TECHVIEW
SOFTWARE

DNA Analysis: New Kid on the Block

Darryl Nishimura

Much excitement and debate has surrounded announcements of competing plans to completely sequence the human genome over the next few years. The resulting wealth of data will prove very valuable to those life scientists who are able to mine it effectively. Identification of genes associated with the development of many human disorders will increasingly become a challenge for the field of bioinformatics. Many large commercial enterprises have the resources to pay an in-house staff to develop and maintain customized analysis tools or to purchase such tools from a third party for deployment on an enterprise-wide level. But without reasonably priced sequence analysis tools, smaller research groups and individual scientists could be left behind.

GeneTool 1.0
BioTools Inc.
Edmonton, AB, Canada.
$1000 (academic),
$1250 (industry).
780-423-1133.
www.biotools.com

To address this segment of the market, the Canadian biotechnology firm BioTools has recently released a powerful DNA analysis product called GeneTool that seeks to challenge the established commercial packages in the field. GeneTool is described by the company as a user-friendly, powerful software product for DNA sequence analysis. One interesting feature of the software is its compatibility across numerous common laboratory platforms. The program's intuitive graphical user interface operates in a virtually identical fashion, independent of the operating system. This is accomplished with a unique programming language called Smalltalk, which is similar to Java.

Unfortunately, universal platform compatibility comes at a cost: reduced execution speed. For the GeneTool program to be used efficiently, it should be run on a Macintosh or PC of relatively recent vintage with a minimum of 48 MB of RAM.

The author is in the Department of Pediatrics, 440G EMRB, University of Iowa, Iowa City, IA 52242, USA. E-mail: darryl-nishimura@uiowa.edu

GeneTool is organized around the GeneTool Launcher, a palette of buttons with links to the different modules of the GeneTool package. Each module is responsible for specific tasks related to the analysis of DNA sequence. These can be imported into GeneTool from a number of common formats, such as Raw (sequence only), GenBank, FASTA, DDBJ, ABI chromatogram, and SCF chromatogram. The ability to see the actual experimental results greatly improves the accuracy of making base calls (assignment of base identity based on gel position) from automated sequencing machines. Once in the program, sequences can be easily transferred between the different DNA analysis modules. DNA analysis modules in the program are called Chromatogram Editor, Assembly Editor, Multi-Align Editor, and the Sequence Editor.

The Chromatogram Editor allows a user to import, view, and modify ABI and SCF chromatogram files. These files contain data produced by automated sequencing machines, from which the actual base calls are derived. Bases from the 5′ and 3′ ends of DNA sequences can be trimmed to remove low-quality data from the analysis. Editing of base calls is available in the Chromatogram Editor, and a useful command is provided that automatically finds the next ambiguous base in the sequence. Once the sequence has been edited, changes can be saved to a file in a GeneTool format. It should be noted that this format is not the same as the ABI or SCF format, so copies of the original chromatograms should be saved, should future access to them be necessary.

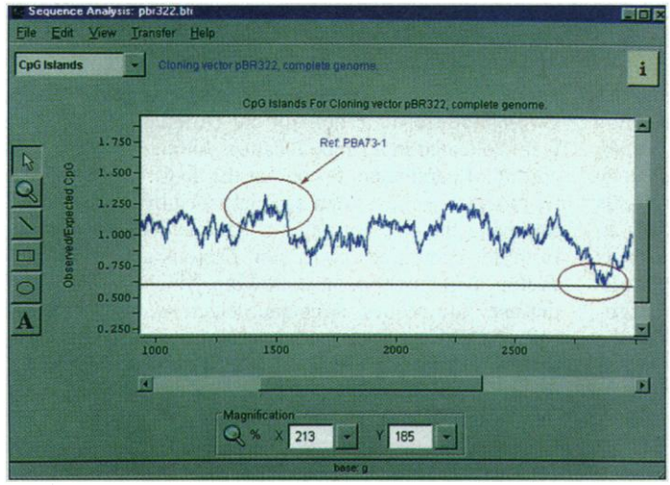

Fig. 1. CpG island analysis of pBR322 sequences. The circled areas identify high (left) and low (right) concentrations of CpG sequences.

The Assembly Editor allows users to assemble overlapping DNA sequences into a longer contiguous sequence. The procedure can be performed on either chromatogram files or DNA sequence files. DNA sequences can be imported into the project individually or as a group of sequences, situated in a common directory. Alignment of the DNA sequences is performed with the XALIGN algorithm (*1*), with user-definable parameters. Bases within an aligned region can be colored on the basis of criteria such as identity, mismatch, property, or feature information. However, the Assembly Editor is the weakest module in the GeneTool package. It would be quite useful to be able to select multiple files for simultaneous viewing to help judge the validity of a base call in a given region. Unfortunately, any editing performed within the Assembly Editor is not transmitted to the chromatogram file, but is only retained in the current Assembly Editor project. Finally, the assembly of sequences is quite slow and is not likely to be very feasible for the assembly of a large number of DNA fragments.

The Multi-Align Editor is very similar to the Assembly Editor. Both modules appear to share many features, and the project window is almost identical. Differences between the two modules are due to the fact that the alignments are being performed with different goals. The purpose of the Assembly Editor is to assemble regions of contiguous sequence from two or more overlapping DNA sequences, while the purpose of the Multi-Align Editor is to look for regions of similarity between two or more different sequences. These sequences may be members of a gene family within the same species or from different species. The user is able to align an entire DNA sequence or, alternatively, selected regions, from each sequence. Then, a pairwise identity matrix can be generated for an aligned region. The same problems found in the Assembly Editor in dealing with chromatograms are present in this module.

The workhorse of the GeneTool package is the Sequence Editor. DNA sequences can be imported in any of the supported formats. The Sequence Editor is packed with a variety of features that are too numerous to list in detail. When sequences are imported in GenBank format, annotated features are shown in color for easy visual reference. Annotations accompanying each sequence can be accessed, new features can be added,

sequence annotations can be edited, and colors that represent features within the sequence can be defined. The module has voice read-back to facilitate error checking and sequence comparison.

The Sequence Editor can also create custom restriction maps or display enzymatic digestions of DNA as a gel simulation. Repeat detection and masking is made possible by the use of organism-specific databases. These greatly aid the specificity of screening for homologous sequences by reducing recognition of repetitive elements as regions of homology. Screening and removal of vector sequence are also functions available in the Sequence Editor. The module can help in primer selection and design, via user-specified conditions. It calculates a melting temperature for each primer template duplex and assigns a quality score to rate its effectiveness.

The Sequence Editor has a number of features that can aid in the identification of potential coding sequence within genomic DNA. First, it can plot the frequency of occurrence of GC, AT, and CG doublets across a DNA sequence, allowing a user to identify CpG islands (CG-rich regions upstream of coding regions of eukaryotic DNA), for example (Fig. 1). In addition, there are several methods for evaluating coding regions within a genomic DNA segment. The module can identify open reading frames within a DNA sequence by traditional start or stop codon methods (Fig. 2). Alternatively, GeneTool offers an exon prediction function, which BioTools claims is a first for a commercially available product. The exon prediction algorithm is claimed to be 93% accurate at the nucleotide level when using a test set of 570 vertebrate genes. To facilitate the evaluation of potential coding regions identified by the program, the program will attempt to match sequences against either a local sequence database (with a FASTA/BLAST-like algorithm) or against the publicly available databases at the National Center for Biotechnology Information (NCBI) via remote BLAST analysis.

In addition to open reading frame and protein identification functions in GeneTool, there are additional modules that support DNA analyses. The Database Search module searches either local or remote DNA databases. Included in the local databases are a BioTools Sequence Database, based on GenBank, and a proprietary DNA sequence motif database. Other databases cover restriction enzymes, repeated sequences, and vector sequences. Sophisticated compression algorithms compress these databases to roughly 3 GB in size. The databases can be updated as often as daily, by special request, via electronic download or on CD-ROM. Remote searching of the Entrez database is also available through NCBI. Local searching can be done by keyword, pattern, or sequence; remote searching is limited to keyword.

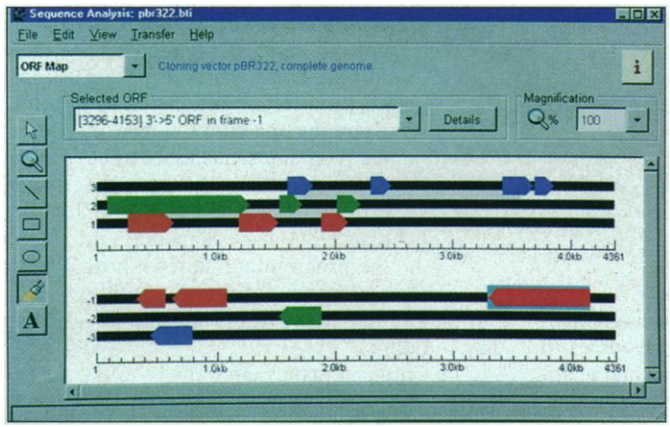

Fig. 2. Identification of open reading frames in pBR322.

There is also a Text Editor with a fully scrollable text-viewing window with autowrapping features. The Preferences module contains user-definable parameters for the different modules. Even though these parameters can often be altered within a given module, it is convenient to have them collected in a single location. In fact, editing of parameters for one of the analysis modules is most often accomplished by retrieving the relevant page of parameters from this centralized collection. There is a Technical Support module through which the user can request assistance electronically. Selection of this module presents the user with a form to enter a description of the problem or make suggestions for program improvements.

GeneTool comes without any printed documentation or manuals. However, the online Help module provides access to the program documentation in a useful and easily navigable format. The help screens are presented in a format resembling a stripped-down Web browser with two frames. The left frame contains navigation information, while the documentation is presented in the right frame. Each screen in GeneTool has a help option within the menu bar. The first line in the help menu is context sensitive, meaning that the page being presented refers specifically to that screen. It is very handy to be able to quickly find the documentation for a particular type of analysis, without having to search through the entire help section.

Another very useful feature of these programs is the option for employing networked parallelism to solve intricate problems. This function divides a complicated problem into segments that can be routed and simultaneously executed on separate, networked computers. Networked parallelism is currently available for Sun Solaris, SGI, or PCs running LINUX. To take advantage of this feature, the user must obtain BioTool's Networked Parallel Server software, in addition to either GeneTool or PepTool. Networked parallelism is available as an additional purchasing option with the programs.

Overall, GeneTool offers many elegant features along with cross-platform compatibility. The program's powerful tools for DNA analysis are quite intuitive to learn and easy to use. GeneTool's use of a specially designed print previewer allows users to edit text and graphics in all figures. Combined with the PepTool 1.1 package (*2*), GeneTool is a powerful tool for DNA and protein sequence analysis at a reasonable price. This initial version of GeneTool does have a few rough spots, however. The two main problems are the slow execution speeds on older systems and inconsistent handling of chromatogram files. Also, users should be aware that running both GeneTool and PepTool simultaneously will likely require at least 64 MB of memory and preferably 128 MB or more.

GeneTool is available for PowerMac (MacOS 7.5 or higher); Windows 95, 98, or NT; Sun Solaris; and SGI operating systems. A minimum of 32 MB of memory is required to run the program, but 64 MB is recommended. This review was conducted on the Macintosh version of the GeneTool 1.0 package. A networked version of GeneTool is available starting at $1400 for academic and nonprofit institutions ($1750 for industry). GeneTool and the protein analysis package PepTool 1.1 can be purchased together for $1500 (academic) or $1875 (industry) for single-license versions and $2000 (academic) or $2500 (industry) for one-user networked versions. The number of concurrent users can be increased incrementally by the purchase of extra keys. Time-limited demo versions of GeneTool and PepTool are available from the BioTools Web site.

References

1. D. S. Wishart, R. F. Boyko, B. D. Sykes, *Comput. Appl. Biosci.* **10**, 687 (1994).
2. B. Basham, *Science* **283**, 1132 (1999).