BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 16 2004, pages 2865–2866
doi:10.1093/bioinformatics/bth302

LUCY2: an interactive DNA sequence quality trimming and vector removal tool

Song Li[1] and Hui-Hsien Chou[1,2,*]

[1]Department of Computer Science and [2]Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA

Received on March 13, 2004; revised on April 22, 2004; accepted on April 23, 2004
Advance Access publication May 6, 2004

ABSTRACT

Summary: LUCY2 is a raw DNA sequence trimming and visualization tool based on the popular command-line LUCY1. Users can change parameters, trim multiple sequences and visualize the results within an integrated, easy-to-use graphical user interface. LUCY2 is designed specifically for non-programmers to use, and is currently available on Windows, Linux and MacOS X. Source code is also available for porting to the other platforms.

Availability: LUCY2 is distributed under the GNU General Public License and can be downloaded from www.complex.iastate.edu

Contact: lucy2@www.complex.iastate.edu

Most of the bioinformatic tools operating on DNA data assume that the sequences are trustworthy. However, raw data obtained from sequencing machines often violate this assumption. The tool LUCY1 was created previously to solve the raw data quality assurance problem (Chou and Holmes, 2001). Since its release, LUCY1 has become popular among genome sequencing centers. TIGR used it almost exclusively since 1998 (M.Holmes, personal communication), and several other projects used or considered LUCY1 in their genomic data processing pipeline (Qiu *et al*., 2003; Sorek and Safer, 2003; Venter *et al*., 2003; Waldbieser *et al*., 2003). Currently, there are over 1000 optionally registered users of LUCY1 (M.Holmes, personal communication).

LUCY1 was originally designed for high-throughput DNA sequencing centers to easily 'plug-in' to their data streams; therefore, it is operated exclusively through a command-line interface. This style of operation was not suitable to individuals who may just like to trim a few sequences, for the following reasons:

- LUCY1 has considerable number of parameters. There are 23 individual parameters and up to 20 window size and maximum error pairs that could influence its outcome.
- Users need to specify input/output files as command-line arguments, but often find it difficult to locate the path leading to their data files in a graphical user interface (GUI) environment.
- Users cannot see the results immediately after the computation. LUCY1 simply saves the results to a text file, which requires considerable efforts to understand them.
- LUCY1 does not physically remove bad sequence regions. This is useful for genome assembly purposes, but often end-users wish to completely delete low-quality regions before submission to GenBank; they need to use a companion program to trim the sequences.

Overall, it is inconvenient for users to operate with LUCY1 command-line interface and to memorize its parameters. We believe it is necessary to develop a GUI-based LUCY2 to provide the full functionality of LUCY1 in a more user-friendly manner.

The user interface of LUCY2 can be seen in Figure 1. Loaded sequence and quality files are displayed in tab-delimited panels. Quality values are reflected by the background color of the DNA characters; the higher a quality value, the lighter its correspondent character background. This gives users an immediate impression of the sequence quality. LUCY2 is capable of loading multiple files and batch processing them. A viewing panel is created for each loaded file and its correspondent panel tab indicates the sequence file name. All viewing panels also display sequence names and a position ruler. Users can identify a base position by pointing their mouse to the base.

As Figure 1 also shows, LUCY2 provides a much friendlier interface to set its parameters. When a parameter field is selected, a helping message will be shown in the lower part of the dialog to explain the parameter. A reasonable default value is always provided so users do not always need to change their parameters. If some parameter values are changed, LUCY2 will memorize them. The next time LUCY2 is restarted, the memorized values will be loaded as the default values. This feature makes it easier for users to work with their favorite set of parameters. Users can reset LUCY2 back to use the original factory default parameter values at any time.

*To whom correspondence should be addressed.

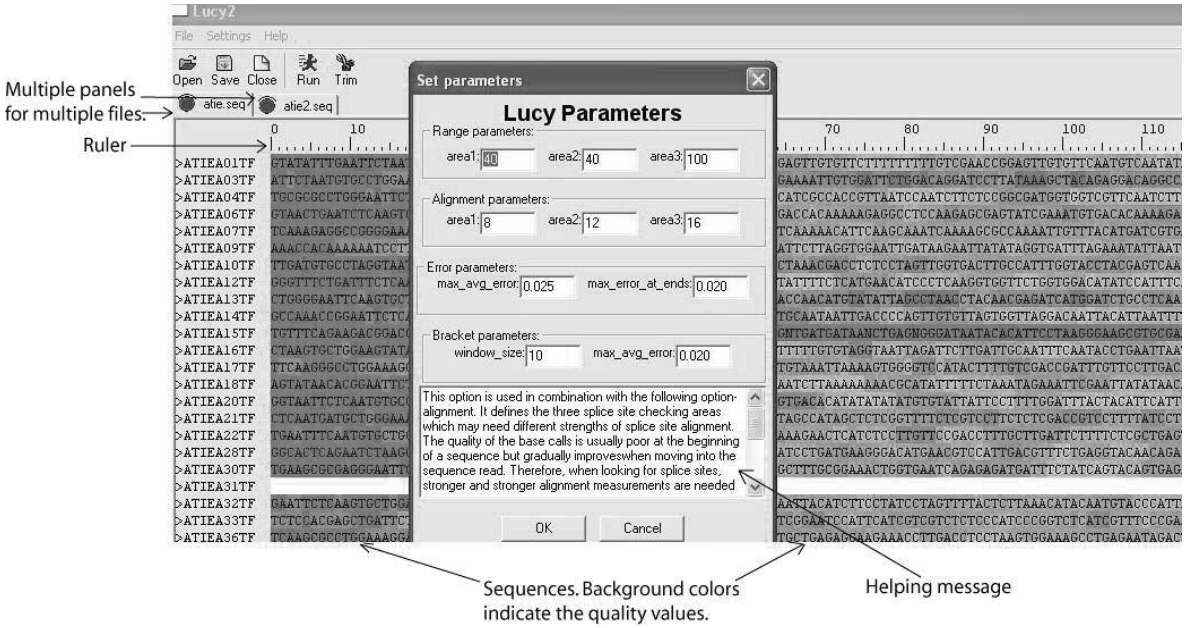

Fig. 1. The GUI of LUCY2.

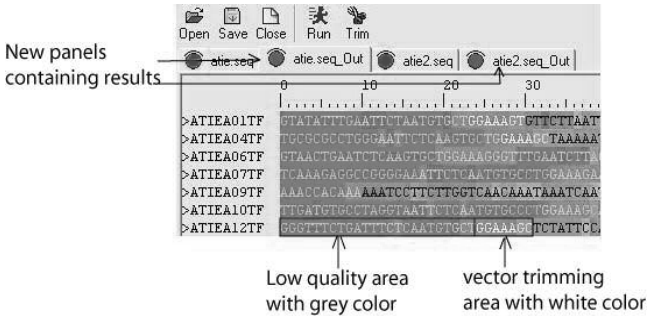

Fig. 2. The LUCY2 output panel.

Sequence trimming starts by clicking on the Run button in the toolbar. After that, the computation results are shown in a new 'Out' panel. Low-quality regions found by LUCY2 will be indicated by gray text color, and vector regions found by LUCY2 will be shown as white text. Trimming results in the output panel can be easily compared with the original sequences in the input panel because scrolling is synchronized between the two panels (Fig. 2). Users can determine what actions to take next: re-trim with a new parameter set; permanently remove the bad regions by clicking the 'Trim' button; or save the results to files with good regions indicated in the FASTA headers. All saved sequence files also include their associated quality files just like what LUCY1 would create.

ACKNOWLEDGEMENTS

We thank Michael Holmes of TIGR (the co-author of LUCY1) for his enthusiastic support of the creation of LUCY2, and for maintaining LUCY1 source code. This project is supported by the NIH grant 4R33GM066400.

REFERENCES

Chou,H.-H. and Holmes,M.H. (2001) DNA sequence quality trimming and vector removal. *Bioinformatics*, **17**, 1093–1104.

Qiu,F., Guo,L., Wen,T.-J., Liu,F., Ashlock,D.A. and Schnable,P.S. (2003) DNA sequence-based "Bar codes" for tracking the origins of expressed sequence tags from a maize cDNA library constructed using multiple mRNA sources. *Plant Physiol.*, **133**, 475–481.

Sorek,R. and Safer,H.M. (2003) A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.*, **31**, 1067–1074.

Venter,J.C., Levy,S., Stockwell,T., Remington,K. and Halpern,A. (2003) Massive parallelism, randomness and genomic advances. *Nat. Genet.*, **33**, 219–227.

Waldbieser,G.C., Bilodeau,A.L. and Nonneman,D.J. (2003) Complete sequence and characterization of the channel catfish mitochondrial genome. *DNA Seq.*, **14**, 265–277.