

PepTool and GeneTool

*Platform-Independent Tools for Biological Sequence
Analysis*

David S. Wishart, Paul Stothard* and Gary H. Van Domselaar

Faculty of Pharmacy & Pharmaceutical Sciences, Protein Engineering Network of
Centres of Excellence and *Department of Biological Sciences, University of Alberta,
Edmonton, AB, Canada T6G 2N8

1. Introduction

PepTool and GeneTool are two new bioinformatics software packages currently being offered by BioTools Inc. (www.biotoools.com). As the names might imply, PepTool is designed for protein sequence analysis and GeneTool is designed for DNA sequence analysis. The combined package is typically priced at \$750 for academic users and \$1500 for commercial users. PepTool is actually based on two public domain programs originally developed at the University of Alberta - SEQSEE (1) and XALIGN (2). These two UNIX-specific programs were later adapted to other platforms, given a graphical user interface (3) and subsequently licensed to BioTools as a commercial package called PepTool. GeneTool was independently developed by BioTools, although it uses some key concepts and algorithms originally found in PepTool. PepTool (version 1.0) was released in December 1997 and GeneTool (version 1.0) will be released in the fall of 1998.

Both PepTool and GeneTool are comprehensive, integrated programs that offer the full range of analytical and graphical features typically found in many advanced commercial bioinformatics products. PepTool and GeneTool also bring some much-needed advances into the bioinformatics arena in algorithm design, graphical interface implementation, data compression, networked parallelism and internet communication. However, probably the most eye-catching innovation lies in the fact that PepTool and GeneTool are both platform-independent software packages. This means that these two programs can run on just about any computer or just about any operating system (Mac OS, Windows, UNIX) without any noticeable change to the programs' overall look and feel. BioTools was able to achieve this by developing the PepTool/GeneTool graphical-user interface (GUI) using a special language called Smalltalk. Smalltalk, which was developed by Xerox's Palo Alto Research Center in the early 1970's, is essentially a more sophisticated version of the better known platform-independent GUI programming

language called JAVA. Smalltalk allows sophisticated GUI's to be prepared without the usual concerns of platform compatibility and back-end design or the constraints of having to work with a slow program interpreter.

In the following pages we will attempt to highlight some of the more useful features offered by PepTool and GeneTool. Particular attention will be paid to the unique or unusual components of each program. Space limitations prevent us from giving a complete overview of both packages. However, it is hoped that this short introduction may offer readers some insight into the design, intent and general utility of these powerful new tools for biological sequence analysis.

2. Methods and System Requirements

Essentially all of PepTool's and GeneTool's complex analytical functions (i.e. the back-end) were written in ANSI C. The graphical-user interface and certain simple analytical functions were written entirely in VisualWorks Smalltalk (version 2.5.2, ObjectShare). PepTool and GeneTool are available for the Power Macintosh (OS version 7.5 and higher), Windows-compatible PC's (Win95, Win98 and WinNT), Silicon Graphics (Irix version 5.0 and higher) and Sun (Solaris version 2.0 and higher) platforms. Other versions for other operating systems may be specially ordered. Networked parallelism (*vide infra*) is available for SUN, SGI and Windows machines and should be available for the Macintosh in late 1998. The combined package (PepTool + GeneTool) without databases, requires 70 MB of disk space (25 MB for PepTool and 45 MB for GeneTool). Both packages come with their own sequence databases and users may arrange to purchase a variety of update options. The specially compressed PepTool protein database requires 60 MB while the compressed GenBank database requires approximately 1.5 GB. Computers running PepTool and GeneTool (plus databases) should have a minimum of 24 MB of RAM (64 MB is recommended) and 2 GB of free disk space. Both PepTool

and GeneTool support WYSIWYG ("what-you-see-is-what-you-get") printing on PostScript-compatible printers, and on any printer running under Windows.

3. PepTool - Specific Program Features

Because of its requirement for universal platform compatibility, PepTool's GUI does not strictly adhere to any single OS interface convention although, as a general rule, it tends to follow many MacOS stylistic tendencies. Depending on the platform being used, the program may be started from either the Finder or Multifinder (for MacOS), by clicking on the Windows Start button (for Win95/98/NT), or by typing *peptool* (for UNIX). After starting, an application "Launcher" (Fig. 1) appears at the top of the screen along with a Sequence Editor window at the center of the screen. The PepTool Launcher allows the user to launch additional windows, to access Help files, to change program preferences or to contact BioTools electronically. In fact, PepTool has at least a dozen different views or windows accessible through either the PepTool Launcher or the Sequence Editor including (i) a Sequence Editor; (ii) an Alignment Editor; (iii) a simple Text Editor; (iv) a Graph Viewer/Editor; (v) a DotPlot Viewer/Editor; (vi) a Helical Wheel Viewer/Editor; (vii) a Structure Viewer/Editor; (viii) a Sequence Motif Viewer/Editor; (ix) a Sequence Statistics Viewer; (x) a Help Viewer; (xi) a Preference Editor; and (xii) a Bug Reporter. Text files, folders or image files created with these different windows can be saved and are automatically marked (with an icon and a three-letter extension) in a format specific to that window. All of PepTool's files, folders and directories can be searched or navigated with a File Chooser which typically resembles the user's system-specific file selector.

3.1 The Sequence Editor

The function of the Sequence Editor (Fig. 2) is to serve as a central workspace from which to enter, edit, retrieve, graph or analyze protein sequences. As such, most of

PepTool's functionality is accessible through this particular window. The Sequence Editor contains a standard set of menu items including: **File** (for file handling and printing functions), **Edit** (for editing the viewed sequence), **Transfer** (for transferring the sequence or selected portions thereof to other applications or windows), **Search** (for finding or retrieving sequences in the database), **Analyze** (for performing statistical or structural predictions), **Graph** (for plotting physico-chemical properties or sequence similarities) and **Help** (for accessing the context-dependent hyperlinked Help system).

Sequences automatically loaded or manually entered into the Sequence Editor can be saved in either SWISS-PROT, PIR, PepTool or ASCII format. The Editor also has the capacity to read "Foreign Format" files including GCG, IntelliGenetics, FASTA, SWISS-PROT and NBRF-PIR as well as other common file types. The Foreign Format reader is both "intelligent" and general, meaning it does not require the user to know or to pre-designate a given sequence file format. Similarly, if the Foreign Format reader encounters a file format it has not seen before, it is usually capable of making a reasonable choice about how to parse the sequence from superfluous text.

As with most sequence editors, the PepTool Sequence Editor supports autospacing, autowrapping and mouse-driven text selection for the usual cutting, pasting, copying and segment deletion operations. It also has a text entry filter (the screen flashes when non-IUPAC letters are entered from the keyboard), a sequence ruler, a "real-time" sequence length monitor and an editable cursor position box which is instantly updated when the cursor position is changed by a mouse-click or text-entry operation. Information about the sequence and the sequence file is displayed at the top of the window and additional data (such as the accession number, journal reference, date, etc.) can be read or entered from a pop-up sequence reference card (accessed by the "Reference" button on the lower right corner of the window).

A particularly useful feature of PepTool's Sequence Editor is its support of color-coded secondary structure display and editing. The buttons located on the right side of

the window allow users to paint secondary structure (if it's known) directly on to a sequence or to pre-cluster certain residues together when performing pairwise sequence alignments. These buttons also serve as a color-coded legend when viewing sequences loaded from PepTool's Structure DB -- a database containing several hundred sequences with known secondary structures.

3.1.1 Database Searching

PepTool permits several kinds of sequence database searches from a variety of databases, all of which are launched from the Sequence Editor (under the **Search** menu item). Results from database searches can be viewed, saved or transferred using a Data Browser (Fig. 3). PepTool supports database queries and sequence retrieval on the basis of (i) keywords (such as organism, protein name, accession number, partial name or logical combinations of the above); (ii) sequence patterns (simple sequence fragments or complex sequence patterns); (iii) subsequence similarity (short stretches of similar sequences); and, most importantly, (iv) global sequence homology. PepTool provides the option of conducting two kinds of global homology searches - a fast one and an exhaustive one.

The fast search [FASTALIGN] (1), which typically takes less than five minutes on a personal computer, is based on techniques similar to those described for FASTDB, FASTA and BLAST, although it uses a specially developed scoring matrix and produces a global alignment instead of a partial local alignment (which is normally done by BLAST). Side-by-side comparisons of FASTALIGN to FASTDB have indicated that FASTALIGN is slightly faster and more sensitive than FASTDB (1).

The exhaustive search [NALIGN] (1), which typically takes several hours on a personal computer (without networked parallelism), is based on the Needleman-Wunsch algorithm (4). Independent tests have shown this to be a very powerful algorithm for remote sequence identification with its performance easily exceeding that of BLASTP,

BLITZ, DFLASH or FASTA (5). Interestingly, the same algorithm now used in PepTool played a key role in identifying a new class of poxvirus-encoded virulence proteins (6) and a novel uracil glycosylase (7).

3.2 The Alignment Editor

The Alignment Editor (Fig. 4) is an intuitive tool designed to permit the viewing, editing and automatic generation of both pairwise and multiple sequence alignments. Typically data is transferred into this window from a Data Browser or Sequence Editor. Sequences may be transferred either individually or in groups. From the **Edit** menu a user can easily add or delete specific sequences or change a given sequence or sequence name. Once the sequences have been loaded and/or edited, the alignment can be computed automatically by pressing the "Compute Alignment" button on the lower right corner. For this operation PepTool uses the XALIGN algorithm (2) which is capable of quickly aligning several hundred sequences using both sequence clustering and secondary structure information in the alignment process. A consensus sequence is automatically generated in the window above the alignment view using the threshold indicated in the Consensus Threshold box. Under the **Display** menu a user can select how the alignment should be displayed with options for coloring by structure (two colors), property (twelve colors) or identity (one color). The pairwise comparison matrix can also be calculated and viewed from the **Display** menu item.

Manual alignment and manual editing of an automatically generated alignment can also be performed by selecting or "painting" over a sequence block. Once highlighted, the entire sequence block (containing one or more partial sequences) can be moved right or left using the mouse-activated arrows at the bottom of the window.

3.3 The Structure Viewer

The Structure Viewer (Fig. 5) displays predicted secondary structure using specially shaded and color-coded "helix" and "beta-sheet" icons. Six different predictions, including the classic Chou-Fasman and Garnier-Osguthorpe-Robson (GOR) methods, are generated. A consensus result is generated based on the weighted average of all six predictions. The consensus result is typically 70% correct based on a simple three-state scoring system. The presence and location of membrane-spanning helices (colored in red) is also predicted using the technique of Klein et al. (8). The order of the individual predictions can be rearranged by toggling a check-box at the bottom of the window and dragging the predicted structures to different locations. Under the **Display** menu it is also possible to selectively turn on or turn off certain predictions. At the top of the Structure Viewer the expected percent content of individual secondary structures is calculated (to allow comparisons with CD or FTIR measurements) and the predicted folding class is identified (it is the one with the highest coefficient).

3.4 The Graph Viewer

This Graph Viewer/Editor (Fig. 6) shares many features with other windows including the Helical Wheel Viewer and the DotPlot Viewer. All three support fully scrollable displays, stepwise or regio-selective zooming and auto-scaling. Furthermore, all three permit the addition or deletion of text, lines, arrows, boxes or circles to the displayed graph using a graphical palette located on the left side of the window. The Graph Viewer is specifically designed to display such functions as hydrophobicity, hydrophobic moments and predicted flexibility. These "protein property" graphs may be further edited through the **Graph** menu, where the user may adjust the graph color, linewidth, graph title and axis titles as well as turn on or turn off the grid lines and residue labels. Through the **Annotation** menu the color, linewidth and line style for any graphical annotation (except text) can also be interactively selected and adjusted.

3.5 The DotPlot Viewer

Dot Matrix or Dot Plot sequence comparisons can be displayed, edited, annotated and evaluated using PepTool's DotPlot Viewer (Fig. 7). Pairwise comparisons between two different sequences as well as simple self-sequence comparisons are possible. The number and length of plotted diagonals can be adjusted using the editable "Stringency", "Window Size" and "Diagonal Filter" boxes. Likewise the color of the plot (as well as the axis and graph titles) can be changed through options listed in the **Graph** menu. PepTool's DotPlot program is unique in that it displays the level of sequence similarity using a simplified color shading scheme, with identical matches appearing brightest and weak matches appearing progressively lighter. The DotPlot Viewer permits the usual zooming and annotation operations found in PepTool's other graphical viewers although, unlike the others, it does allow the sequence for selected diagonals to be viewed in the lower sequence window. This is done by first clicking on the "ATGC" button on the annotation palette and then clicking on a specific diagonal line in the DotPlot window. The pairwise sequence alignment corresponding to that diagonal then appears highlighted in the lower window from which it can be easily viewed or inspected.

4. GeneTool - Specific Program Features

GeneTool shares many basic design and layout features with PepTool. However, it also has a number of important enhancements (many of which are expected to make their way into PepTool, version 2.0). In particular, GeneTool supports resizeable windows, resizeable fonts, multi-feature display, multi-feature editing, print-preview annotation and audio play-back. It also handles database searching, preference selection, reference information, window zooming and window management in a more intuitive fashion. Just as with PepTool, GeneTool may be started from either the Finder or Multifinder (for MacOS), by clicking on the Windows Start button (for Win95/98/NT) or by typing *genetool* (for UNIX). After starting, the GeneTool Launcher appears at the top of the

screen along with a Sequence Editor window at the center of the screen. GeneTool has over 20 different views or windows accessible through either the GeneTool Launcher or its Sequence Editor including: (i) the GeneTool Sequence Editor; (ii) a Translation Viewer/Editor; (iii) a Chromatogram Viewer/Editor; (iv) an Alignment Editor; (v) a Contig Editor (vi) a simple Text Editor; (vii) a Layout or Presentation Editor; (viii) a Graph Viewer/Editor; (ix) a DotPlot Viewer/Editor; (x) a Restriction Map Viewer/Editor; (xi) a Feature/Exon/Sequence Motif Viewer/Editor; (xii) a PCR Primer Designer; (xiii) a Gel Simulation Viewer, (xiv) a Sequence Statistics Viewer; (xv) a Help Viewer; (xvi) a Preference Editor; and (xvii) a Bug Reporter. All of GeneTool's files, folders and directories can be searched or navigated with a File Chooser similar to the one in PepTool.

4.1 The Sequence Editor

Just like the PepTool Editor, the GeneTool Sequence Editor (Fig. 8) serves as GeneTool's central operation window or central sequence worksheet. Consequently, most sequence-specific operations can be launched from this window. The GeneTool Editor maintains a similar arrangement of menu options (**File, Edit, Format, Analyze, View, Transfer**) and it permits the same wide choice of sequence formats to be read or saved (including EMBL, GenBank and DDBJ formats) as the PepTool Editor. To limit the proliferation of file types found in PepTool, the designers of GeneTool have consolidated many of the multiple file types typically generated from a given sequence analysis into a single sequence file. The previously calculated graphs, plots, simulations or other analysis functions associated with a given sequence file can be selected and viewed using the **View** menu. As with most other DNA sequence editors, the GeneTool Editor permits variable character grouping (1, 3, 5, 10 etc.), single or double strand display, DNA to RNA conversion, strand reversion, strand complementation, upper and lower case display, audio playback, autospacing, autowrapping and mouse-driven text selection for

cutting, pasting, copying and segment deletion operations. It also supports the degenerate DNA alphabet (and flashes when non-IUPAC letters are entered from the keyboard), as well as continuously updated sequence length, reading-frame and cursor position boxes.

There are a number of layout or design differences in the GeneTool Editor relative to the PepTool Editor. In particular, the sequence name, sequence length, cursor position and reading-frame boxes now appear in a Sequence Status bar. Furthermore, the reference information button has been moved to the top and replaced with an "I" icon, which also permits more comprehensive annotation and reference display. Perhaps the most noticeable change is the fact that the GeneTool Editor supports a sophisticated feature display and mark-up system using an editable, scrollable "Feature Legend" box. With this system, GenBank, EMBL or DDBJ sequences can be loaded and their feature tables automatically displayed using color-coded text selectors. The full name of the feature (as well as its corresponding color) can be viewed in this expandable Feature Legend box. Individual feature coloring in the text window can be toggled off and on using the colored radio button attached to each "Feature Name" button. By holding down the "shift" key and clicking on the "Feature Name" button (or alternately by clicking on the Feature button at the top of the Legend box), a dialog box containing additional information about that feature (and all other features) is displayed. This dialog box allows the user to add, reorder, edit, annotate or prioritize overlapping features. A key advantage with this feature rendering method is that it allows users to add their own features to new sequence data (in a manner similar to the way PepTool permits secondary structure to be added or removed in its editor). This is simply done by: (i) adding a new feature button to the Legend or editing an existing feature button to the desired feature name; (ii) highlighting the featured sequence in the text window; and (iii) clicking on the corresponding feature button to color the highlighted text.

4.2 The Chromatogram Viewer

Raw sequence data generated from automated DNA sequencers can be read, edited and saved in a variety of formats using GeneTool's Chromatogram Viewer (Fig. 9). In particular, data can be read directly from ABI or SCF formatted chromatogram files as well as GeneTool's own chromatogram format. Individual chromatogram traces can be toggled off or on using the colored A,C,G,T buttons located on the left side of the window or, alternately, by checking off the Base Trace selections under the **Format** menu. Individual trace colors can be changed through a color palette presented in the Preferences window. Each trace can be selected and dragged up or down to help clarify base calls. A 5'/3' trimming feature (located under the **Edit** menu) is also available to eliminate unwanted or unreadable data at the extreme ends of a chromatogram. The vertical scale of all four chromatogram traces can be adjusted using a scaling bar on the right side of the screen. Base calls made by the sequencer can be changed or deleted in a manner similar to most standard text editors. However, insertion of a base or bases must be done through a modal change in the **Edit** menu. The Chromatogram Viewer also supports two types of "Find" functions, one designed to locate ambiguous base calls ("Find Next Problem") and the other to locate specific subsequences ("Find...").

4.3 The Exon Finder

GeneTool uses a unique method for identifying exon/intron locations in eukaryotic DNA based on the reference point logistic (RPL) method developed by Dr. Peter Hooper at the University of Alberta. RPL is similar to a sophisticated neural network and can be trained to recognize very complex patterns and signals, such as those found at exon/intron boundaries. Performance evaluations using the test data (containing some 570 vertebrate genes) of Burset and Guigo (9) indicate that RPL can predict the location of exons and introns with a correlation coefficient of better than 0.85 (P. Hooper, personal communication). This is substantially better than most other gene-finding algorithms, including such popular programs as GRAIL and GRAIL 2 (9). Furthermore, the RPL

prediction only takes a few seconds on a standard desk-top machine. BioTools has enhanced this RPL technique by adding a database search method to fine-tune the initial exon/intron predictions. This typically improves the predictive performance by an additional seven or eight percent, although it adds another four minutes to the analysis time.

When "Find Exons/Introns" is selected from the Analyze menu, a dialog box is presented in which the user is asked to select either the "fast" search (which is the pure RPL method) or the "exhaustive" search (which combines RPL with a fast database scan). Graphical results are presented in a window like the one shown in Fig. 10. Individual exons or the complete set of exons may be selected by a mouse click and transferred to a Sequence Editor. Alternately, the displayed set of exons may be spliced together using the "splice" operation under the **Edit** menu. It is also worth noting that this window (and other graphical windows for motif or feature viewing) permits zooming all the way down to the sequence level so that the full gene sequence can be viewed and inspected.

4.4 The PCR Primer Designer

The Primer Designer (Fig. 11) is both an interactive and an automated tool for PCR primer selection and design. It may be launched either from within the Sequence Editor or from the GeneTool Launcher. To simplify primer analysis, sequence data is always presented in a double-stranded format, with an option to display the amino acid translation between the two strands. PCR primers may be created manually by clicking and dragging on the upper strand (for the "forward" primer) or the lower strand (for the "reverse" primer). During this operation, a primer sequence is automatically generated above (or below) the selected region while the primer length, product length, melting temperature and primer score are calculated and updated in real time in the parameter boxes below. The primer score is an indication of the potential of the primer to form a good PCR oligo. High scores indicate a good primer while low scores with asterisks

indicate the presence of potential false priming sites, hairpin turns or incompatible melting temperatures. Primers generated through this interactive mode can be subsequently edited (to introduce point mutations) in the same manner one would edit characters in a standard text editor. Changes to a primer sequence automatically cause a corresponding change in the translated amino acid sequence (including a change in color) and an update to the primer's calculated melting temperature and PCR score. Note that the original DNA sequence and the translated sequence are not editable -- only the PCR primer sequence is editable.

Automated PCR primer selection is also available under the **Analyze** menu. When the "Find Primers..." operation is selected, a series of compatible primers for both the upper and lower strands is calculated and presented in two data browser boxes that appear at the bottom half of the window. Forward (or upper) primers are shown on the left side and reverse (or lower) primers are shown on the right. These lists may be scrolled through and individual primers selected by a simple mouse click. Selecting a primer in this way brings the primer into the sequence view and updates the parameter boxes located in the center of the Primer Designer window. These primers may then be edited, lengthened or shortened using the same primer editing techniques described earlier. Note that PCR Primer parameters for both the manual and automated mode may be set in the Primer Parameters dialog box.

GeneTool's PCR Primer Designer also supports functions to find sequences or subsequences in both the upper and lower strands; to sort identified primers by their length, position, melting temperature or score; to check primers for specific problems; to rename primers and to save selected primers to a text file.

4.5 Restriction Map Viewer

Essentially every gene sequence analysis package has some kind of graphical restriction map viewer, and GeneTool is certainly no exception. Restriction digests are normally

performed from the Sequence Editor (under the **Analyze** menu) although they may be initiated from the Layout Editor and the Gel Simulation Viewer as well. Both linear or circular DNA can be processed and presented. GeneTool comes with a database of some 400 restriction enzymes although it is possible for users to create their own sub-libraries of enzymes, as well as add new enzymes. When the Restriction Map function is selected, a dialog box is presented which allows the user to select an enzyme library (the default is the full enzyme library) and to choose which enzymes in that library will be used. Specific enzyme selection may be done either on the basis of the overhang produced by the enzyme (5', 3', blunt end), the enzyme cut frequency within the DNA sequence being processed (single cutters, double cutters, etc.) or the enzyme name. Selecting enzymes by name is done using a scrollable check-box list located on the right side of the dialog box. This particular list allows any number of enzymes to be selected or deselected by name. It also indicates which enzymes have been chosen when the user has performed a selection-by-enzyme-type operation.

Once a restriction digest has been performed, a graphical map is generated as shown in Fig. 12. If sequence features have been previously identified, they are displayed as colored bars or semi-circles. Clicking on any colored feature leads to that feature's information being displayed in a Status bar at the bottom of the window. Once activated, that same feature may also be transferred to a Sequence Editor for further analysis. In addition to the sequence feature display, enzyme cut-sites are also displayed. Unique restriction sites are displayed in blue while multiple restriction sites are displayed in black. Clicking on any restriction enzyme label leads to a pop-up box displaying a zoomed-in region of the sequence with the enzyme recognition sequence highlighted in red. Enzyme labels (with the attached site line) may be moved or dragged to any position on the screen to make for a more readable or symmetric presentation. Clicking on two enzyme names, while holding down the "shift" key, allows one to select the DNA

sequence between the two cut sites. This "graphical digest fragment" may then be cut, copied or pasted into another sequence or into another Sequence Editor.

Additional annotation (lines, circles, arcs, arrows, text, etc.) can be added to the map using the annotation icons on the left side of the window. Additional formatting or presentation changes can be performed through the **Format** menu where it is possible to selectively show or hide the sequence rulers, grid lines or enzyme labels. Under the **Format** menu it is also possible to show a complete tabular summary of the restriction digest which includes the enzyme names, frequency of cuts, position of cuts and the recognition sequences. Under the **Help** menu, a user may view the complete GeneTool enzyme library with a full alphabetical listing of the enzyme names, recognition sequences and commercial suppliers.

4.6 The Layout Editor

The Layout Editor offers users the opportunity to create textually complex layouts or text figures (Fig. 13). These complex textual representations of DNA sequence data are commonly presented in published manuscripts, but typically require many tedious hours on a word processor. In an effort to reduce the difficulty associated with generating these kinds of text figures, BioTools has developed a specific Layout Editor to accelerate and simplify the editing process. As seen in Fig. 13, this editor essentially resembles the GeneTool Sequence Editor (minus the Feature Legend box) although it does have additional controls for adjusting the output. By selecting sections of the DNA sequence to be formatted (using the mouse) and then clicking either the "Group By", "Capitalization", "Strands", "Translations" or "Restriction Digest" buttons, it is possible to alter or annotate the highlighted sequence. The Translation button permits multi-frame (one, three or six) translation using either the single letter IUPAC amino acid code or the three-letter code. Likewise, the Restriction Digest button permits a textually annotated

representation of restriction enzyme cut-site locations using the same dialog box and selection procedure found in the Restriction Map Viewer.

Additional formatting and annotation options are also available through GeneTool's Print Previewer. This particular window conveniently allows the user to view the text as it should appear on the printed page and to add or overlay text, lines, arrows, boxes or other useful annotations to selected regions of the PostScript image.

5 General Program Features - Networked Parallelism

Both PepTool and GeneTool offer a unique "speed-up" feature called networked parallelism. Networked parallelism allows a user to run a single program or a process simultaneously on several networked computers. The advantage to running a program on many computers as opposed to a single computer is that the program execution time can be accelerated by a factor roughly equal to the number of computers being used.

Networked parallelism is actually a far cheaper alternative to purchasing a multi-million dollar supercomputer. Indeed, given that many laboratories, universities and private companies already maintain networks of many personal computers, the use of networked parallelism means that it is relatively easy to get the power of 10's or 100's of networked computers for free (without disrupting the operation of other users in the network).

BioTools has implemented networked parallelism (using PVM) in its most rigorous and time-consuming database searching routine (the Needleman-Wunsch algorithm). Our understanding is that in the near future BioTools plans to extend this very useful option to other time-consuming operations such as multiple alignment, contig assembly, secondary structure prediction and exon identification.

5.1 Database Compression

Protein and gene sequence databases are growing faster than hard drive capacity. The April 1998 release of GenBank (the last for which CD's were issued) required 12 CDs to

hold all of the sequence data. Fortunately internet access to the BLAST servers at the NCBI or EBI now allows many researchers quick access to these huge databases without having to find a place to store 7.9 Gigabytes of data or to read a dozen CD's at a time. However, these public servers are somewhat restricted in the types of searches that can be performed and the way that data can be saved, presented or downloaded. Furthermore, a growing number of university researchers and private companies are becoming increasingly concerned about internet security and firewall breaches that may occur when querying publicly accessible databases. The question is: how do you permit flexible database access and maintain security without the headache of purchasing a new hard drive every six months or a new CD every week?

One answer is to use data compression technology. BioTools has made use of the fact that most biological sequence data uses only a restricted "alphabet" of either four (for DNA) or 20 letters (for proteins). This means the size of the ASCII character set can be reduced from 8 bits per character to roughly 2.3 bits for DNA sequence data and 5 bits for protein sequence data. Further, by removing blanks, empty spaces or redundant information from the database text fields and replacing common words with special characters, a good deal more compression can be achieved without significant loss of information. Finally, by combining multiple databases with duplicate entries (as there are for protein sequences) into a single non-redundant database it is possible to gain even more space savings. Using these and other data compression techniques, BioTools claims it has reduced the size of the protein sequence databases from 300 MB to 60 MB and the GenBank database from 7.9 GB to 1.5 GB. This means that the complete set of databases can be delivered on 3 CD's (instead of 12) and easily stored on a regular 2 GB hard drive.

While maintaining a local sequence database offers considerably more convenience, flexibility and security than a remotely accessible database, it is likely that researchers will continue to demand regular access to the NCBI's or EBI's super-fast facilities and highly integrated database features. To maintain this important database

access route, BioTools also offers integrated WWW access to the NCBI server through its GeneTool package.

Summary

Although we are unable to discuss all of the functionality available in PepTool and GeneTool, it should be evident from this brief review that both packages offer a great deal in terms of functionality and ease-of-use. Furthermore, a number of useful innovations including platform-independent GUI design, networked parallelism, direct internet connectivity, database compression and a variety of enhanced or improved algorithms should make these two programs particularly useful in the rapidly changing world of biological sequence analysis. More complete descriptions of the programs, algorithms and operation of PepTool and GeneTool are available on the BioTools web site (www.biotoools.com), in the associated program user manuals and in the on-line Help pages.

Acknowledgements

The authors wish to thank Scott Fortin, Ann Leins and Debby Waldman for their helpful comments and critical reading of the manuscript. We also thank the staff at BioTools Inc. for their assistance in preparing a number of the figures. GH Van D is supported by a PMAC-MRC graduate scholarship and PS is supported by an NSERC post-graduate scholarship and an AHFMR studentship stipend.

References

1. Wishart, D.S., Boyko, R.F., Willard, L., Richards, F.M. and Sykes, B.D. (1994) SEQSEE: a comprehensive program suite for protein sequence analysis. *Comput. Applic. Biosci.*, **10**, 121-132.
2. Wishart, D.S. Boyko, R.F. and Sykes, B.D. (1994) Constrained multiple sequence alignment using XALIGN. *Comput. Applic. Biosci.*, **10**, 687-688.
3. Wishart, D.S., Fortin, S., Woloschuk, D.R., Wong, W., Rosborough, T., Van Domselaar, G., Schaeffer, J. and Szafron, D. (1997) A platform-independent graphical user interface for SEQSEE and XALIGN. *Comput. Applic. Biosci.*, **13**, 561-562.
4. Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443-453.
5. Cattell, K., Koop, B., Olafson, R.S., Fellows, M., Bailey, I., Olafson, R.W. and Upton, C. (1996) Approaches to detection of distantly related proteins by database searches. *Biotechniques*, **21**, 1118-1122.
6. Upton, C., Mossman, K. and McFadden, G. (1992) Encoding of a homolog of the IFN- γ receptor by myxoma virus. *Science*, **258**, 1369-1372.
7. Upton, C., Stuart, D.T. and McFadden, G. (1993) Identification of a poxvirus gene encoding a uracil DNA glycosylase. *Proc. Natl. Acad. Sci. USA*, **90**, 4518-4522.
8. Klein, P., Kanehisa, M. and DeLisi, C. (1985) The detection and classification of membrane-spanning proteins. *Biochim. Biophys. Acta*, **815**, 468-476.
9. Burset, M. and Guigo, R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353-367.

Figure Legend

Fig. 1. The PepTool application Launcher as seen on a Win95 platform. A number of different applications or windows may be accessed from this program launcher.

Fig. 2. The PepTool Sequence Editor as seen on a UNIX platform. Note how the secondary structure can be marked and viewed.

Fig. 3. A PepTool Data Browser as seen on a Power MacOS platform. Database hits may be selected in the upper window and the alignments viewed in the lower window.

Fig. 4. The PepTool Alignment Editor as seen on a Win95 platform. Multiple alignments can be viewed and colored according to sequence identity as shown here.

Fig. 5. The PepTool Structure Viewer as seen on a UNIX platform. The coils indicate helices and the arrows indicate beta-strands.

Fig. 6. A PepTool Graph Viewer/Editor as seen on a Power MacOS platform. An annotated plot of the helical hydrophobic moment is illustrated.

Fig. 7. The PepTool DotPlot Viewer as seen on a Win95 platform. The strong diagonal lines indicate the presence of multiple internal repeats in this protein.

Fig. 8. The GeneTool Sequence Editor as seen on a Power Mac platform. Note the way in which sequence features can be marked and viewed.

Fig. 9. GeneTool's Chromatogram Viewer as seen on a Win95 platform.

Fig. 10. An exon/intron map as generated by GeneTool's Exon Finder (Win95).

Fig. 11. The PCR Primer Designer as seen on a Win95 platform. Note the selectable list of primers on the lower half of the window.

Fig. 12. A Restriction Map of the pBR322 plasmid prepared using GeneTool's Restriction Mapper (as seen on a Win95 platform).

Fig. 13. An example of the textual figures that can be generated in GeneTool's Layout Editor (as seen on a Win95 platform).