

GCG

July 31, 2002

Sequence Analysis & Consulting Service

Seminar Series

Susan Johns

(johns@cgl.ucsf.edu)

Launch a telnet session on socrates by clicking on the socrates icon in the long window on the left hand side of the screen and responding with your account name and password.

Then, copy over to your account some data to work with by entering the following commands. [The first one does the actual copying. And the second one moves you into this area with the data.

```
gcg_setup<rtm>  
cd explore<rtm>
```

As a result of these commands, you are now in a new sub-directory called **explore** in your account where the necessary files are located. Activate the GCG software.

```
gcg<rtm>
```

GCG

GCG (the Wisconsin Package) is an integrated package of over 130 programs that allows the user to manipulate and analyze nucleic acid and protein sequences. Scientists from all over the world have collaborated to develop and refine the Wisconsin Package, making it a flexible tool with which to analyze your sequence data.

GCG is a widely used **commercial** sequence analysis software package. Since the software is licensed, there are restrictions on its usage. The software may only be used by authorized accounts of the license holder. This is the reason that SACS has a secure web site for doing sequence analysis tasks requiring username and password.

GCG General Program Areas

GCG has sequence analysis programs in the following general areas.

- Appendices
- Comparison
- Database Searching
- Editing and Publication
- Evolution
- Fragment Assembly
- Gene Finding and Pattern Recognition
- Importing and Exporting
- Mapping
- Primer Selection
- Protein Analysis
- DNA/RNA Secondary Structure
- SeqLab
- Translation
- Utilities

GCG Basics

GCG is an evolving set of programs, some of which are graphical, most of which produce their results as text files. The software has extensive documentation written from a user's perspective.

- The software suite needs to be activated only once during a computing session.

This is done by entering the term **gcg** at the socrates prompt.

The pause between entering **gcg** and the return of the prompt is caused by assigning all the terms necessary to make the software suite work.

- Programs are started by entering their name.

GCG Program Operation

GCG programs are started by entering their name. The program's first statement is a general description of what the program does. If the program doesn't do what you want it to, abort by entering **^c** at its first user prompt. [It may be necessary to press the return key to get back to the socrates prompt.]

GCG programs all follow the same general process. They prompt the user for the sequence to be used and have the user decide what portion of the sequence to use in that program. If there are any parameters to be selected, these are prompted for next. The user is asked for a name for the generated

output file containing the results of the analysis.

Some GCG programs are much more complex than others, having numerous parameters that can be user selected. In most cases, GCG runs an analysis with default settings which are appropriate for most uses of that program. It also, only asks the user about the most commonly changed parameters even though it may have a lot more available to change. To see all the possible selectable parameters, use the command switch **-check**.

```
socr<xx>:program_name -check<rtm>
```

GCG requires

- The user to have a general idea of what needs to be done.

[Use package documentation to locate the names of likely programs to use for a given purpose. GCGmanual has information listed via general topics of interest, GCGhelp lists information by program name.]

- Have a sequence to work with.

[Get a sequence by entering it into the machine with an editor, extracting it from a database or modifying an existing one.]

The sequence must be compatible with GCG. However, GCG now will use fasta formatted data files in some programs if the **-fasta** command switch is used.

[Modify the format, if necessary, so it can be used.]

Where to Run GCG Programs

GCG programs can be run in one of four ways on socrates. The means selected depends on the end user. Running this software suite on the web is the easiest, using the command line allows the most flexibility.

All the GCG programs can be run from the command line in your SACS account.

If you have xterm emulation software on the machine you use to connect to socrates, all the GCG programs can be run through the SeqLab gui interface.

A selected sub-set of GCG programs can be run from the SACS GCG Web Interface page.

Using the W2H Interface permits a wider selection of GCG programs than the SACS page, but the learning curve is higher. The W2H Interface can use a user's home directory as the source of its sequence files.

Sequence formats

Sequence analysis can be done with a number of different programs or software suites. Since these were all developed independently of one another, there is a wide variety of sequence formats available. As stated earlier, GCG will accept sequence information in its own format and in fasta format.

Usual sequence formats:

```
GCG format
GenBank format
fasta format
```

GCG provides a number of tools to allow the conversion of one format into another.

Example tasks - format conversion

In the SACS web pages file conversion is handled automatically for the user. When working on the command line, this is a point of user concern. Since there are some many possible formats, it is important that a user knows how to do format conversions.

```
genbank to gcg
fromgenbank
```

```
fasta to gcg
fromfasta
```

```
gcg to fasta
tofasta
```

Example tasks - finding a sequence

have sequence information

```
reference database searching
GCG's stringsearch
command line - local data
socr<xx>:stringsearch<rtm>

web - local data
```

have an access code (accession number)

```
database sequence retrieval
command line
GCG's Netfetch sequences from NCBI
GCG's Fetch sequences from local
databases
```

have the sequence - want to find similar ones
sequence database searching
command line
GCG's netblast(NCBI),
gcgfasta(local), wordsearch(local),
ssearch(local)

Example tasks - finding a pattern in a sequence

have actual pattern

GCG's findpatterns
command line - local data
socr<xx>:findpatterns<rtm>

web - local data

seeking established patterns

command line
socr<xx>:findpatterns -dat=genmoredata:tfsites.dat<rtm>
GCG's motifs
GCG's profilescan

web
GCG's motifs basic
GCG's motifs advanced

What are list files and how are they used?

GCG output files from various database searching programs produce what can best be described as lists of database hits. These lists can then be used in other programs as a source of desired database files to use in the analysis.

Assume that through the use of stringsearch you found a list of a number of proteins containing heme groups that you were interested in. By using this output file in the program pileup, a multiple alignment would be generated of all the proteins in the original list file. A @ symbol is used by GCG to denote that the name that follows is that of a list file.

How do I display GCG graphics on socrates?

GCG contains a number of programs that produce graphical output. The software suite expects the end user to know if a program is going to produce some sort of graphical output and to have the software's plotport set properly so that it can be displayed.

The NCSA telnet program being used on the Macs in S165A can support the display of Tektronix's color output (TEK4105). To set the plotport for this type of output, run the setplot program and select the "TEK4105" option from the list presented there. This option is the 13th one of the 130 plus given in the list. It is just

one position down from what is displayed in the window that comes up.

Now run a program that produces graphical output like isoelectric using the ccho.pir1 file as your input sequence. You will need to close the window after you are finished looking at it.

Another popular form of graphical output are postscript files. This type of output can be created by entering the term postscript at the prompt, selecting the type of postscript file desired and giving a name for the resulting postscript file. Any graphical program run after setting this up will generate postscript output having the name entered in the setup procedure.

How do I run GCG programs in the background?

There are times when the desired task is the search of a database or a series of databases. When a task is going to take a long period of time it is best to run it in the background.

Normally such a process would be set up to use the batch queue, however, this doesn't work on socrates. Therefore a user has really only one option. That is to create an input file that contains all the responses to the prompts that the program expects and use this as the input file to the desired process.

Assume that a findpatterns run for the pattern cxxch was desired on the GenPept database. There are close to 600,000 protein sequences in that database and findpatterns is a slow program.

The best way to figure out what to put in such an input file is to run through the program a number of times with dummy data to determine what all the necessary responses are.

example input file

After activating a gcg session a user would then enter the following command to have the process take place.

```
socr:<l0> findpatterns < findpatterns.in -nomon > /dev/null &<rtm>
```

Note: If you have a number of large jobs to do, create a script so that they run sequentially and you won't take over socrates.

How do I use SeqLab on socrates?

If you are using the proper type of emulation software, you can run x-term software such as SeqLab. The Macs in S165A have such x-term emulation software on them.

Start the MacX software on your Mac. The Macs in S165A have a MacX Application folder in the SACS folder contained in the long window on the left hand side of the screen. Double click on that folder to gain access to the MacX icon to launch the program. Close the folder when this process is finished.

Start up SeqLab by entering `seqlab &<rt>` at the prompt.

Two windows result from this process, click the OK button in the "About SeqLab" window (the smaller of the two) to start the SeqLab session. Normally a user is working from a `working.list` when they run SeqLab. If the program does not start in "Editor" mode, click on the mode button and select "Editor".

In this example, sequences will be files from the PIR database. Click on "File" to get the menu and select "Add Sequence from" option followed by "Databases". Move the scroll bar down until PIR becomes visible on the list. Click on the PIR name to have a black block appear at that position. In the "Database Specification" window replace the * with the term `ccho` and then click on the "Add to Main Window" button. The sequence appears in the main window. Repeat this process with the terms for `cca` and `ccms` until you have these three sequence displayed in the main window. Then click on the "Close" button of the SeqLab Database Browser window.

Notice how similar these three sequence are. By clicking on the first name in this window, that data becomes active. Position the cursor on the first character of the sequence and press the

period key. A dot appears in the sequence. Repeat this process with the second sequence. Use the scroll bar at the bottom of the main window to move through the sequences and check just how similar the three sequences now are to one another.

Highlight all three sequences by moving the cursor to first name in the list and while keeping the mouse button pressed down move down the list. All three sequence names are now highlighted. From the "Functions" menu, select "Multiple Comparison" and then "Pileup". This gets you to the data entry window for the pileup program. Click on the "Run" button to get the program to run the multiple alignment on these three sequences.

The main window is returned, followed quickly by three additional windows that placed on top of the main window. The bottom window contains a dendrogram of the alignment. In the middle window is a listing of the files produced by the process. And the top window contains the pileup program's results.

First, look at the dendrogram and when you are finished click on that window's "Close" button. Highlight the figure file in the list window and click on the "Delete from disk" button to remove it from your account and then select the "Close" button of that window. Scroll down the alignment data in the top window. When finished, click on the "Close" button for that window.

To get out of SeqLab, select "Exit" from the "File" options. Don't save your results. The window disappears and there is a done statement in your telnet session window.

SeqLab is a very powerful interface to the GCG package. It does have a learning curve and can be tricky for any beginning user. If you use Macs in your lab contact SACS about getting MacX.

URL's used in the seminar:

GCG home page: <http://www.gcg.com/>
GCG Manual: <http://www.sacs.ucsf.edu/Documentation/gcghelp/gcgmanual.html>
GCG Help: <http://www.sacs.ucsf.edu/Documentation/gcghelp/gcghelp.html>
SACS GCG Web Interface: <http://www.sacs.ucsf.edu/Resources/webgcg/index.html>
SACS's W2H Interface: <https://www.sacs.ucsf.edu/secure/cgi-bin/w2h/w2h.start>