UC Davis
UC Davis Previously Published Works

Title
The Genetic Data Environment. A user modifiable and expandable multiple sequence analysis package.

Permalink
https://escholarship.org/uc/item/36x2x7bt

Journal
Methods in molecular biology (Clifton, N.J.), 70

ISSN
1064-3745

Author
Eisen, JA

Publication Date
1997

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

The Genetic Data Environment: A User Modifiable and Expandable Multiple Sequence Analysis Package

(A GUIDE for the Graphical User Interface (GUI) GDE)

Jonathan A. Eisen

Department of Biological Sciences
Stanford University
Stanford, CA 94305-5020
jeisen@leland.stanford.edu

1. Introduction

The Genetic Data Environment (GDE) is a software package designed by Steve Smith and colleagues for molecular sequence alignment and analysis (Smith et al. 1994). Four features make GDE stand out relative to other similar programs:

1. It is free.
2. It has a user friendly and visually powerful multiple sequence alignment editor.
3. Analysis can readily be performed on any sequence(s) or region(s) of sequences simply by selecting the sequence(s) or region(s) of interest and choosing the desired function from the pop-up menus.
4. Although it comes with a variety of powerful sequence analysis tools, any additional programs of the users interest or updates for programs in use can be incorporated quickly and easily into the menu system.

The current release of GDE includes a variety of sequence analysis tools including methods for sequence alignment and editing, conversion between sequence formats, nucleic acid translation, identification of restriction sites, RNA secondary structure prediction and drawing, database searching, dot plots, phylogenetic analysis, consensus determination, and printing and formatting. Instructions for how to use many of these features are presented here. However, since GDE is user expandable, the main focus of this chapter will be on how to use the core GDE alignment window. In addition, a brief guide on how to add additional programs to the GDE menu system is included. Learning to use this type of program may be of more use in the future - other programs will likely adopt such a user expandable system. For example, currently, work is in progress to incorporate many of the features of GDE into the incredibly powerful but somewhat cumbersome software package GCG (Smith, personal communication).

2. Materials

2.1. Hardware

The GDE software package is designed to run on the Sun family of computer workstations. However, it can also be run with some modifications on other Unix based workstations such as DecStations and SGIs. The sequence alignment editor is designed to be run in an X-Windows or OpenWindows environment and can be displayed locally (on the machine running the GDE software) or remotely, on any machine capable of X-window emulation (e.g. MacX can be used for displaying on a Macintosh). Although most of the features and programs of GDE are designed to be run from the alignment editor, many can also be run from the Unix prompt. A working knowledge of Unix and X-windows is helpful for using GDE but not necessary. Whenever possible I include all instructions needed.

The core GDE package requires about 15 megabytes of disk space. Additional space is required for sequence database files. The amount of RAM needed varies a great deal depending on the size of the sequence files being viewed and the number and type of programs used to manipulate or analyze these sequences at any one time. The GDE system can be run with either color or black and white monitors. However, to best use the sequence alignment window it is

helpful to have color. This is because amino acids are colored by chemical type (all acidic are one color, all basic are another, etc.). Thus regions of sequence similarity can be quickly identified by blocks of particular colors. In addition, some of the highlighting features of particular GDE programs work best when viewed in color.

2.2. Software

The current GDE package (version 2.2) can be obtained from a variety of computer archives. URL addresses for some sites are given below.

1. http://golgi.harvard.edu/ftp/
2. http://www.dl.ac.uk/SEQNET/gde.html
3. gopher://megasun.bch.umontreal.ca/11/GDE
4. gopher://rdpgopher.life.uiuc.edu/11/programs/Editor_GDE
5. ftp://ftp.sunet.se/pub/molbio/unix/GDE
6. ftp://fly.bio.indiana.edu/molbio/unix/GDE
7. ftp://solomon.technet.sg/pub/NUS/Z2/indiana/molbio/unix/GDE

The GDE package is usually found at archive sites in compressed archive format as a single file (e.g., gde2.2.tar.Z). This file must be copied to a local machine, decompressed and unarchived. In addition, the *.cshrc* file of all users who want to run GDE must be modified slightly. Below are the instructions that can be used set up GDE for a Sun Sparcstation (once the file has been copied from an archive site). The commands in *italics* should be typed from the Unix prompt and followed by a carriage return. For other machines some modifications of these instructions may be necessary. The specifics will depend on the machine, the type of Unix being run, and the type of X-windows being used for display. Instructions for setting up GDE on a variety of other machines are available at many of the above archive sites.

% *mkdir /usr/local/GDE* <return>
(makes a directory for the GDE program)
% *mv gde2.2.tar.Z /usr/local/GDE/* <return>
(moves the file to the directory)
% *uncompress gde2.2.tar.Z* <return>
(uncompresses file)
% *tar -xvf gde2.2.tar* <return>
(unarchives file)

For each user, the following lines should be added to the *.cshrc* file found in their home directory. The additions can be made using a text editor like *vi*, *emacs* or *textedit* or by using the *cat* command (type *cat >> .cshrc* from the Unix prompt and any text typed will be added to the *.cshrc* file, when done type control-D).

```
set path = ($ path usr/local/GDE/bin)
setenv GDE_HELP_DIR
/usr/local/GDE/GDEHELP
```

2.3. Databases

The GDE package comes with two database comparison programs - *fasta* (Pearson 1990) and *blast* (Altschul et al. 1990). To make use of these programs, the desired databases must be set up in specific formats and locations. All should be set up in subdirectories within the GDEHELP directory (/usr/local/GDE/GDEHELP). Instructions for doing so are given below. Special programs are required to format databases for the *blast* programs, and these are included with the GDE package. To run these programs, simply type their name followed by a carriage return from the Unix prompt. If the appropriate databases are already set up elsewhere on a local system, aliases for the locations of these files can be set up in the directories described below instead of copying the entire databases.

1. For *fasta* protein searches, copy PIR to the GDEHELP/FASTA/PIR/ directory
2. For *fasta* nucleotide searches, copy Genbank to GDEHELP/FASTA/GENBANK/ directory.
3. For *blast* protein comparisons, copy PIR to GDEHELP/BLAST/PIR/. Use the *pir2fasta* program to convert to temporary FASTA format. Then reformat the database using the *setdb* program.
4. For *blast* nucleotide comparisons, copy Genbank to BLAST/GENBANK/ in the GDEHELP directory. Then use the *gb2fasta* program to convert to temporary FASTA format. Finally, use the *pressdb* to reformat the database.

3. Methods

3.1. GDE Basics

3.1.1. Starting the program

Prior to starting GDE, the user must set up for displaying in an X-window or equivalent environment. If GDE is to be run locally on a workstation, usually the windows environment

will be started when you log on to the machine. If it is not started, try typing *x* or *openwin* from the Unix prompt. GDE can also be run remotely by setting up to display on a local machine but running the program elsewhere. There are many ways to do this depending on the machine on which you will be displaying. In general, what you have to do is instruct the local machine that you are allowing the remote machine that will be used to run the GDE software to be a X-windows host (for many Unix systems type *xhost + remote_machines_address* from the Unix prompt replacing *remote_machine_address* by the name or IP address of the machine from which you will run GDE). Then you have to instruct the remote machine that you will be displaying GDE elsewhere (for many Unix systems type *setenv DISPLAY local_machines_address:0* from the Unix prompt replacing *local_machine_address* by the IP address or name of the machine used as the display).

Once everything is set up, to start GDE type *gde* or *gde filename* (where "*filename*" is replaced by the name of the file one wants to open), followed by a carriage return. The GDE alignment window should appear. An example window is shown in Fig. 1. This window includes many of the features that will be referred to later.

3.1.1. Using the Mouse and Menus in GDE

GDE is a menu driven X-windows based system. Like with other windows environments, in X-windows, pop-up/drag down menus are used to access a variety of commands. The most obvious difference between X-windows and Mac or PC-based windows is that there are three buttons on the mouse to become familiar with. For GDE, the buttons are used for different functions including:

Left Button: placing cursor, selecting sequences and regions of sequences, scrolling, resizing windows, splitting screen
Middle Button: extending text selection
Right Button: opening pop-up menus and scroll-bar menus

The most important mouse skill in GDE is selection of items from the GDE menus. To select an item in a menu, such as the *File* menu (in the upper left in Fig. 1):

1. Point the mouse cursor at the menu button of interest and click with the right mouse button. This will expose the items in the drag down menu.
2. Select one of the items in the menu by pointing and clicking with the left mouse button.
3. Menus can be "thumbtacked" to the screen by first selecting the menu with the right mouse button and then clicking on the thumbtack with the left mouse button.

For most GDE menu items, a dialog box will appear after the command has been selected. These boxes are used for various types of input that define exactly how the command will be executed. GDE uses five types of input formats in these dialog boxes - text lines, sliders, chooser buttons, pop-up menus, and check-boxes (the first four are demonstrated in Fig. 2).

1. Text lines. To enter text in a text line, point the mouse cursor to the text line, click with the left mouse button, and then type the text.
2. Sliders. To modify values in sliders point the cursor to the rectangular box on the slider line and then click and drag it (with the left mouse button) to the left or to the right to get to the desired number (which is shown in the text line to the left). Sliders can be altered in increments of 1 by pointing and clicking with the left mouse button to the right or left of the slider box, along the slider line.
3. Chooser buttons. To alter selections in chooser buttons simply click with the left button one of the boxes to the right of the choice. The selected box will be highlighted.
4. Pop-up menus. Pop-up menus can be altered as described above for GDE menus.
5. Check-boxes. Boxes are checked by simply pointing and clicking with the left mouse button.

In general, once the dialog box has been "filled out" to the users interest, the command is started by clicking the *OK* or *DONE* buttons. As mentioned above, one of the most powerful aspects of GDE is the ability to quickly add new programs. A dialog box like the one in Fig. 2 can usually be added in about 30 minutes with no programming experience except a little knowledge of Unix commands. The dialog boxes are helpful because once they are programmed the user does not have to remember the code line instructions for each program. Thus any program can be run on specific sequences or regions with the click of a button. Instructions for adding menu features are given in section 4.

3.1.2. Sequence Input and Sequence Types

GDE uses four different types of sequences: DNA/RNA, protein, text, and masks. The sequence type is important in determining what characters are allowed to be entered into the sequence, as well as how external programs handle the sequence when it is selected for analysis. DNA/RNA and protein sequences use the standard single characters to denote the specific nucleotides and amino acids. Characters denoting degenerate positions are also allowed but may interfere with some external functions. Text sequences allow any characters and are particularly useful for keeping notes along with an alignment (e.g., intron positions, transcription start sites, mutation spots, etc.). Masks are used to direct external programs to use only subsets of a sequence alignment. This can be particularly useful in phylogenetic analysis (section 3.5.5) but is also helpful in other functions (section 3.1.16.).

There are three ways to get sequences into a GDE window. Short descriptions for each method are given below. Combinations of these can be used to load up multiple files and sequences into one window (remember to check the file name prior to saving if multiple files have been opened or imported).

Direct input (for sequences in GDE, flat, or Genbank format)

1. Choose the *Open* ... command from the *File* menu.
2. In the dialog box, the local directory is shown. Click on the name of the file to be opened or move through the directories to find the file of interest.
3. Once the file is selected, click the *Open* button.
4. The sequence(s) will be added to the ones currently in the GDE Window.

Loading Sequences in Other Formats

1. Choose the *Input Foreign Format* command from the *File* menu.
2. A text line for inputting the name of the file to import will appear in the dialog box. If the file of interest is in the directory from which the GDE program was started, type in the file name (e.g. gde.pir). If the file is in another directory you need to type the path name as well (e.g. /GDE/gde.pir). Sometimes it is easier to move the file to the directory in which GDE was started rather than typing the entire path name.
3. Click the *OK* button.
4. The sequences will be converted to GDE readable format and added to the ones already in the GDE window.
5. This function uses the *readseq* program to convert between sequence formats and thus has all of the features and bugs of this program. It is important to be careful when importing sequences that have been received by email from sequence databases. Depending on the way they were received and the email system used, sometimes the email headers can interfere with the importing functions. In addition, only some sequence information fields will be converted, others may be left out, or merged into the same field. Instructions for accessing sequence information fields are in 3.1.5.
6. Formats readable include Genbank, IG/Stanford, NBRF, EMBL, GCG, DNA Strider, Fitch, Pearson/Fasta, Zuker, Olsen, Phylip, Plain text, ASN.1, PIR, MSF, and PAUP.

New sequences

1. Choose the *New Sequence* command from the *File* menu
2. Choose the sequence type (DNA/RNA, protein, text, mask) from the dialog box.
3. Type in a name.
4. Click the *OK* button.
5. A sequence name (with no sequence yet) will be added to the sequences already in the GDE window. The sequence can then be typed in directly (see 3.2.1)

3.1.3. Selection of Sequences or Regions for Analysis

In general, functions selected from GDE menus are performed only on the sequence(s) or region(s) that have been selected by the user. The ability to quickly select different sequences and regions of interest allows the user to perform desired analyses quickly and specifically. For example, to compare a small segment of the N-termini of a protein to a sequence database, just select that region and then choose one of the database searching options from the GDE Menu (see 3.3.8).

Sequences and regions can be selected either directly using the mouse or indirectly using menu functions. The currently selected sequences or regions are highlighted in the GDE window (see Fig. 1). It is important to note that region and sequence selection are independent - changing selected regions has no affect on which

sequences are selected and vice-versa. However, for some commands, sequence and region selection can be in conflict. This occurs when the command chosen can be performed on either sequences or regions (e.g., multiple sequence alignments). In these cases, a selection window will appear asking the user to choose whether the function is to be performed on the region(s) or sequence(s) selected. Some functions can only be performed on either regions or sequences but not both (e.g., grouping, see 3.1.8) and thus a chooser window will not appear in these cases.

Sequence Selection

1. Click on the short name of the sequence with the left mouse button.
2. To select multiple sequences use mouse dragging (click and hold the left mouse button while dragging the mouse cursor across the names of the sequences to be selected and releasing after the last name) or shift clicking (hold the shift key while performing additional selections with the mouse button).
3. Use the *Select All* or *Select by name* commands from the *Edit* menu to select sequences indirectly.
4. Deselection of sequences is done either by selecting other sequences without holding down the shift key or clicking the mouse in the region immediately to the right of the sequence names (but to the left of the sequence text).

Region Selection

1. With the left mouse button held, drag the mouse cursor across the region to be selected and release the button when at the end of the region.
2. Alternatively, "embrace" the region to be selected by pointing and clicking at one side of the region with the left mouse button and then point and click on the other side of the region with the middle mouse button. This method allows the use of scroll bars to move to the second edge of the region to be selected (which makes selection of long regions of sequence easier than with mouse dragging).
3. Both of the above methods can be used to select a region from one sequence or comparable regions from multiple sequences.
4. Selection of regions can be complicated by grouping of sequences (see 3.1.8) - selecting a region in one member of a sequence group will automatically select that region in all members of the group.
5. To deselect regions, select another region or point and click the mouse anywhere in the text.

3.1.4. Saving Sequences and Alignments

GDE allows for sequences and alignments to be saved in a variety of formats. Be careful when switching between formats - some formats do not retain all of the reference information retained by other formats (such as authors, accession numbers, etc.). In addition, some formats do not retain all of the sequence data. For example, some formats require all sequences to be of equal length so longer sequences may be truncated when converting to these formats. Thus, unless there is some need to switch to a different format, it is probably best to save things either in GDE or Genbank format. The GDE format retains essentially all of the information seen in the GDE window, such as alignment, sequence groups, and sequence notes. For more information on the GDE format, see the GDE manual (described in section 5). The three different means of saving sequences and alignments are described below. If you have many sequence files be careful not to overwrite files of interest.

Saving Everything (Entire Alignment)

1. Choose *Save As* from the *File* menu.
2. Select format (GDE, Genbank, or Flat) from the dialog box.
3. Enter a new name or leave the original name.
4. Click the *OK* button.
5. The file will be saved in the directory where the GDE program was opened.

Saving Specific Sequence(s) or Region(s)

1. Select sequence(s) or region(s) to be saved.
2. Choose *Save Selection* from the *File* menu.
3. Select sequence format (GDE, Genbank, Flat).
4. Enter file name.
5. Click the *OK* button.
6. The file will be saved in the directory where the GDE program was opened. Alignment information will be retained.

Saving in other formats

1. Select sequence(s) or region(s) to be saved.
2. Choose *Output Foreign Format* from the *File* menu (this uses the *readseq* program like the *Input Foreign Format* Command described above).

3. Select output format from the pop-up menu (Genbank, IG/Stanford, NBRF, EMBL, GCG, DNA Strider, Fitch, Pearson/Fasta, Zuker, Olsen, Phylip, Plain text, ASN.1, PIR, MSF, PAUP, Pretty).
4. Enter a name for the file.
5. Click the *OK* button.
6. The file will be saved in the directory where the GDE program was opened.

3.1.5. Sequence information

GDE allows storage of a variety of information for each sequence. Under normal conditions the majority of this information is kept hidden. Access to this information is gained via a dialog box (see Fig. 3). This information can be useful for sorting functions (see 3.1.6), for future reference, and for a variety of special functions. For example, strand and direction will influence translation functions and sequence type will influence allowable modifications.

1. Select the sequence of interest.
2. Choose *Get Info* from the *File* menu.
3. Change or enter text for short name (the name shown in GDE Window), full name, ID number, description, author, comments.
4. Set pop-up menus for sequence type, strand, direction.
5. Click the *OK* button when done.

3.1.6. Sorting and Ordering Sequences

In order to aid multiple sequence alignment and analysis it is frequently helpful to have specific sequences next to each other. Reordering of sequences can be done in two ways - either by cutting and pasting or using sorting functions.

Manual

1. Select sequence(s) to be moved.
2. Choose the *Cut* or *Copy* commands from the *Edit* menu, or use built in cut/copy keyboard function keys.
3. Select the site at which the sequences are to be placed (by selecting the sequence immediately above the site).
4. Choose the *Paste* command from the *Edit* menu.
5. The sequence(s) will be added in the same order (top to bottom) as they were before cutting.

Computer based

1. Select sequence(s) to be sorted.
2. Choose the *Sort* command from the *Edit* menu
3. Choose primary and secondary sort fields (group, type, name, sequence ID, creator, offset) in the dialog box.
5. Click the *OK* button.
6. A new GDE window with the results will appear.

3.1.7. Extracting sequences/Regions

Sometimes it is helpful to extract subsets of sequences or regions of sequences into a new alignment window. This can be done in either of the following two ways:

Direct

1. Select sequence(s) or region(s).
2. Choose *Extract* from the *Edit* menu.
3. A new GDE window with the results will appear.

Indirect

1. Select sequence(s) or region(s).
2. Choose *Save Selection* from the *File* menu (see 3.1.9).
3. Use the *Open* command to re-open this saved selection (see 3.1.2).

3.1.8. Grouping Sequences

Grouping of sequences allows editing functions to be performed on all members of the group at the same time. This feature is particularly useful for aligning sequences by hand. For example, if one had separate alignments of 30 gamma globins and 30 beta globins and wanted to align them together manually, it might be easiest to group all of the beta globins into one group and all of the gammas into another. Then, alignment gaps could be placed in all gammas at the same time and all betas at the same time by entering the gap into only one of the members of the group. If one then wanted to put a gap in only one or a few of the beta globins, they could be ungrouped and the gap could be placed in just those few. When editing functions are attempted on one member of a group, only those modifications that are permitted for all members of the group will be allowed (see 3.1.9). Regions cannot be grouped, only sequences can. To change sequence groups:

1. Select sequence(s) to be grouped or ungrouped.
2. Choose *Group* or *Ungroup* from the *Edit* menu.

3. If any of the sequences selected are part of another group, the user will be asked whether to merge the groups or to create a new one.
4. A number will be placed to the left of the short sequence name(s) to indicate group status.

3.1.9. Sequence Protections

GDE allows for the protection of sequences against accidental modification. There are four different types of modifications allowed during editing. The default (which is used when an alignment is first opened) is to allow only modification of alignment gaps and translations. Depending on the type of sequence (DNA, protein, text, mask) "ambiguous" characters are different. For example, N is ambiguous for DNA and RNA but is not for protein. All modifications are allowed for new sequences (when the New Sequence command is used (3.1.2). To modify sequence protections:

1. Select sequence(s).
2. Choose *Protections* from the *File* menu.
3. Select modifications allowed (unambiguous characters, ambiguous characters, alignment gaps, translations).
4. Click the done button when done.

3.1.10. Repeat Counts

Repeat counts allow the user to repeat a keystroke any number of times by typing the number corresponding to the desired number of repeats immediately prior to typing the key to be repeated. This is very useful for manual sequence alignment (for inserting or removing multiple gap characters) and for moving the cursor a defined number of spaces (see 3.2). Repeat counts will not work when the cursor is in a text or mask sequence because numbers can be used as part of these sequences. However, repeat counts can be used indirectly for text or mask sequences if these sequences are grouped with protein or DNA/RNA sequences and the repeat counts are typed when the cursor is in the protein or RNA/DNA sequences.

3.1.11. Printing

GDE has two means of printing sequences or alignments. "Normal" GDE printing allows printing of sequences and alignments with a variety of Unix commands as well as viewing and editing the file to be printed. Sequences can also be printed with the PrettyPrint format of the *readseq* program. PrettyPrint output is designed for publishing and presentation of alignments and can produce very nice figures. Both printing commands are accessible from the *File* menu.

3.1.12. Cursor position

The cursor is identified by the flashing horizontal line in the sequence text section of the GDE window. It is used in essentially the same way as the cursor in most word processing programs. First and foremost the cursor marks the spot at which editing commands are performed and text selections begin. In addition, it can be used to mark a place for quick returns if the screen is scrolled to another page. Information about the cursor position is displayed in the status line (see Fig. 1). To move the cursor, either point to a new region and click with the left mouse button or use the arrow keys (repeat numbers can be typed before the arrow keys to move a specific number of positions). If the cursor is moved past the edge of the screen, scrolling will be activated and the next page of sequence will be shown. Since scrolling can be performed without moving the cursor (see 3.1.13), the cursor may not always be visible in the GDE window. The cursor may be hidden from view if the scroll bars are used to scroll pages. To return the screen to the region of sequence where the cursor is located, type one of the arrow keys. This function (which I will refer to later as the *return screen* function) is helpful but can lead to some confusion. If you want to keep the view on the sequences you have scrolled to, remember to change the cursor point to that region using the mouse.

3.1.13. Scrolling

Only a portion of most sequences will be viewable in a single GDE Window. The rest of the sequence can be viewed by scrolling to another page (to the right or left). In addition, if an alignment contains many sequences it may be necessary to scroll up and down to see different sequences. Scrolling can be performed in a variety of ways including:

1. Click with the left mouse buttons on the arrows on the scrolling elevator (see Fig. 1).
2. Click and drag in the center of the elevator.
3. Use the scrollbar menu (which is opened by clicking with the right button on the scrollbar)
4. Click on the scroll bar edges (the vertical lines at the edge of the scroll bar). This moves the window all the way to the beginning or end of an alignment.

5. Use the cursor arrows to move the cursor past one edge of a screen page (see 3.1.12).

3.1.14. Split Screens

A split screen allows the viewing of discontinuous regions of a particular alignment. This can be used for example, to insert gaps in the upstream portion of a sequence while simultaneously monitoring the alignment of the downstream portion, even 1000s of positions away. Be careful not to have different vertical positions for different screens - this will make comparisons of the different screens difficult Vertical scrolling can be locked in the screen properties menu (see below) to minimize such problems.

The region of the alignment shown in a particular screen can be changed in three ways, by upstream manipulations of the sequence (such as insertion of gaps) in another screen, by using the scrolling functions, or by using the return screen function described in 3.1.12. The *return screen* function can lead to much confusion when using split screens because this function only operates on the active screen. The active screen is determined by the screen in which the mouse pointer is pointing. Therefore, be sure to know which screen the mouse is pointing to before you use the *return screen* function. For example, imagine you are using the right screen to view the C-termini of a protein alignment and the left screen to view the N-termini, and the cursor is in one of the proteins in the N-termini. If you want to insert a few alignment gaps in this proteins' N-termini be careful that the mouse is pointing to the left screen. If it is pointing to the right screen, when you type the alignment gaps, the right screen will return to the position of the cursor and thus you will have two screen showing the N-termini. Below are descriptions of the two ways to make and remove split screens. Any number of split screens can be used at one time.

1. Point mouse cursor at edge of scroll bar.
2. Click and drag to create or remove split screens.

1. Point and click right mouse button on the scroll bar.
2. Select from pop-up menu *split views* or *unsplit views.*

3.1.15. Screen Properties

Many of the screen features can be altered using X-windows functions and thus are specific to the type and version of X-windows being used. In addition, GDE allows the user to modify a variety of the properties specific to the GDE window (see below). Becoming familiar with these functions is important because they can be used to aid in analysis and alignment of sequences. In addition, some of the programs run through GDE menus may change the screen properties. For example, the *Variable Positions* command (3.5.8) changes the sequences to black and white to better emphasize differences in degree of conservation of different alignment positions. Therefore, to return the screen to color, the screen properties must be reset. Below is a description of how to alter screen properties.

1. Choose *Properties* from the *File* menu.
2. In the dialog box, enter or alter:
   -Font size for sequence names and text.
   -Editing mode (insert or check).
   -Color type (monochrome, character=color, alignment color mask) .
   -Message panels (activates a variety of messages displayed on screen).
   -Screen inversion (inverts color patterning, very useful for manual alignments).
   -Vertical scroll locking (keeps vertical positions of split screens together).
   -Key clicks.
   -Insertion point (to the right or left of cursor).
   -Scale.
3. Click the *OK* button when done.

3.1.16 Using Sequence Masks

Sequence masks are used to determine which alignment positions of the selected sequence(s) or regions(s) will be used by programs selected from the GDE menus. When a sequence mask is selected along with sequence(s) or region(s) of sequence(s), GDE first filters the sequence(s) prior to running whatever external programs are selected. The filter removes all alignment positions at which there is a 0 in the sequence mask. Sequence masks are particularly useful for phylogenetic analysis (3.3.5). Sequence masks can be generated either manually (by creating a new sequence of the mask type and typing in the 1's and 0's) or by running the *sequence consensus* function (3.3.11) and using it to generate a mask by degree of conservation. Masks can be incorporated into any function of interest by simply including a line in the *.GDEmenus* file to tell GDE to use a mask if selected. Some programs will not use masks and thus masks

that are selected will be treated as any other sequence and no filtering will occur.

3.2 Editing and Aligning Sequences

3.2.1. Manual Alignment and Sequence Editing

GDE allows simple and easy editing, entering and manipulation of sequences and alignments. Some of the tools to remember when attempting manual alignments with GDE include

1. Amino acids are color coded by chemical group.
2. Sequences can be grouped and ungrouped to allow modifications to many at once.
3. Split screens can be used to view the affects of upstream changes on downstream alignments.
4. The keyboard may have Cut, Copy, and Paste keys that can be used instead of menu commands.
5. Repeat counts can be used to avoid overtyping and to allow for precision with large numbers.
6. Screen properties can be adjusted to aid viewing.
7. Gaps can be inserted with -,~, or the space bar.
8. Check sequence protections (if sequences are grouped then the modification being done must be allowed for all of the sequences).
9. The insertion point (before or after the cursor) can be controlled from the *Screen Properties* menu (3.1.15).
10. Save often.

3.2.2. Automated Gap Removal

This function is particularly useful when sequences being studied have been extracted from an alignment containing many additional sequences. This may leave gaps in all of the sequences being examined which, depending on the gap size, may make analysis and viewing somewhat burdensome. In addition, it is sometimes helpful to remove gaps in sequences prior to running external programs - some programs are sensitive to gap position and gaps may influence results. Be careful to save prior to removing gaps from an important alignment because compression cannot be undone. The *preserve alignment* choice can be used to remove gaps only at positions where all selected sequences have gaps.

1. Select sequence(s)
2. Choose *Compress* from the *Edit* menu
3. Choose *Preserve alignment* or *Remove all dashes*

3.2.3. Finding sequences

The *Find* function allows a user to find specific sequence string or sequences with similarity to a particular search string. The method is described below (the *Find* dialog box is shown in Fig. 2).

1. Select sequence(s) to search.
2. Select *Find all* from *DNA/RNA* menu.
3. Type in search string into text line.
4. Select search features: percent mismatch allowed, case sensitivity, if U=T, match and mismatch colors.
5. Click the *OK* button.
6. Matches are highlighted in alignment window. Multiple pages of an alignment must be scanned through manually to find highlighted regions.

3.2.4. Clustal Alignment

The *clustalv* multiple sequence alignment program (Higgins et al. 1992) is included as part of the GDE package. It allows multiple sequence alignments to be done with any number of sequences and allows the user to choose from a variety of alignment and output parameters. When run from the GDE window, *clustalv* will align sequences in the background and return the alignment in a new GDE window. Unfortunately, when doing this some of the reference information in sequence files may be lost because *clustalv* has to convert between formats. Below are the instructions for running *clustalv* from the GDE window. It can also be run from the Unix prompt by typing *clustalv* followed by a carriage return. More information about *clustalv* is found in the help file included with GDE (accessible by clicking help from the *clustalv* dialog box). A new and improved version of this algorithm, *clustalw* is now available for Unix machines (Thompson et al. 1994) and can be readily incorporated into GDE (see section 4).

1. Select sequence(s) or region(s).
2. For DNA alignments choose *Clustal* from the *DNA* menu.
3. For protein alignments choose *Clustal* from the *Protein* menu.
4. Enter alignment parameters in the dialog box (Ktuple size, Window size, Gap penalties).
5. Click the *OK* button.
6. When the alignment is done, the results will return in new GDE window.

3.2.5. Other Alignment and Editing Tools

GDE comes with a variety of additional sequence analysis tools which will not be described in detail here. They include methods for:

1. Sequence reversal (choose *Reverse* from the *Edit* menu).
2. DNA complementation (choose *Complement* from the *DNA /RNA* menu).
3. Changing text case (choose *Change Case* from the *Edit* menu).
4. Diagramming alignment on one page (choose *Strategy View* from *Seq. management* menu).
5. Contig assembly (choose *Assemble contigs* from *Seq management* menu).
6. Finding restriction sites (choose *Restriction sites* from *DNA/RNA* menu).

3.3. Sequence analysis

3.3.1. Translation

1. Select sequence(s) or region(s).
2. Choose *Translate* from the *DNA/RNA* menu.
3. In the dialog box, choose minimum ORF size, reading frame(s), genetic code, aa abbreviation, and whether ORFs should be entered as one or separate sequences.
4. Amino acid sequences will appear as new sequences in the same window. They will be given a name based on the name of the sequence they were translated from with a number indicating the reading frame (see Fig. 1).

3.3.2. Secondary structure prediction

The *MFOLD* program is an RNA secondary structure prediction program designed by Michael Zuker (e.g. (Zuker 1994)). GDE is set up to use the default settings for this program, and to pass the output through the *Zuk_to_gen* program which converts the predicted structure to a series of nested brackets. This notation can then be used for the *highlight helix* (3.3.4) and *draw secondary structure* functions (3.3.3). Depending on the size of the sequence the process may take a long time. To run *MFOLD* from the GDE Window:

1. Select sequence(s) or region(s).
2. Choose *MFOLD* from the *DNA/RNA* menu.
3. Choose linear or circular RNA.
4. Click the *OK* button.
5. Results will appear in a new GDE window like the one in Fig. 4.

3.3.3. Secondary structure drawing

This function invokes the LoopTool program to convert an alignment of a sequence and a series of brackets identifying base-pairs to a drawing of a secondary structure. The helix information must be coded in a text file with the base pairs coded by a nested series of left and right brackets. This information can be typed in manually or generated by a secondary structure prediction program (e.g., see 3.3.2). The sequence with the helix information should be named HELIX. To run LoopTool from the GDE window:

1. Select the DNA or RNA sequence and text sequence with the helix information.
2. Choose *Draw Secondary Structure* from the *DNA/RNA menu.*
3. Structure will show up an LoopTool window (see Fig. 5).
4. Many parameters can be modified from within the LoopTool window using the pop-up menus.

3.3.4. Highlighting helix

This function is used to identify regions of a sequence that do not fit into a predicted secondary structure.

1. Select DNA or RNA sequence(s) and the text file with the helix information.
2. Choose *Highlight Helix* from *DNA/RNA* menu.
3. Click the *OK* button.
4. Non-canonical base-pairs (including G:U) will be highlighted.

3.3.5. Phylogenetic Analysis

The GDE package comes with two phylogenetic programs - the least squares method of De Soete (De Soete 1983) (as implemented by Mike Maciukenas in the *lsadt* program) and the comprehensive *Phylip* package (Felsenstein 1989). The De Soete method is a distance-based phylogenetic reconstruction method. Phylip includes programs for distance, parsimony, and likelihood based phylogenetic reconstructions as well as a variety of other tools related to phylogenetic analysis. The first step in any molecular phylogenetic analysis is alignment of homologous sequences. This is an important step because each alignment position is assumed to include residues that are homologous in the

different sequences (that is, not only are the sequences assumed to be derived from a common ancestor, but so are the individual alignment positions). Therefore, regions of an alignment for which homology is ambiguous should be excluded from phylogenetic analysis. This can be done easily with GDE through the use of sequence masks (3.1.6). Multiple masks can be made for a particular alignment and different ones can be used simply by selecting the mask of interest along with the sequences to be analyzed. Below is a brief description of how to run *Phylip* parsimony or likelihood phylogenetic methods as well as the De Soete method. Running *Phylip* distance methods includes essentially the same steps as for the parsimony methods and thus will not be described.

Phylip

1. Select sequences or regions to be analyzed.
2. Select mask (if desired).
3. Choose *Phylip* from the *Phylogeny* menu.
4. Choose program to run (for details see *Phylip* help files accessible from the *Phylogeny* menu).
5. Select if bootstrapping is desired.
6. Select if a consensus tree should be made if multiple trees are generated (such as by bootstrapping or if multiple equally parsimonious trees are found in parsimony analysis).
7. Choose method of viewing result.
8. Click OK when done.
9. Depending on which items were selected, a series of windows will be opened and the user will be asked to input instructions into the *Phylip* programs that have been launched. For instructions for these programs see the *Phylip* help files.
10. When each *Phylip* programs' instructions are completed, the program will be run, and when it is done the next *Phylip* program needed will be launched. For example, if bootstrapping is selected, first the *seqboot* programs menu will be opened and once completed, *seqboot* will be run to generate the bootstrapped sequence files. When this is done, the output will be loaded into the phylogeny program that has been selected and this programs' menu will appear. When this menu is completed the phylogeny program will be run. When this is done, if consensus has been selected, the trees will be loaded in by the consensus program and so on. It may seem complicated, but doing this through GDE is a lot easier than doing it through the Unix command line.

DeSoete

1. Select sequences or regions to be analyzed.
2. Select mask (if desired).
3. Choose *DeSoete* from the *Phylogeny* menu.
4. Choose distance correction method (Olsen, Jukes Cantor, none).
5. Choose initial parameter estimate.
6. Choose random number seed
7. Choose method of viewing (Treetool or text).

3.3.6. Treetool

Treetool is a phylogenetic tree drawing and manipulation program that has been incorporated into the GDE package. It allows viewing, rerooting, coloring, reshaping and many other activities to be performed on phylogenetic trees. In addition, it allows trees to be saved in a variety of formats (including PICT format for import into graphics programs). The *treetool* menus and functions are accessed in essentially the same way as those of the GDE window. A comprehensive help file is included and is accessible from the help menu button. An example treetool window is shown in Fig. 6.

3.3.7 Email servers

While it is sometimes useful to conduct all sequence analysis on a local machine, analysis on remote machines has become an important tool for many researchers. Using remote computers, especially those set up by government or private institutions can be very beneficial -- databases are updated and improved constantly, programs are maintained and modified, and the computers are usually very fast and powerful. There are now probably hundreds of locations set up for performing anonymous remote sequence analysis. Some of these have been incorporated into the current GDE release. One of the great advantages of GDE is that new email servers and WWW servers can be incorporated almost instantly by modifications to the *.GDEmenus* file (see section 4). The email servers currently built into GDE include *blast* searches and sequence retrieval from NCBI, *fasta* and *blitz* searches through EMBL and GeneID and Grail searches. To perform these searches, simply select the sequence(s) or region(s) of interest and select the desired function from the Email menu. Be careful - some of the email services are designed for only one sequence at a time, while others can use single sequences or alignments of sequences.

3.3.8. Local Database Searches

GDE includes two programs for local database searches - *blast* and *fasta*. In addition to running them from the GDE menu they can also be run from the Unix prompt.

1. Select sequence(s) or region(s).
2. For DNA or RNA searches choose *fasta*, *blastn* or *blastx* from *DNA/RNA* menu.
3. For protein searches choose *fasta*, *blastp*, *tblastn*, or *blast3* from the *Protein* menu.
3. Choose matrix, number of alignments, database, and other parameters from the dialog box.
4. Click the *OK* button.
5. Results will appear in new GDE window.

3.3.9. Variable positions

This function allows the user to identify and highlight regions of a DNA or RNA alignment of different degrees of sequence conservation. To run the *Variable Positions* program, select the sequence(s) or region(s) or interest and select Variable Positions from the DNA/RNA menu. An example result is shown in Fig. 7.

3.3.10. DotPlot

Dotplots are a way of quickly identifying regions of sequence similarity either within or between sequences.

1. Select sequence(s) or region(s).
2. Choose *Dot Plot* from *DNA/RNA* menu.
2. Click the *OK* button from the dialog box.
3. Results will return in Plot window (see Fig. 8).
4. Properties can be altered from within this window, by clicking the *Properties* button.

3.3.11. Sequence Consensus

This program can be used to determine a consensus sequence from an alignment of DNA or RNA sequences. In addition it can be used to generate a sequence mask. The mask generated by this program will include 0's at sites of low conservation and 1's at sites of high conservation. The cut off degree of conservation can be set from within the dialog box. To use the *sequence consensus* program select *Consensus* from *DNA/RNA* menu.

4. User modification - adding additional programs

The majority of the GDE window is set up by an easily modifiable file called *.GDEmenus*. The GDE program will look for a *.GDEmenus* file first in the local directory in which the GDE program was started, then in the home directory of the user, and finally in the GDE_HELP_DIR. It will use the first *.GDEmenus* file it finds. The *.GDEmenus* file provides three pieces of information for the GDE program: 1) the instructions for the name and appearance of the GDE menus and dialog boxes that are shown once commands are selected 2) the Unix commands that will be run after a command is selected and the dialog box is completed and 3) the format that the selected sequence(s) or region(s) will be used as input into the Unix command.

The language used by the *.GDEmenus* file is a simple menu description language that is read by the GDE program. The general approach is to set up *items* within a *menu* (or make a new menu) and then to include instructions for the desired dialog box inputs (*args*) that will be used to provide specific components of the Unix commands of each item. Multiple *args* can be used for each *item* and multiple *items* for each *menu*. In addition, it is necessary to specify the input and output formats that will allow the chosen programs run by the Unix commands to analyze the sequence(s) or region(s) selected. Below is a description of the line commands that can be placed in the *.GDEmenus* file. Text in bold must be typed in as shown. Text in normal print can be modified by the user. Sections in italics are descriptions of what the line is used for.

Any Unix commands are allowed from the *arg* line. For more complicated programs it may be necessary to invoke a window from which to run the program. For example, when the Phylip option is selected from the Phylogeny menu, first a GDE dialog box appears asking for a variety of inputs (see 3.3.5). Once selected, these commands are used to determine which Phylip programs to launch from a new window. In a new window, a Phylip menu for the Phylip program that has been selected will appear. When this is completed, the Phylip program will be run and the next line in the Unix command will be executed.

itemmenu:menu_name

This defines the name of the menu header (e.g., File, Edit, etc. in Fig. 1) All lines between this line and the next itemmenu: line are used for this menu.

item:item_name

This defines each selection within a menu (e.g. Save, Open, etc), all lines between this line and the next item: line are used for this item. As many item lines as are desired can be used.

itemmeta:meta_key

This defines meta_keys for machines that can use them.

itemhelp:help_file

This defines the name of a help file. It should include either the file name or the path plus file name if the file is not in the GDE_HELP_DIR directory.

itemmethod: Unix command

This tells GDE what Unix instructions to use once the command is selected. It can be up to 256 characters in length and can include multiple commands (separate by a ;) embedded $ variable names (defines by the arg function described below) as well as shell scripts, backgrounds processes, etc. If no arguments are specified no pop-up menu or dialog box will appear and the Unix instructions will be run automatically.

arg:variable_name

This defines the name of a variable that will appear in the itemmethod line. To include in the Unix command line use a $ before the variable name. Be careful not to have text in the Unix line with the same characters as in the variable name as the arg instructions may be placed there as well. Therefore it is usually best to use as a variable name something other than a word or a common abbreviation. The variable itself is determined by input into the popup window defined by the argtype choices below. Multiple args can be included in one item (e.g., there are six in the Find command shown in Fig. 2)

argtype:slider, chooser, choice_menu, or text

These different types are defined in section 3.1.1.

arglabel:label

This is the label that will be used for the argtype in the pop-up window (e.g. Search String in Fig. 2)

argmin:#

This defines the minimum value for a slider (e.g., 0 in Fig. 2)

argmax:#

This defines the maximum number for a slider (e.g. 75 in Fig. 2)

argvalue:#

This defines the default value for a slider or the default choice for choosers, choice_menus (0=first choice, 1= second choice, etc.) (e.g. 10 in slider in Fig. 2)

argtext:default

This defines the default text that is placed in a text line. It is useful for things such as email addresses, file names, printer names, etc.

argchoice:displayed:passed

This is used for choosers and choice_menus. The text replacing "displayed" is the label given to the button or menu choice and the text replacing "passed" is the actual value that is passed to the variable if that button or menu choice is selected. As many argchoices as desired can be entered.

in:input_file

This is the name to use in the Unix command line to represent the file of the selected sequence(s) or region(s). GDE will replace this name with a randomly generated temporary file name (invisible to the user in most cases).

informat:file_format

This is the format that the sequence(s) or region(s) will be written in the input_file. Can be either Genbank or flat.

inmask:

This tells GDE that the alignment positions of the sequences selected can be regulated by a selected sequence mask. If a mask is selected along with sequences, all alignment positions with a 0 in the mask will be removed prior to analysis. This is particularly useful for phylogenetic analysis programs.

insave:

out:output_file

This is the name to use in the Unix command line for output by external functions. GDE will replace this name in the Unix command line by a randomly generated temporary file name (invisible to the user in most cases). It is up to the external function to place results in this file if it is to be read back into GDE.

outformat:file_format

The data in the output file will in this format (colormask, Genbank, or flat). This tells GDE what format the file will be in when it is time to read it back.

outsave:

Do not remove the file after reading.

outoverwrite:

This is used to instruct GDE to overwrite sequences currently in the GDE Window. It is useful for sequence alignments.

.GDEmenus modifications and lines generated by others around the world are available from a variety of sites including some of those listed in section 1. A demo is also given in the GDE manual (see section 5). The GDE Menu Building program allows for quick additions of new instructions to an existing file. It is available from http://golgi.harvard.edu/.

5. Help and Information

A manual for GDE is included in the GDE package. The manual comes in Microsoft Word, Text, and postscript formats. Also included with the GDE package are help files for most of the external programs run through GDE. Finally, additional help can be found through an electronic discussion group (send email to Tim Littlejohn tim@bch.umontreal.ca to be added to the list) and from some of the Web sites described above.

6. Notes

In part because it is shareware, and in part because it combines such a variety of programs, GDE is not without its share of bugs. Some problems that may be important to remember include:

1. Some of the instructions incorporated into the GDE menus may try to execute programs not available on the computer being used or they may be in a different location than expected. Error messages generated by these commands can be found in the window from which GDE was started. If you get an error like "Command not found" it should be preceded by the command that was tried (e.g., textedit:Command not found). There are many solutions to this problem. First, if the program does exist on your system, then the path to find that program was probably not set up in the users *.csrhc* file. Add it. If the program is not available on your system your can try and get it. Alternatively, if you have an equivalent program you can edit the *.GDEmenus* file and insert the replacement programs name wherever the other program was listed.

2. Sequence reference information may be lost when converting between formats. Some sequence formats do not contain lines for all of the information stored by other formats

3. When using the DeSoete phylogenetic program sequences cannot have a | character in their short sequence name (the one displayed in the GDE window). The tree generating program uses this same character as part of the distance matrix input

4. When loading in multiple sequence files, the default name of the GDE file may change to the name of one of the new files. Make sure to check the name when saving.

5. The input foreign format command may be confused if their are headers (such as email, headers) on sequence files. It is best to remove such headers prior to importing.

7. Acknowledgments.

I would like to thank Steven Smith for making the GDE software package freely available and continuing to improve it. I would also like to thank all of the researchers who have made their sequence analysis programs into shareware and allowed people like myself to perform highly powerful sequence analysis without much expense.

8. References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403-410

De Soete G (1983) A least squares algorithm for fitting additive trees to proximity data. *Psychometrika* **48**: 621-626

Felsenstein J (1989) PHYLIP -- Phylogeny inference package (Version 3.2). *Cladistics* **5**: 164-166

Higgins D, Bleasby A, Fuchs R (1992) Clustal V: Improved software for multiple sequence alignment. *CABIOS* **8**: 189-91

Pearson WR (1990) Rapid and sensitive sequence comparison with fastp and fasta. *Meth Enzymol* **183**: 63-98

Smith SW, Overbeek R, Woese CR, Gilbert W, Gillevet PM (1994) The genetic data environment an expandable GUI for multiple sequence analysis. *CABIOS* **10**: 671-675

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673-4680.

Zuker M (1994) Prediction of RNA secondary structure by energy minimization. In: Griffin AM, and Griffin, HG (ed) Methods In Molecular Biology, Vol. 25. Computer Analysis Of Sequence Data. Humana Press Inc., Totowa, New Jersey., pp 267-294.

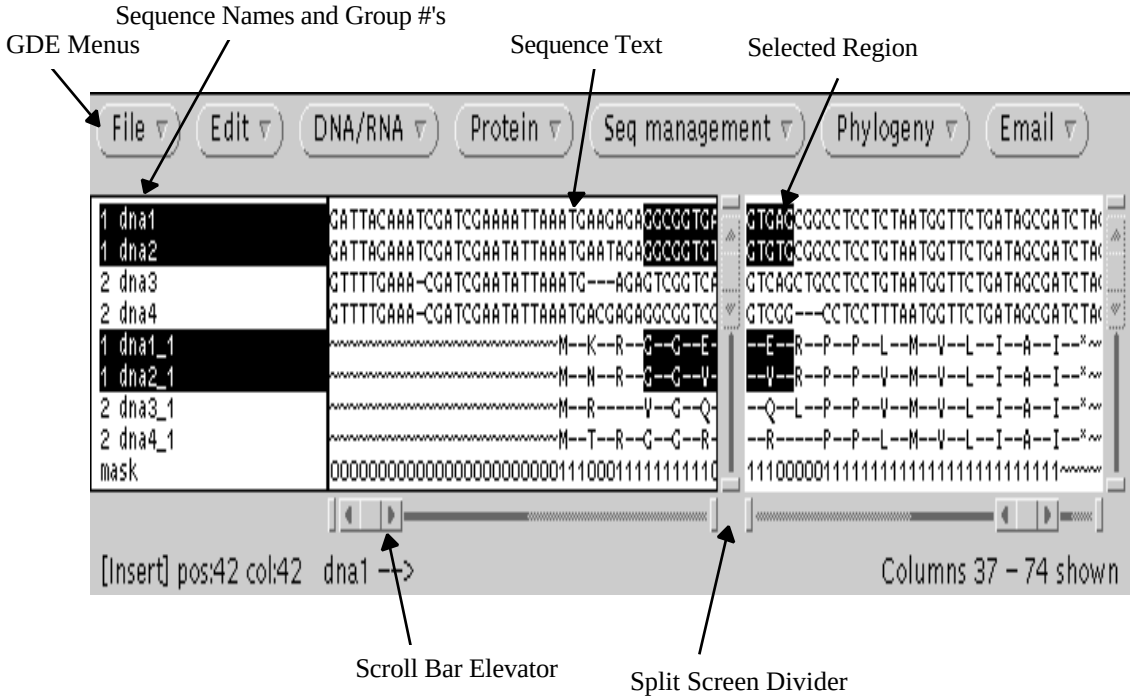

Fig. 1 GDE Window.

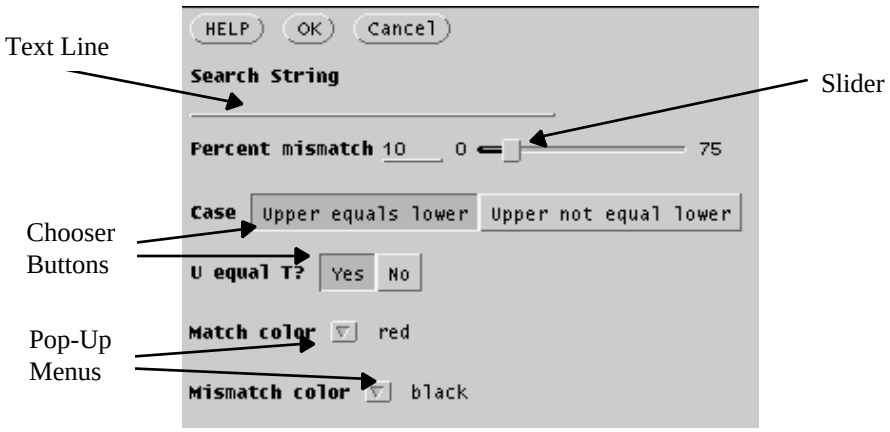

Fig. 2. Dialog Box for Find Command

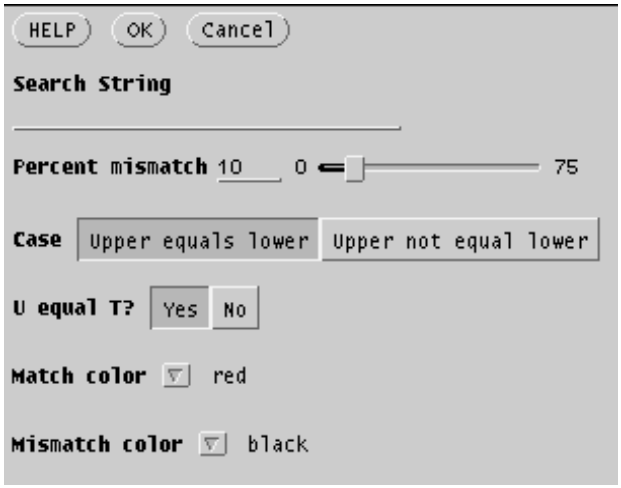

Fig. 3. Sequence Information Dialog Box

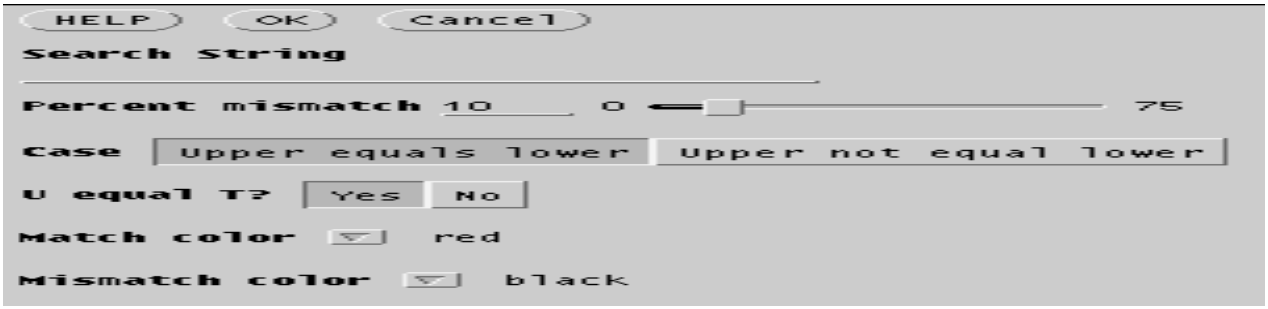

Fig. 4. Alignment of sequence and secondary structure brackets.

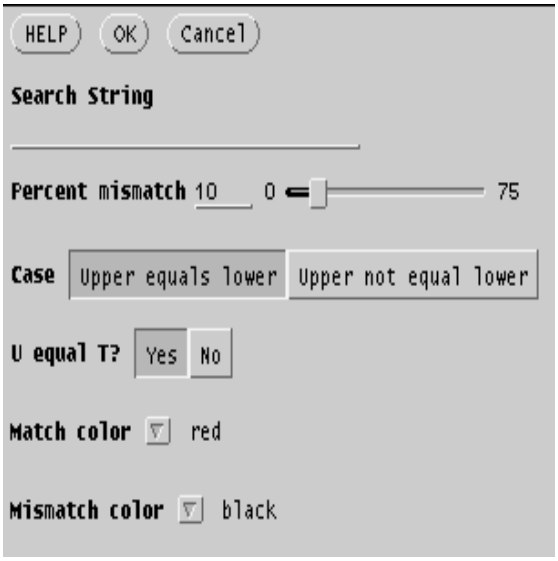

Fig. 5. LoopTool Window

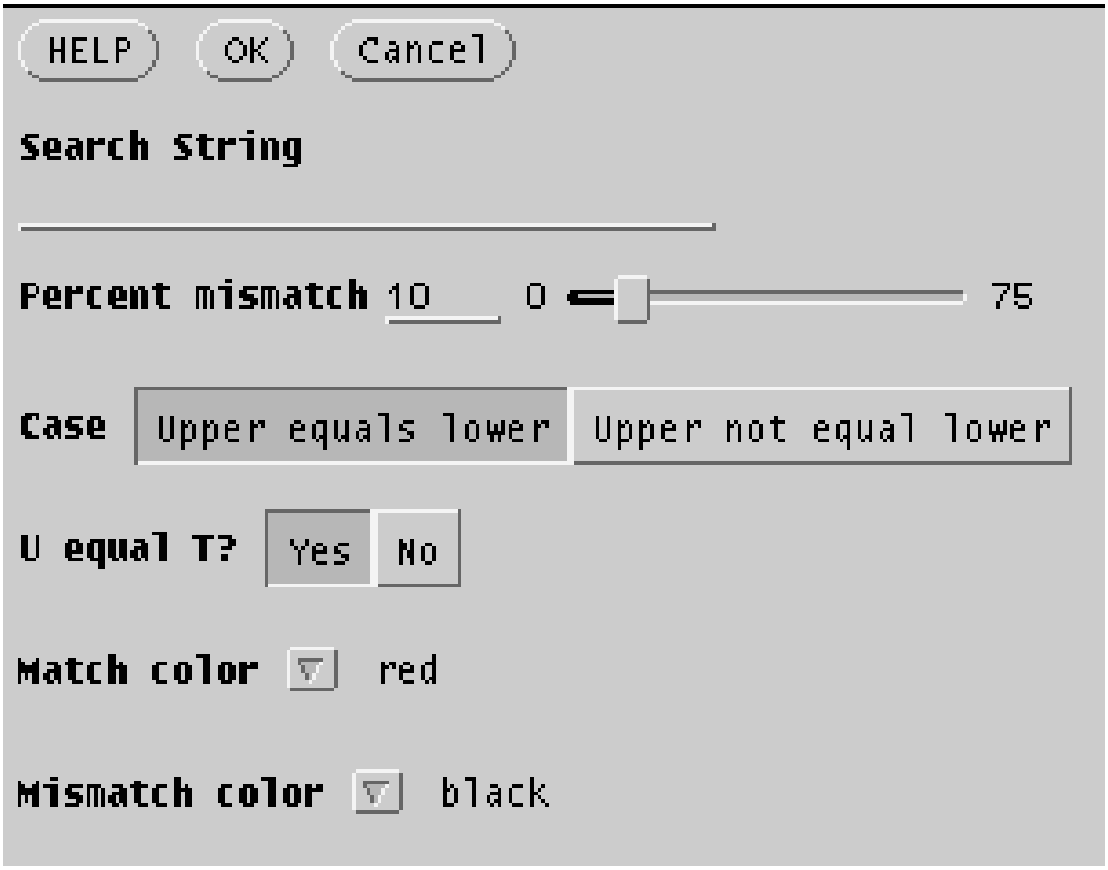

Fig. 6. Treetool Window

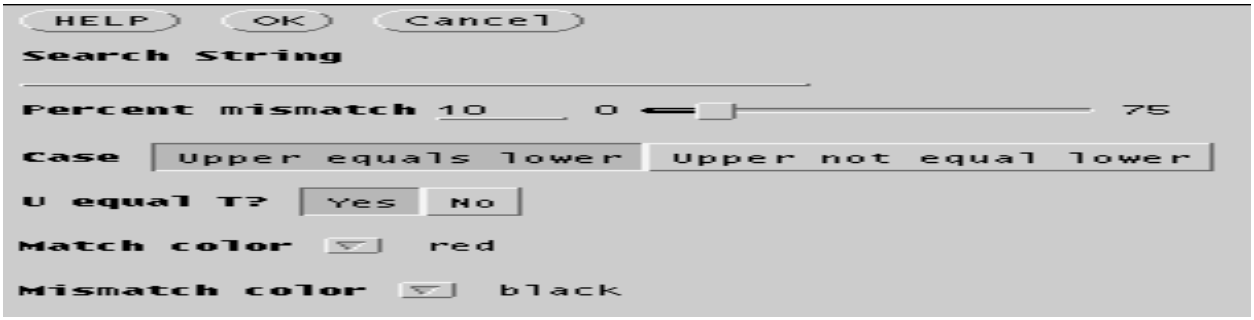

Fig. 7. Variable positions.

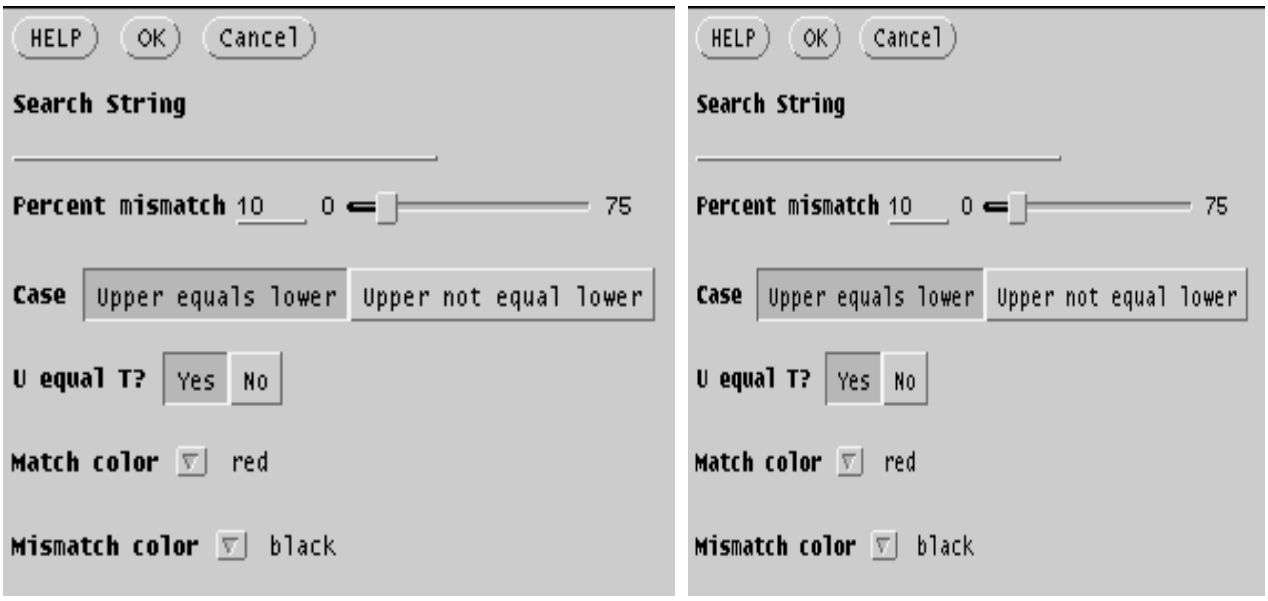

Fig. 8. DotPlot Window and Properties Dialog Box