



INTRODUCTION TO GCG

The aim of this class is to run a typical GCG programs. Since GCG programs all work in the same way, you should be able to run almost any GCG program after this session.

Note: Older accounts vary greatly in the way they are set up. All instructions in this tutorial assumes that you have an account created after 1993. To update your account, refer to the instructions in the ACHS-305 "Introduction to UNIX" handout.

Login to your account

GCG operates from the command line of your UNIX account. Login to your UNIX account. If your default is to run **Umenu**, choose "**Go to UNIX**"

Set up your default printer

GCG will print to your default printer, and if it is not set, some of the programs may stall. In this session we will not use the printer, so you do not need to set it now. For instruction on how to set your default printer, refer to the appendix at the back of this document. The default printer needs to be set only once and will remain set for future sessions.

Create a new directory

It is good practice to keep all files associated with one project together in one directory. UNIX commands are available at any time during the session.

Create a new directory called "tutorial1" by typing

```
mkdir tutorial1
```

To see if the directory is there type

```
ls
```

Tutorial will be listed as "**tutorial1/**"
The "/" symbol indicates that it is a directory.

To move to the "tutorial1" directory type the command

```
cd tutorial1
```

The word "tutorial1" will be added to your prompt to show that you are there. (To move back to your home directory type "**cd**" or "**cd ..**" at any time.)

Start GCG

To start GCG type **gcg**

After a pause, you will be rewarded with the GCG banner. Take a moment to look at it. It provides information on the latest updates of the databases and which databases are available.

```

This program sets up GCG and Phylip to run in a ksh subshell.
Type "exit" when you are finished working with GCG and Phylip.
Initializing GCG now....Please wait....

      Welcome to the WISCONSIN PACKAGE
      Version 8.0.1-UNIX, September 1994
      Installed on aix

Copyright 1982, 1983, 1984, 1985, 1986, 1987, 1989, 1991, 1992, 1994
Genetics Computer Group, Inc. All rights reserved.

Published research assisted by this software should cite:

      Program Manual for the Wisconsin Package,
      Version 8, September 1994, Genetics Computer Group,
      575 Science Drive, Madison, Wisconsin, USA 53711

Databases available:
GenBank           Release 86.0 (12/94)
PIR-Protein       Release 43.0 (12/94)
SWISS-PROT        Release 30.0 (10/94)
PROSITE           Release 12.1 (10/94)
Transcription Factor (TFD) 7.3 ( 9/93)
Restriction Enzymes (REBASE) ( 8/94)

Help is available with the command % genhelp or by
calling (608) 231-5200 or sending e-mail to Help@GCG.Com

Plotting Configuration set to:
      Language: psd
      Device: LASERWRITER
      Port or Queue: |lpr -h -P achs_l2
avery: /home/jdr8n/tutorial11 >GCG>

```

This is your default printer

Note that >GCG> was added to my prompt to show that the GCG shell is active

Running a typical program

There are over 135 programs available in the GCG package. A list of GCG programs, including short descriptions of each, is available from Academic Computing/Health Sciences (document ACHS-310).

To see a list of all the available programs, type

genhelp

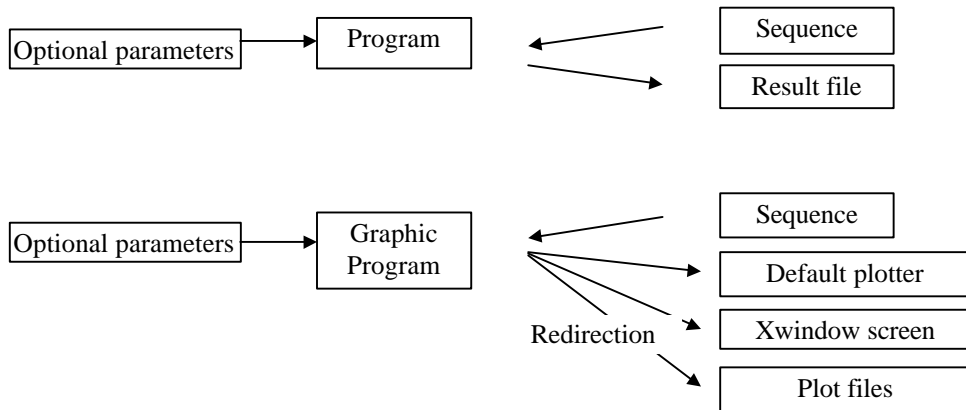
To see a list of all available programs, grouped by function, type

genmanual

A complete online manual is available at:

<http://www.med.virginia.edu/achs/molbio.html>

All GCG programs work basically the same. The program will operate on a file, normally a sequence file, and write the results to another file. The results are not written to the screen. To see the output of a program, you need to view the output file, normally the UNIX command "more filename".



The following is an example of a text file with a heading.

You may add as much text and <return>s as you want.
The heading and the sequence is separated by the
double-dot symbol. Any text after that is regarded
as sequence

```
..  
accccagagtgggtgctcccactgctgtgaaacataaaacaaaaagacctca  
atcgcaaccacagtcacaaaatggcaagatatagacgacacagcaggagcc  
ggagtaggagcagataaccgacgccggaggagaagaagaagcagacacccat  
aatcgaaggaggacatacaggagaagccggagacattcgagaaggagaag  
aggaagaaggagaggttaattgtttgggctggagaggtaatagtgggtttg
```

The reformatted file will look as follows:

```
You may add as much text and <return>s as you want.  
The heading and the sequence is divided by the  
double-dot symbol. Any text after that is regarded  
as sequence  
enigma Length: 631 March 23, 1995 11:36 Type: N Check: 3662 ..  
  
1 accccagagt ggtgctccca ctgctgtgaa acataaacia aaagacctca  
51 atcgcaacca cagtcaaaaa tggcaagata tagacgacac agcaggagcc  
101 ggagtaggag cagataccga cgccggagga gaagaagaag cagacacccat  
151 aatcgaagga ggacatacag gagaagccgg agacattcga gaaggagaag  
201 aggaagaagg agaggttaatt gtttgggctg gagaggtaat agtgggtttg  
251 cagttactgc tcctagaccc cttaatagag aggaacggtt cctaaggtct //etc.
```

Fetching a sequence from the Genbank database

To fetch the X01832 sequence from the database type

```
fetch X01832
```

The file X01832.gb_ro will be placed in your directory. (gb_ro refers to the rodent subdivision of GenBank where the sequence is located.)

If the GenBank file contains a translated peptide sequence, a second file called X01832.gp_ro will be placed in your directory. (gp is the GenPept database that contains translated protein sequences from GenBank.). *Not all GenBank sequences contain translated protein sequences and entries in GenPept.*

To type X01832.gb_ro every time you want to use the sequence is awkward. To fetch the sequence and redirect the output to a more sensible file name type

```
fetch gb:X01832 -out=sample.seq
```

The “-out=sample.seq” option works with all commands.

HINT1: An accession number always consists of a letter of the alphabet followed by 5 numerals or 2 letters of the alphabet followed by 6 numerals (e.g., SR004562). Anything else is not an accession number. The letter “O” is never used.

HINT2: To fetch the translated protein sequence of X01832 type `fetch gp:X01832`, or better still `fetch gp:X01832 -out=sample.pep`.

Mapping restriction sites

To map restriction sites in the sequence “enigma” type

```
map enigma
```

The program will prompt for the range of the sequence to analyze.

```
Begin (* 1 *) ?          Press <return> to accept the default of 1, or enter your own value.
```

```
End (* 631 *) ?        Press <return> to accept the default of 631, or enter your own value.
```

```
Enzyme(* * *):        <return> will select all enzymes  
?                   provide information on the available enzymes  
EcoRI               will select a particular enzyme Eco R1 (Note the format!)
```

```
n                   To select no translation type
```

The program will prompt for the name of the output file. Press <return> without typing a name, to use the default file name “enigma.map.” When the program is finished, you will return to your prompt.

Type **more *enigma.map*** to see the results of the analysis. (To print this file to your default printer, type **lpr *enigma.map***.)

The real power of the GCG programs are hidden in the command line options. To see which options are available type

```
genhelp map          A table of available information will be displayed. You can look at any of these  
                      items by typing the first couple of characters of the item and <return>.
```

To see the available optional parameters type

```
opt                 Hint: typing genhelp map opt<return> is faster.
```

If you want to limit your search to restriction enzymes that have 6-base recognition sites, type

```
map enigma -six
```

To limit your search to 6-base cutters, that cut only once, type

```
map enigma -six -once
```

Many more options may be set, e.g.

- **-miss=1** to introduce one base mismatches, useful for point mutations.
- **-app** provides information on the restriction enzymes used, including their cutting sites and suppliers.
- **-d** accepts all the defaults of the program.
- **-out=** redirect the output to another file name.

If you want to select the defaults of all the prompted parameters add “**-default**” to the command line (**-default** or **-d** is a global parameter that will work on all programs). The output can also be redirected with **-out=** (see the example of the fetch command) Type

```
map enigma -six -once -out=sixbase.map -d
```

Other programs that map restriction enzymes:

- **mapsort** arranges restriction fragments on order of size.
- **mapplot** prints a graphic representation of the cutting sites.

Searching Databases

Searching databases for text

To search the databases for a particular author or enzyme name type

lookup

Choose a library or accept the default (**All libraries**) by pressing **<return>**

The following screen will be displayed:

```
Complete the query form below:

      All text:
      Definition:
      Author:
      Keyword:
      Sequence name:
      Accession number:
      Organism:
      Reference:
      Title:
      Feature:
      On or after (dd-mmm-yy):          On or before (dd-mmm-yy):
      Shortest sequence length:        Longest sequence length:

      Inter-field operator:  AND          Form of output list:  Whole Entries
```

Press <Ctrl>D to continue.

Use single words, or words connected with “&,” “|” or “!” to fill in the blanks and press <Ctrl>d

Keep the following guidelines in mind as you write LookUp queries:

All queries are case insensitive.

& specifies AND. A & B means find all entries that contain both A and B

| specifies OR. A | B means find all entries that contain either A or B.

! specifies BUT-NOT. A ! B means find all entries that contain A but not B.

() Parentheses group expressions to evaluate first. For example, Smithies & (Slightom | Blechl) searches for sequences containing Slightom or Blechl. Then, out of those sequences it searches for those which also contain Smithies.

? any single character. The value s?ith includes Smith, Slith, Sjith, etc., but not Sith.

***** anything or nothing. The value *smith* includes Smith, Hocsmith, Smithies, etc.

literal query Adding a pound sign (#) to the value, for instance pseudo#. will only find those entries where the word pseudo occurs by itself.

An example of a list file called lookup.list:

```
LOOKUP in: swissprot,pir,embl,genbank,em_tags,gb_tags of: "[SQ-ALL: histone* &
h1*]"

492 entries January 19, 1996 15:22 ..

SWISSPROT:B4_XENLA ! ID: 730d0001
! DE B4 PROTEIN (HISTONE H1-LIKE PROTEIN).
SWISSPROT:H101_CHICK ! ID: 303a0001
! DE HISTONE H1.01.
SWISSPROT:H103_CHICK ! ID: 313a0001
! DE HISTONE H1.03.
SWISSPROT:H10_HUMAN ! ID: 333a0001
! DE HISTONE H1' (H1.0). //etc.
```

To fetch a sequence e.g. B4_XENLA type

fetch SWISSPROT:B4_XENLA

To fetch all the sequences in the lookup.list file, type

fetch @lookup.list

Searching databases for sequence similarity

To find out more about our sequence, we can search the databases for related sequences. This will allow us to find out what gene we have and what kind of animal it came from. Also, the positions of exons can normally be determined by looking at closely related, mapped genes.

BLAST is a fast search that can find closely related sequences quickly. Fasta is a slower search, but is more sensitive in finding more distantly related sequences.

To run BLAST type
blast enigma

Select option 7 “nr”

Accept all the other defaults, except limit the number of sequences in the output to 50.

Note: BLAST connects directly to NCBI and search their database. All the other GCG programs use our local database. It may therefore be possible to find sequences with BLAST that are not available with Fetch or Fasta.

To see the results of our search type
more enigma.blastn

FASTA is more sensitive, but slower program and may take several hours to complete. It is therefore best to submit a search as a batch job. (See Computer resources and running programs later in the document.)

To submit a fasta search as a batch job type
fasta enigma -bat

Accept the defaults except for the following.

Reply to “**Search for query sequence(s) (* GenEMBL:* *)**” with “**genbank:***”

Reply to “**List how many best scores**” with “**20**”

Note: This is only for to speed up the process for the tutorial, normally choose a number suitable for your sequence.

The job will be automatically submitted and the file containing the results will appear in your directory.

Warning messages refer to sequences found that may be similar to your query sequence, but were not reported due to the parameters you set.

Available Databases

For up-to-date information on local databases check the header that appears when you activate GCG. The “Short Descriptions of GCG Commands” handout (ACHS-310) contains more detailed information. The Uva Molecular Biology website (<http://www.med.virginia.edu/achs/molbio.html>) is a valuable resource.

DNA Databases

A copy of the GenBank DNA database is maintained on site and is updated daily. To specify GenBank in Fasta use **genbank:*** or **gb:***.

Note: The expressed sequence tags division of GenBank contains expressed, but unannotated sequences. To search GenBank including the est sequences, specify **genbankplus:*** or **gbp:***.

EMBL is a European version of GenBank. These two databases are essentially the same and are reconciled daily. Non redundant (nr) refers to the fact that only one copy of each sequence is kept, eventhough the same sequence may appear in more than one database. The EMBL database on site contains sequences that occur in EMBL but not in GenBank.

Protein Databases

We have already used **GenPept**, which is a protein database consisting of the translated protein sequences, contained in the GenBank sequence headers. Most sequences are hypothetical and not confirmed experimentally. Updated daily.

PIR is a protein database with experimentally confirmed proteins as well as hypothetical ones. Updated Quarterly. The nrl_3d subdivision contains sequences that have solved 3-D structures

SWISSPROT is a well annotated protein database. Updated weekly.

Translating a sequence

Type

```
translate enigma
```

Select the following options:

```
Begin:70
```

```
End:214
```

```
A) Add another exon from this sequence
```

```
Begin:438
```

```
End:484
```

This will write the translated protein sequence `enigma.pep`. Note that the name of the original sequence and the positions of the exons are retained in the header.

Exercises

- (a) Fetch the protamine P1 DNA sequence of the Tasmanian devil (*Sarcophilus harrisii*) to your account. Use `mapsort` to find an enzyme that cuts twice within the first 100 bases of the sequence.
- (b) Fetch the file called "tutorial1.txt" to your account. Remove any unwanted text and reformat the sequence to `gcg` format. Reverse and complement the sequence. Now translate the reversed and complemented sequence into a protein sequence.

The following commands are available from the command line:

[] indicates optional parameters.
n position of the nucleotide
s start position
f finish position

	EDit seqname	- get a new sequence file to edit
[n]	Include [seqname]	- insert another sequence [at position n] (SeqEd prompts for range and strand)
s,f	Delete	- delete a range of bases
[s]	Check [/Blind]	- check (or proofread) a range of bases [beginning at s]
	37	- go to base 37
	REDraw	- redraw the screen
[n]	COmment comment	- insert a comment [at position n]
[n]	COmment	- enter comment editing mode [at position n] If you do not supply a position number, the default will be the last position of the cursor in the sequence.
[n]	HEAding	- edit documentary heading [at line n]
	OVERstrike	- enter overstrike mode
	INSert	- enter insert mode
[n]	Mark markcharacter	- mark the sequence [at position n]
	PERFect	- require finds to be perfect matches
	PROtein	- set sequence type to PROTEIN
	NUCleotide	- set sequence type to NUCLEOTIDE
[s,f]	Write [seqname]	- SAVE AS , write [a part of] the sequence to a file
	Help	- show commands in screen and command modes
[s,f]	EXit [seqname]	- SAVE [a part of] the sequence AND QUIT
	Quit	- QUIT the editor WITHOUT SAVING the sequence

To edit the header, type at the command mode

head<return>

Move around with the arrow keys and make any changes you want.

To return to the command mode type <control>d

To export the portion of the sequence between position 550 and 600, type

550,600 write fragment.seq<return> (only the “w” of “write” is really necessary)

The file “fragment.seq” will be placed in your directory.

To include (or insert) a sequence at a specific position in the sequence go to screen mode and move the cursor to position 572. Type <control>d to move to the command mode. Type

include fragment.seq<return> (Only the “i” of “include” is necessary.)

If you did not want to move the cursor type 572 i fragment.seq

You will be prompted for the range and the strand of the sequence you would like to include.

Note: a comment is automatically added to keep track of insertions.

Comments are a good way to keep track of a modifications and can be added by typing at the command line

comment<return>

(Only the “co” of “comment” is really necessary.) The program will default to the last position of the cursor.

Unlike the header, comments are attached to a position in the sequence. You can supply the position by typing

484 co<return>

To **save** (or write) the current sequence, type at the command line

w<return>

To **save as** a new file name type

w filename<return>

To **exit** the editor **and save** the sequence, type

ex or exit<return>

To **quit** the editor type at the command line.

q or quit<return>

This will **NOT save** the sequence.

APPENDIX II

Setting up your default printer

To tell if your printer is set, type
`echo $PRINTER`

(UNIX is case sensitive and spaces are important. Type in commands in upper and lower case *exactly* as shown.)

If your printer is already set, skip the setup procedure.

To set your printer through Umenu type
`umenu`

Uva's Umenu system will appear.

Choose "System Customization"

Choose "Printer preference"

Choose "Type in your own selection (ksh & Umenu)."

Enter the name of your local networked printer. If you do not know the name of your local printer, you may enter the name of our local printer here at ACHS by typing

`achs_13`

Type "y" to change your UNIX command option as well.

Type "m" to return to the main menu.

For the change of printers to take effect, you have to logout of your account and log back in again.

(You do not have to do it now, but next time you start a session, your printer will be set.)

Choose "Go to UNIX"

Computer resources and running programs:

Most programs, e.g. map, will execute rapidly and can be run in real time—the default for GCG. However, some programs, such as fasta, may take a long time to run. You can run these programs in real time and watch the output. During this time you cannot run any other command and the process will be terminated when you logout or exit. There are two solutions:

1) Run the program as a batch file by including "**-bat**" or "**-batch**" in the command line, e.g. `fasta enigma -bat`.

The program will prompt you for the parameters and submit a batch job when it is complete. You may logout and log back in at a later time to retrieve your results. This is useful for people who connect through telephone lines, or if you want to run a program overnight. **This procedure is convenient, efficient, and highly recommended. You are limited to two batch processes at any one time.**

2) Run the program as a background job by including a `&` in the command line e.g. `fasta enigma &`.

This is useful for getting on with other work while a process is running. The program will be terminated when you logout. Running several programs in the background slows down the processor for everyone on that computer. The processor-watchdog (called cron) may terminate your programs if you abuse the system. Hint: make a note of the process number displayed when the program is submitted. The only way to stop a background process is through the kill command, e.g. `kill 605784`, if the process number is 605784. To see what background processes are running, type `ps -a`. **You are limited to two background processes at any one time.**