TECHVIEW

SOFTWARE

Keeping Track of Bases

Ellen M. Quardokus

Nucleic acid sequence determination and analysis are essential for most molecular biology research. Genome sequencing projects have generated immense volumes of data, so that the challenge to programmers has been to develop software to perform extensive sequence analyses and to organize the information in a coherent form. The Gene Inspector (GI) provides an array of tools for integrating common computing functions in molecular biology, such as analysis and editing of sequences, with graphics and presentation. With GI, users can manage sequence analysis projects from beginning to end. The unique electronic notebook format of GI gives researchers the ability to organize diverse yet interrelated types of information. For example, sequence data from and analyses of a particular clone may be linked to scans of agarose gels showing restriction digests of the clone. These, in turn, can be linked to a description of how the clone was made. All of this information may be kept in one place, printed, and presented from the GI package.

The Gene Inspector
Textco

West Lebanon, NH.
$2599 or
$1899 (academic); $299 (upgrade). Satellite licenses available, $198 or $108 (academic) per computer per 18 months.
603-643-1471
www.textco.com

To facilitate data analysis, the search strategies of frequently performed analyses may be stored together in "analysis suites." These may be recalled and used to analyze new data, ensuring that researchers working together on a project use exactly the same parameters during sequence analyses. Data and the results of the analyses can be linked together by what is known as a "hotlink" so that when data is updated the results will reflect the changes made.

There have been substantial improvements in GI since its debut 2 years ago. In the newest version (1.5), Internet database analyses of nucleic acid and peptide sequences have been incorporated. They include BLAST searches to identify homologous DNA or protein sequences from NCBI databases at http://www.ncbi.nlm.nih.gov/BLAST, BLOCK searches (protein motif identification) from http://www.blocks.fhcrc.org/, and FASTA searches (DNA and protein sequence relationships) from http://alpha10.bioch.virginia.edu/fasta/. For nucleic acids, GI performs GRAIL searches (neural network analysis for identification of consensus sequences) at http://bioweb.pasteur.fr/seqanal/interfaces/grailclnt.html. Also new are SignalP searches (identification of amino acid sequences involved in protein export) at the Web site http://www.cbs.dtu.dk/services/SignalP/. Now, users can also send and receive e-mail from within the program. Much effort has been spent to make possible the manipulation of multiple sequences for alignment, analysis, and presentation. One of the biggest headaches in presenting multiple sequence alignments is shading the appropriate areas of the sequence to emphasize similarities, identities, differences, or special features in the sequence. With noticeable ease, GI provides hundreds of options to meet these presentation needs. In addition, GI provides drag-and-drop options that enable users to move objects (such as text or images) between GI windows as well as into other programs (Adobe Photoshop or Illustrator, www.adobe.com, and Microsoft PowerPoint or Word, www.microsoft.com).

Updates to GI are available at the Textco Web site at www.textco.com/updates.html. Documentation is logically organized and easy to understand, and the manual is spiral-bound to lie flat while being used. Twenty-one short tutorials orient new users to the features of GI and greatly enhance productivity. Detailed step-by-step instructions lead users through the various nucleic acid and protein sequence analyses and provide references to the particular methodology or algorithms used. A free, downloadable copy of GI with an abbreviated manual is available from the company Web site (www.textco.com). The demo is fully functional except that it does not allow users to save, print, or export analysis results; however, users may import their own sequences into the program for a trial run.

GI's electronic notebook provides the framework for storing the diverse types of information generated during a typical laboratory project. The many features found in the notebook make it a powerful tool for organizing day-to-day progress. As sequence data is updated, analyses must also be updated. The ability to hotlink data to analyses hastens the process of reanalyzing data. Another convenient feature of the notebook is the ability to mark an important or frequently used location with the Bookmark option. After Bookmarks have been defined, they may be chosen from the toolbar menu. Large amounts of information, such as recipes for buffers and other solutions or Internet search results, may be stored in Appendices. These can be opened in a separate window from the notebook. Another kind of marker called an Alias points to locations in the Appendices and is placed in the notebook instead of the Appendix itself. The program provides a special type of text, called conditional text, that may be either hidden or revealed upon printing. This feature allows users to present graphics for posters without showing supplementary information that they may want to keep on record in the notebook. Displays may be customized in many ways by changing the font, color, and style of text, axis ranges, tick marks, divisions, labels, and object titles. Such customized formats may be saved as style sheets (similar to the style sheets in a word processor) that allow users to customize the appearance of a particular analysis output once and then easily reapply this format to subsequent analysis.

There are two ways to select objects in a GI notebook: a single click of the mouse button (called a selection) or a double click (called a target). Users should be aware of this distinction because the different types of selection determine which menus exist and which functions are available. Likewise, it is important to recognize that there are two types of files that GI uses: GI sequence files and GI notebook files. Only the notebook files can hold different types of data.

Sequences can be entered into GI in two general ways. If manual entry is required, it is possible to reassign keys on the keyboard to facilitate entering nucleotides. Input sequences can be confirmed in three ways. The program may be set to read the sequence back as it is entered with the Speak Typing function. Alternatively, the entire sequence may be entered before it is read back, or the sequence may be re-entered and compared against the first entry. GI can also import and export sequences in 10 formats: DNAStrider, EMBL, Fitch, GCG, GenBank/GB, IG/Stanford, NBRF, Pearson/FASTA, PIR/CODATA, and Plain Text. Files saved in Textco's Gene Construction Kit, Gene Construction Kit 2, and DNA Inspector IIe formats may also be imported by the program.

GI provides standard editing functions and allows copying and pasting to alter sequences. The GI editor is composed of two main parts: (i) an overview pane at the top of the sequence window showing a graphi-

The author is in the Department of Biology, Indiana University, 1001 East 3rd Street, Bloomington, IN 47405–3700, USA. E-mail: ellenmq@indiana.edu

cal representation of the location of any portion of the sequence in relation to the whole sequence and (ii) an editing pane that contains the sequence itself. The overview pane may be used to quickly advance the sequence in the editing pane by sliding the "segment indicator" box along the line representing the sequence. Multiple sequence alignments may also be edited in a similar way.

The Sequence menu of GI provides a way to enhance sequence alignments with boxing, shading, or inversion of the text color (Fig. 1). The menu enables the researcher to choose the way the sequence is presented by customizing the Score Adornments dialog. Together with the Format menu, this dialog allows one to set the color, pattern, box, or shading used to bring attention to either consensus sequences or nonmatching areas. After sequences are formatted in the sequence file, one can use other data in the GI notebook with the drag-and-drop or copy-and-paste methods. Additional sequence formatting options are available from within the GI notebook through the Features menu. The Features menu lets the user translate nucleic acid sequences, show restriction endonuclease recognition sites, and display protein digestion results.

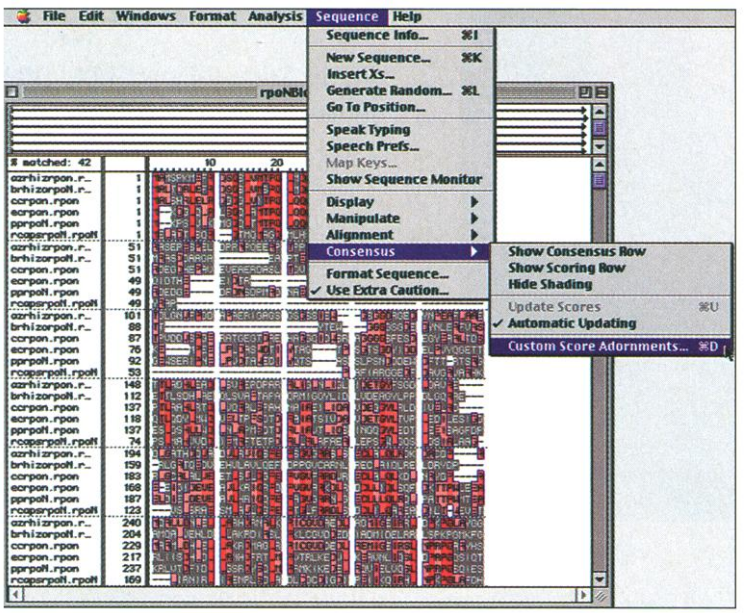

Fig. 1. Multiple sequence alignment window in Gene Inspector.

The general approach for performing analyses on nucleic acid or amino acid sequences in GI consists of choosing the input sequence, the type of analysis to be performed, and the notebook in which to store output if it is different from the one currently open. Explanations of the parameters for each analysis are presented in an easy-to-understand manner in the GI manual. The most powerful feature of the program is its ability to define suites of analyses that can be saved to perform multiple analyses simultaneously. With it, a standardized set of analysis parameters can be set up once and used again on other input sequences.

GI contains Codon Preference Tables for many model organisms. Users can update these tables as new information becomes available and create user-defined Codon Preference Tables as desired. GI's functions for analysis of nucleic acid and peptide sequences include the ability to align two or more sequences, to perform dot matrix analysis. With the Find Sequences feature, users can locate specific sequence segments within a larger sequence. The Clustal V algorithm is used to align multiple sequences (Fig. 1).

Nucleic acid–specific analyses in GI include algorithms for finding specific sequence motifs, such as inverted repeats, direct repeats, and restriction enzyme sites. Other options allow users to determine information such as base composition or combination of bases as a function of position along a DNA sequence. Standard nucleic acid analyses in GI include identification of coding regions (open reading frames, ORFs) in a DNA sequence, Fickett's TestCode algorithm (an alternate way to identify ORFs), codon preference determinations, and a GC coding prediction algorithm for organisms with genomes that are rich in GC sequences.

Protein sequences can also be entered in two ways. GI can translate DNA sequences into protein sequences. Alternatively, protein sequences may be imported from online databases or sequence files. Available protein sequence analyses include amino acid composition, identification of sequence repeats, pI (isoelectric pH) prediction, physical characteristics, Prosite motif search (which identifies common protein sequence elements), predictions of protease cleavage sites, and side-chain flexibility predictions. Users may perform a number of so-called "sliding window" analyses on peptide sequences. This type of analysis examines the peptide sequence by calculating values for a "window" of adjacent amino acids, plotting the data and moving or "sliding" along the sequence one character at a time, calculating a new value and plotting it until the end of the sequence is reached. The output is given as a plot of the property being examined. Sliding window analyses for proteins include the following protein properties as predicted by the program: antigenicity, hydropathy, hydration, regions embedded in membranes, sites internal to the protein, signal sequence, side-chain protrusion, surrounding hydrophobicity, temperature factor, and transmembrane helix predictions. A number of useful secondary structural analyses in the program include Chou-Fasman (CF) structure prediction; Garnier, Osguthorpe, and Robson (GOR) structure prediction; Helical Wheel analysis of predicted α helices; membrane-buried regions; and transmembrane helices.

GI provides a friendly environment for elegantly handling the mounds of information typically gathered during sequence analysis. In contrast to programs that are designed purely for analysis (e.g., MacVector and LaserGene), GI's notebook design gives users an environment where they can better define their own custom analyses and format the results in one simple step. The program uses familiar Web browser–like functions for creating hotlinks, Bookmarks, and Aliases to navigate within the notebook environment. The Internet database search features are a welcome addition to this version of the program. The tutorials are well chosen to illustrate commonly needed and unique features of the electronic notebook format. The program's ease of use and its functions for defining and saving standard, complex sequence analysis routines for reuse are particularly helpful in any workplace where individuals are inexperienced in performing sequence analysis. A companion plasmid-mapping program called Gene Construction Kit 2 from Textco complements GI nicely. One notable shortcoming is that GI lacks a contig (contiguous or overlapping DNA sequences) assembler for short sequences, requiring users who need this function to purchase another product.

System requirements for this software are as follows: Macintosh System 7.0.1 or later with 8 MB RAM available for the application, 14 MB hard disk space and a Thread Manager (which is built into System 7.5 and later). A PC version of GI is not currently available.