

Locating Data

January 30, 2002

Sequence Analysis & Consulting Service

Seminar Series

Susan Johns

(johns@cgl.ucsf.edu)

How do I locate data to use in sequence analysis tasks?

Log into your socrates account by clicking on the socrates icon in the small window on the upper left-hand side of the screen and responding with your account name and password.

Now, copy over to your account the necessary data. To do this, enter the following commands. The first one sets up some variables. The second one moves you into this area.

```
locate_setup<rtm>  
cd find<rtm>
```

As a result of these commands, you are now in a sub-directory called find where the necessary files are located. Activate the gcg software.

```
gcg<rtm>
```

The means used to locate sequence data depends on what information the process is starting with. Usually this process begins with information from a paper, a general topic or a known sequence. How the search is conducted depends on the tools available to use.

paper search process

Before an article is accepted for publication, the sequence information in that manuscript needs to have been submitted to a database and have an accession number assigned to it. The database and accession number are then given in the article so that interested readers can locate the data.

general topic search

Topic searches usually mean the looking through of the reference part of databases for

keywords. A user needs to be aware of the terminology used in the field. Databases evolve with time and so do descriptive terms. Also, not all database entries get revised when changes occur and so a list of search terms may be needed to find all the information on a given topic.

similar sequence searches

Finding similar sequences requires doing database searches with the initial sequence of interest. Format conversion may be necessary to get the sequence in a form compatible with the tools available.

Before databases can be searched, you need to know what they are and what they are composed of.

The databases used in sequence analysis are composed of primary sequence data. They may or may not contain reference information.

Databases can and are formatted for specific purposes.

complete or full databases

```
[contains both sequence and reference material  
for each data entry]  
(all GCG formatted databases)  
(slower character searching needed to find  
information)
```

minimal databases

```
[contains only actual sequence information]  
(databases used by BLAST or FASTA)  
(fast small window matching finds similar  
sequences)
```

paper search process

example information:

GENOMICS 45, 368-378(1997)

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under the following Accession Nos.: human MRP, L05628, AF017145, AF022824-AF022853; murine mrp, AF022908; human MOAT, U49248; rat MOAT, L49379; rabbit EBCR, Z49144; human SUR, U63421; YCF-1, L35237; Yor1/Ysr1, 1245963; Leishmania PgpA, X17154; human CFTR, M28668; human SUR, L78208; human CFTR, M55106.

Assuming that information highlighted in blue is the desired data, the choice remaining is where search for the data since that will determine the tools to be used.

The article was published in 1997 so it won't make any difference whether the search is made locally or at NCBI. Some times for very recent papers the only place the data is available is at NCBI.

local (gcg command line):
local database short names:

gb: GenBank
gp: GenPept
pir: PIR protein
sp: SwissProtein

view the sequence only:
`typedata gb:L78208<rtm>`

bring into your account (local data):
`fetch gb:L78208<rtm>`

bring into your account (NCBI data):
netfetch

NCBI (through Entrez):

From the black line on the Entrez window, click on "Nucleotide" to get to the nucleotide data searching page. In the "for" box enter the term L78208 and click on the "Go" button. The results of the search will be shown in the updated window. To see the data, click on the link. To save the data on your own machine, pick the desired format from the button currently saying "Default" and then click on the "Save" button. Depending on how your browser is set up, you will be prompted to take the necessary steps to save the data on your local machine. The most useful data formats in the menu list are GenBank and FASTA.

There is a complete documentation on using Entrez at NCBI. It is a very powerful search tool that can search PUBMED as well as databases. Click on the "Limits" link on the main Entrez page to get to a more complex searching page.

general topic searches

example information:

melittin proteins

The question now is where to search for the data. Since this is a topic search, the reference section of the databases will be looked at.

local (gcg command line):
stringsearch
lookup

local gcg web:
stringsearch
lookup

NCBI (through Entrez):

From the black line on the Entrez window, click on "Protein" to get to the protein data searching page. In the "for" box enter the term melittin and click on the "Go" button. The results of the search will be shown in the updated window. Notice that there are 57 hits. This is because the initial basic parameters for the search looks at the entire reference section of the entry, not just restricted definition lines. See the previous section for instructions on how to save data to your own machine.

similar sequence searches

Looking for similar sequences can be done on the web and locally. The software usually used for this process is either BLAST or FASTA.

NCBI's BLAST site is the most heavily used site for this purpose. It has daily updated databases, basic and advanced searching modes and different output formats. There is a web [tutorial](#) as well.

The FASTA site offers similar services using a slightly different algorithm to do the searching.

Similar sequences database searching can be done locally either on the command line or through local SACS web interfaces. The web interfaces are the easiest to use.

The major difference between using SACS web interfaces to BLAST and FASTA, and the public sites listed previously, is a matter of security. However, the SACS site may be running an older release of the software.

A minor difference is the available databases. SACS uses the term **nr** for the non-redundant protein database and **nt** for the non-redundant nucleotide database.

Locally searchable databases formatted for BLAST and FASTA searches are given at the end of the handout.

Similar sequences database searching on the command line can be done using either the GCG implementation of the BLAST and FASTA programs or by running the stand alone versions of the programs.

GCG has both a version of BLAST that runs remotely at NCBI, [netblast](#), and one that runs using the local databases, [blast](#).

The GCG version of FASTA is known as [gcgfasta](#).

example information:

FIGSALKVLAGVLPISIVSMVKQA

Assume that the sequence given above is the one that you are interested in. Problem one is to get it in shape so that you can use it.

This is a small sequence and can be easily cut and pasted into an editing session to make the conversion. The desired format is up to you. Creating a gcg usable file is a two step process, a fasta formatted file one.

create the data file

pico filename
fasta format

```
>myseq  
FIGSALKVLAGVLPISIVSMVKQA
```

gcg format

```
myseq  
..  
FIGSALKVLAGVLPISIVSMVKQA
```

reformat raw file

With the properly formatted file in hand, it is time to search with the sequence against either local or web databases.

local (gcg command line)
local gcg program - NCBI data netblast
local gcg program - local data blast

GI number searches

As part of NCBI's mission to coordinate efforts to gather biotechnology information worldwide, a new unique identifier for sequences was developed. Known as GI numbers, they were a series of digits that are assigned consecutively by NCBI to each sequence it processes.

GI, which stands for "GenInfo Identifier" was an early system used to access GenBank and related databases. A GI number was assigned to each nucleotide and protein sequence accessible through the NCBI search systems, and was a means of tracking changes to the sequence. However, GI numbers were not used uniformly across the collaborating databases (GenBank, EMBL, DDBJ). They instead served as an internal tracking system for the databases that chose to implement them.

When the collaborating databases began to formalize use of sequence identifiers, they created a new, separate field called NID (nucleotide identifier) in the database record. Similarly, the gi number for each protein sequence was named PID, and placed above each amino acid translation in the field: FEATURES/CDS/db_xref="PID:gNNNNNN". Hence, there became two types of gi numbers: NID and PID.

In February 1999, GenBank/EMBL/DDBJ implemented a new "accession.version" system of sequence identifiers that runs parallel to the gi number system. Unlike the gi number system, in which sequence identification numbers were not necessarily consistent across the databases (e.g., GenBank and EMBL could each assign their own gi number to a sequence), the new system is designed to ensure consistency. It is also designed to show a relationship between a sequence identification number and the accession number of the record in which it is found. In contrast, GI numbers are assigned consecutively and bear no resemblance to the accession number. Finally, the new system allows the assignment of alphanumeric protein IDs to proteins translations within nucleotide sequence records. The protein IDs contain three letters followed by five digits, a period, and a version number.

There are actually only two types of sequence identification numbers: gi and accession.version. Within each identifier system there are two sub-types of ID numbers, one for nucleotides and one for proteins:

NID - gi number of a nucleotide sequence
accession.version of a nucleotide sequence
contains two letters followed by six digits, a dot, and a version number

PID - gi number of a protein sequence
accession.version of a protein sequence
contains three letters followed by five digits, a dot, and a version number

The two systems of identifiers run in parallel to each other. That is, when any change is made to a sequence, it receives a new GI number AND an increase to its version number.

You see GI numbers as the result of doing NCBI Entrez searches and sometimes BLAST searches. As a real effort is being made to get users to think in terms of consistent sequence identifiers rather than other forms of sequence naming.

example:

The GenBank accession code for bacteriophage lambda is LAMCG. When this is entered in a BLAST search, the result is an error message. If an Entrez search is done for LAMCG and the found GI number used, the process works.

```
LOCUS       LAMCG             48502 bp    DNA    circular    PHG             31-OCT-2000
DEFINITION  Bacteriophage lambda, complete genome.
ACCESSION   J02459 M17233 M24325 V00636 X00906
VERSION     J02459.1 GI:215104
KEYWORDS    DNA-binding protein; circular; coat protein; complete genome;
```

BLAST and FASTA are used for database searching because they are fast. They use slightly different approaches to the problem of finding similar sequences. They both make refinements in the searching process to increase their searching speed.

BLAST is faster than FASTA, but it is also less sensitive. The default window with BLAST is 11, while it is 6 in FASTA when searching nucleotide databases. Matches can be found with FASTA that BLAST misses.

The slowest, but, most accurate database searching is done with ssearch. GCG contains an implementation of this program.

Both BLAST and FASTA can be run on the command line outside of the GCG frame work. To do this requires that a user be much more knowledgeable about the operations of the programs, their command switches and the location of usable databases on socrates. The input file used must be in fasta format.

BLAST generic command line:

```
blastall -p blastp -d swissprot -i ccho.tfa -o ccho-sw.blastp
```

FASTA generic command line:

```
fasta34 ccho.tfa s > ccho.fastap2
```

fasta's database abbreviations:

protein databases

```
N: Non-Redundant Protein no EST
H: NRDB90 Lisa Holm's 90% ID NRDB
G: GenPept GenBank's CDS's translated
P: Protein Information Resource
S: SWISS-PROT
O: SWISS-PROT (with updates)
M: Proteins New in Last Month
B: PDB seqs of known 3D structures
Y: Yeast (S. cerevisiae) proteins
C: E. coli Protein Sequences
F: Transcription Factor Database
```

AutoBlast?

There are times when what is needed is to run a BLAST search at NCBI at regular intervals against a desired database for a period of time. When this is the case, use the AutoBlast web page to set up the process.

A system process is notified that a user wants to run the autoblast process at described intervals for a period of time with a selected sequence. An **.auto** sub-directory is created in the user account to house the sequences to be used and the necessary control files. An email is sent informing the user when a job has been run and where to find the results.

For a more complete discussion of using BLAST and FASTA check out this prior SACS seminar presentation by Chris Botka.

References:

WU BLAST is developed at Washington University at St. Louis (WUSTL) by Warren Gish.

Altschul, SF, Gish, W, Miller, W, Myers, EW, and DJ Lipman (1990). Basic local alignment search tool. Journal of Molecular Biology 215:403-10.

NCBI BLAST is developing separately at the National Center for Biotechnology Information (NCBI) from WU BLAST roots.

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

FASTA software

W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

URLs used in this seminar:

Entrez

<http://www.ncbi.nlm.nih.gov/Entrez/>

Entrez help page

<http://www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html>

SACS stringsearch page

<https://www.sacs.ucsf.edu/secure/cgi-bin/strings.pl>

SACS lookup page

<https://www.sacs.ucsf.edu/secure/cgi-bin/lookup.pl>

SACS BLAST (local)

<https://www.sacs.ucsf.edu/BLAST/>

SACS FASTA (local)

<https://www.sacs.ucsf.edu/FASTA/>

NCBI BLAST

<http://www.ncbi.nlm.nih.gov/BLAST/>

NCBI BLAST tutorial

<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/tut1.html>

BLAST man page

<http://www.sacs.ucsf.edu/Training/locate/blast-man.html>

BLAST family of programs

<http://www.sacs.ucsf.edu/Training/locate/search4.html>

FASTA man page

<http://www.sacs.ucsf.edu/Training/locate/fasta-man.html>

FASTA family of programs

<http://www.sacs.ucsf.edu/Training/locate/search5.html>

AutoBlast page

<https://www.sacs.ucsf.edu/secure/cgi-bin/autoblast2.pl>

Previous SACS seminar

<http://www.sacs.ucsf.edu/Training/dbsearch/dbsintro.html>

SACS partial listing of available databases for BLAST and FASTA searching:

peptide:	
nr	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences)
nrd90	Lisa Holm's 90% ID NRDB
month.aa	All new or revised GenBank CDS translation+PDB+SwissProt+PIR+PRF released in the last 30 days.
ecoli.aa	E. coli genomic CDS translations
pataa	Patented proteins
pdbaa	Sequences derived from the 3-dimensional structure Brookhaven Protein Data Bank
pir	Protein Information Resource database
swissprot	the last major release of the SWISS-PROT protein sequence database (no updates)
yeast.aa	Yeast (<i>Saccharomyces cerevisiae</i>) protein sequences.
nucleotide:	
nt	All Non-redundant GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences)
month.na	All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
alu	Select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences. It is available by anonymous FTP from ncbi.nlm.nih.gov (under the /pub/jmc/alu directory). See "Alu alert" by Claverie and Makalowski, Nature vol. 371, page 752 (1994).
ecoli.nt	E. coli genomic nucleotide sequences
epd	Eukaryotic Promotor Database
est	Non-redundant Database of GenBank+EMBL+DDBJ EST Divisions
est_human	Non-redundant Database of GenBank+EMBL+DDBJ EST Divisions
est_mouse	Non-redundant Database of GenBank+EMBL+DDBJ EST Divisions
est_others	Non-redundant Database of GenBank+EMBL+DDBJ EST Divisions
gss	Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
htgs	High Throughput Genomic Sequences
mito	Database of mitochondrial sequences"
pdb.nt	Sequences derived from the 3-dimensional structure
sts	Non-redundant Database of GenBank+EMBL+DDBJ STS Divisions
vector	Vector subset of GenBank(R), NCBI, in ftp://ncbi.nlm.nih.gov/blast/db/
yeast.nt	Yeast (<i>Saccharomyces cerevisiae</i>) genomic nucleotide sequences