Nucleic Acids Research, Vol. 19, No. 23 6481–6483

A trace display and editing program for data from fluorescence based sequencing machines

Timothy Gleeson and LaDeana Hillier[1],*
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK and [1]Washington University School of Medicine, Department of Genetics, Box 8232, 4566 Scott Avenue, St Louis, MO 63110, USA

Received August 26, 1991; Revised and Accepted October 30, 1991

ABSTRACT

'Ted' (*Trace editor*) is a graphical editor for sequence and trace data from automated fluorescence sequencing machines. It provides facilities for viewing sequence and trace data (in top or bottom strand orientation), for editing the base sequence, for automated or manual trimming of the head (vector) and tail (uncertain data) from the sequence, for vertical and horizontal trace scaling, for keeping a history of sequence editing, and for output of the edited sequence. Ted has been used extensively in the *C.elegans* genome sequencing project, both as a stand-alone program and integrated into the Staden sequence assembly package, and has greatly aided in the efficiency and accuracy of sequence editing. It runs in the X windows environment on Sun workstations and is available from the authors. Ted currently supports sequence and trace data from the ABI 373A and Pharmacia A.L.F. sequencers.

INTRODUCTION

Time involved in sequence editing is extensive, and anything easing that burden will improve the efficiency of any major sequencing project. Having sequence and trace data available online in easily-manipulable form is invaluable. Ted (a Trace-EDitor) was developed to fill this role in the *C.elegans* genome sequencing project (1). Ted was originally developed as a stand-alone module for trace editing; portions of it were later incorporated into the sequence assembly program, xdap (2), for viewing of trace data.

METHODS

Computing design and implementation. When designing ted, we had a number of specific computing goals in mind including portability and adaptability. For portability, we chose to write ted in ANSI C using the X windowing system and the Xaw toolkit. X provides basic capabilities for the creation and use of windows, and the toolkit contains a number of prepackaged components, such as the 'sliders' used for scrolling. X also allows site, user and per-run defaults to be set. Adaptability is also an important goal since we are providing a new function to research groups who are constantly adding new requirements.

Stylistically, we have followed an 'Abstract Data Type' discipline. In this discipline, a program is split into a number of modules which provide separate, well-defined functions. We separate the interface of a module from its implementation. For example, a unified internal sequence format is used. This can store a varying amount of information. However, there is a clear and simple interface by which the rest of the program accesses this module. Such a style is not well supported by C, but its adoption has been very successful. The addition of new sequencing machines, and thus new external data formats, may cause some changes in the internal representation of the sequence but should not affect the rest of the program.

Ted accepts a large number of optional command line arguments, many of which can also be specified as system defaults. This supports a mode of working whereby ted is invoked not directly by the user but instead by a script or another application which supplies arguments appropriate to the editing task.

Graphical interface. Ted currently accepts data from two fluorescence based sequencing machines, the Pharmacia A.L.F. and the ABI 373A. The sequencing machine data consists of four traces of fluorescence levels together with the machine's interpretation, which is a sequence of bases. Ted displays the traces and the machine-generated base list. A second, initially identical, list of bases is provided for correction by the user.

Ted has an X windows based graphical interface. The trace file can either be input from the command line or by clicking on the INPUT button after the program has been invoked. Other parameters which the user may specify on the command line include: the output file name; a base position or sequence string on which the trace is to be centered; a default trace magnification; a 5′ vector sequence for automated elimination of the sequence

* To whom correspondence should be addressed

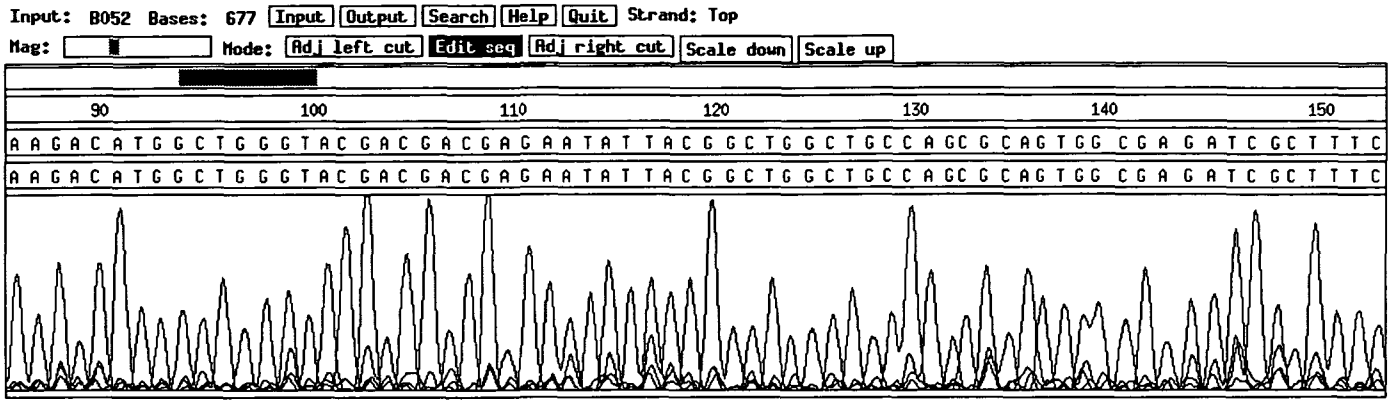

Figure 1. Figure 1 shows a 'screen dump' of the ted graphical interface. The display consists of the control panel and the synchronized view of the base position information, original and edited sequence data, and graphical representation of the trace (with each nucleotide's trace being represented by a different color). The control panel allows the user to read in new trace files (in either bottom or top strand orientation) as well as to search for a string of nucleotides or a certain base position. Scroll bars allow the user to adjust the magnification of or scroll through the sequence and trace data. The user may also choose to change the vertical magnification of the trace data. Further, sequence on the head (vector) or tail (uncertain data) of the sequence may be 'cutoff' using the adjust left and right cutoff buttons. Bases can be inserted, deleted, or replaced as with any ordinary word-processor in the sequence data window. Finally, the sequence may be written to an ascii file using the output button on the control panel.

head (vector); top or bottom strand orientation; or any of the usual X-window parameters (e.g. display, geometry...).

The graphics display (Figure 1) consists of the control panel, the base position information, the original and edited sequence data, and the graphical representation of the trace. The user may begin by using the control panel INPUT button to input a new trace file at which time the user selects whether to view the sequence and trace in top or bottom strand orientation. The trace file is displayed and, if a 5′ vector sequence has been specified on the command line, the program attempts to select a cutoff point corresponding to the vector sequence at the 'head' of the trace file. The bases beyond the 'cutoff' point are displayed on a shaded background. The user may modify the cutoff position by clicking on the 'Adj left cut' button and clicking on the position of the desired cutoff. Similarly, the user may adjust the right cutoff of the sequence (chosen by starting at the 5′ end of the sequence and looking for the first occurrence when 2 out of 5 bases are 'N') by scrolling along the sequence to that point, clicking on the 'Adj right cut' button, and clicking on the appropriate base. Automation of the 'cutoff' process is optional; the user may compile the program with that feature turned 'off'.

Clicking on the 'Edit seq' button allows the user to enter the edit mode. The 'Search' button can be used to skip from 'problem' to 'problem' (i.e., ambiguity to ambiguity) or to look for runs of identical bases (e.g., TTTT) which are often miscalled by the machine software.

Bases can be inserted, deleted, or replaced as with any ordinary word-processor. In difficult-to-read areas, the trace may be vertically or horizontally scaled by dragging or clicking on the magnification scroll bar or by clicking on the vertical scaling buttons ('Scale down', 'Scale up'), respectively. Finally, the edited sequence is saved to an ascii file using the 'Output' button. A history of the editing session can also be saved along with the sequence. The 'Quit' button is used to exit the program. When reinvoking ted on an edited trace file the edited base sequence, rather than the original sequence, is shown in the edited base window. The user may invoke ted by calling in any one of the previous editing sessions.

APPLICATIONS AND CONCLUSIONS

In the *C. elegans* genome sequencing project, data from the ABI or A.L.F. sequencing machines' computers are transferred to Sun workstations. The user invokes a Unix shell script that calls ted systematically on each of the new set of trace files creating a set of sequence files. The sequence files that are deemed to be of acceptable quality are then entered into the sequence assembly program xdap (2) where the sequences are assembled into contigs. Portions of the ted trace-editor have been incorporated into the xdap 'trace manager', which is used in conjunction with the contig editor to view sets of aligned traces at sites of discrepancies in the aligned sequences.

Ted is also used at the stage of choosing oligo primers for the 'walking' stage of the sequencing project. It can be invoked directly from the oligo selection program, osp (3), to allow examination of the trace data in the region of the primers so that integrity of the sequence data can be verified.

Currently, few programs are known to be available which support displaying and editing of both the ABI 373A and Pharmacia A.L.F. trace data. Further, the modular design of the program should allow support for new types of sequencing machines, with new data formats, to be implemented in a straightforward fashion.

AVAILABILITY

Ted is freely available from the authors or from Rodger Staden and Simon Dear (MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK, CB2 2QH) for use on Unix workstations running X-windows. For OpenLook users the executable included in the distribution will run without modification. However, if the user wishes to recompile the programs, for any reason, Release 4 of the X11 windowing system (X11R4) must also be installed.

ACKNOWLEDGEMENTS

The authors would like to thank all members of the *C. elegans* sequencing project with special thanks to the following people:

John Sulston, Bob Waterston, Phil Green, Rick Wilson, Richard Durbin, Simon Dear, and Rodger Staden for their helpful suggestions for improvements in the ted interface and for their parts in the development of ted. The authors also wish to thank ABI and Pharmacia for their co-operation in allowing 'ted's' routines to access the output files of the 373A and A.L.F. instruments. This work was supported by the Medical Research Council and NIH grant R01-HG00136. UNIX is a trademark of AT & T Bell Laboratories. X Window System is a trademark of MIT.

REFERENCES

1. Sulston,J. *et al.* (1991), submitted
2. Dear,S. and Staden,R. (1991) *Nucl. Acids Res.* **19**, 3907–3911.
3. Hillier,L. and Green,P. (1991) *PCR Methods and Applic.* **1**, 124–128.