193 50 Ord, Bin 75193 00.6 Educ2 (F3) FOED, 50 (A) 75191 50 Ord, Bin 75193 00.6 Educ2 (F3) FOED, 50 (A) 7511 50 Ord, Bin 75193 00.6 Educ2 (F4) FOED, 50 (A) 7511 50 Ord, Bin 75193 00.6 Educ2 (F4) FOED, 50 (A) 7511 50 Ord, Bin 75193 00.6 Educ2 (F4) <t< td=""><td></td><td></td><td>FDED_BC_04</td><td></td><td></td></t<>			FDED_BC_04		
Scaladid Index Real Fundad Index 55 (30-0) 19560 (30-66 gas(2) F2) F056 (30-64 gas(2) F2) F					
Scaladid Index Real Fundad Index 55 (30-0) 19560 (30-66 gas(2) F2) F056 (30-64 gas(2) F2) F					
Description Index Status Status Status Index 55 Oct allion 5500 (Status) <					
Description Index Status Status Status Index 55 Oct allion 5500 (Status) <		J			
D_1O_0 PSD2 DOEE_SO2_P32 PSD2_C CA PSD3 D_1O_0 <tdd_1o_0< td=""> D_1O_0 D_1O_0</tdd_1o_0<>		ladar	Make	Readfold	Index
Discription PISB Dock@ADC PISD					
Co. Co. 0 15191 Co. Bock 2 (*1) FPE0 (* C. 4) 1586 Co. Co. 0 11117 Co. 0 111117 Co. 0 <t< td=""><td>ED_8Cv9_08</td><td>75183</td><td>XO-RE-BAC2_P23</td><td>FOED_SC_04</td><td>75201</td></t<>	ED_8Cv9_08	75183	XO-RE-BAC2_P23	FOED_SC_04	75201
165_00<0,00				FDED_SC_04	73676
ED_SCH_08 F1285 VOEKCT_F15 F0ED_SC_04 71311 ED_SCH_08 681311 VOEKC1-ITI-SP6-2_B01 F0ED_SCH_00 681716 ED_SCH_08 68484 1020 F0ED_SCH_03 68489	ED_SCV9_00		PIO-BAC3_013	FDED_SC_04	
ED_SCV9_08 68731 00-BAC-1-T7-SP6-2_B01 F0ED_SCV9_03 69716 ED_SCV9_08 69484 1020 F0ED_SCV9_03 69469	ED_SCV9_08			FDED_SC_04	
ED_SC/4_08 69484 1020 FDED_SC/4_03 69489					
	ED_0073_00		1030		
tzontal: Vertical: 150 v name: FDED SCv9 08 Location: 192 1255 Scaffold: All		Vertical: 150			

(d) Concatenating two contigs

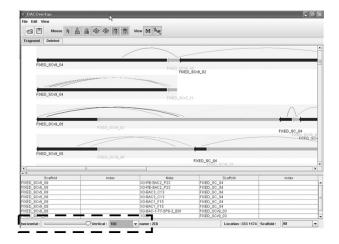


Fig. 3. Zooming features.

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-ofthe-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.