

# The WEB and Sequence Analysis

May 1, 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.

First, copy over to your account the necessary data. Move over to a place in your account where you want to want this data. Then enter the following commands. [The first one does the actual copying. And the second one moves you into this area.]

```
web_setup<rt>  
cd web_analysis<rt>
```

As a result of these commands, you are now in a new sub-directory called **web\_analysis** in your account where the necessary files are located.

### Basic Definitions

**Internet**:- the huge network that links many of the world's scientific, research and educational computers, as well as some commercial networks; also called the NET.

**WWW**:- world wide web - that portion of the Internet that uses hypertext (html) and graphics to display information.

**browser**:- the software that allows access to the information on the web.

### Accessing the Internet

- **mail servers** - submitting tasks to automated servers
- **ftp** - means of transferring data from one place to another
- **web browser** - data access, searching and task submission

### Mail Server Access

There were a number of computers with specialized software on them for doing sequence analysis tasks.

When these computers initially came on-line the only way to conveniently submit requests to them was through email.

Therefore, mail server utilities were created that allowed the loading of a sequence into the program. From the nature of the sequence, the program decided which set of possible known mail servers to use (protein or nucleotide server sites). Since the requirements for submitting something to these servers are known, the software prompted for the necessary information and sent off the request.

On socrates there is a mail server utility running called, **msu**.

```
socr:<x>msu<rt> -- fasta formatted files  
or  
socr:<x>msu -g<rt> -- GCG formatted files
```

msu screen trace

Your results are returned via e-mail.

Today, while these servers still are in service, web interfaces rather than mail server utilities are used to submit requests.

### FTP Access

The File Transfer Protocol (FTP) was developed as a means of moving data between computers on the Internet.

**ftp** software is available on socrates. It requires that a user have an account on both machines involved in the data transfer. The second account may be an anonymous one supplied by the remote machine for information available for distribution.

To use ftp, a user needs to know the name of a desired ftp site which contains information that they want to transfer.

```
socr:<x>: ftp site_address<rt>
```

FTP functions, such as automatic downloads, are being incorporated into web sites, however, major databases still depend on this process to provide access to their data.

### Browser Access

A web browser is an Internet utility that allows you "browse" the information space on the WEB. The web's fundamental building blocks are hypertext documents. To locate items of interest on the web a

search engine is used to look through the collection of materials to find what you are looking for.

examples of popular search engines:

<http://www.google.com>  
<http://www.hotbot.com>  
<http://www.altavista.com>  
<http://www.yahoo.com>

examples of commercial sites with lots of advertising and chances to win money:

<http://www.iwon.com>  
<http://www.luckysurf.com>

[these require registration and result in lots of junk email]

### Finding information on the web

You have read an article in the June 1999 Discover magazine about research being done by Catharina Svanborg at Lund University in Sweden. She has found that breast milk causes cancer cells to die.

You want to see if she has any information on this up on the web and what her email address is. Use one of these search engines listed above to help you in this task.

What about doing sequence analysis on the web?

There are a large and growing number of sequence analysis servers on the web. The trick is finding the ones you want and learning how to effectively use them. Doing a search on the web can be frustrating as well as rewarding in the locating interesting bioinformatics sites. A number of well established sites are listed on the SACS Links page and under Miscellaneous on that page.

SACS has placed a number of sequence analysis programs on its web server. This is a secure server and you need to provide your account name and password in order to use this resource. Access is through the SACS's Sequence Analysis on the Web page.

SACS has a growing number of GCG programs on its web server. GCG is commercial software and its licensing prohibits open access to the software. Local users can access it via the GCG Web Interface link on the Sequence Analysis on the Web page. There is another interface to GCG software on that page as well, W2H. This allows you to use files in your own account while working via the web.

When working on the web there are a number of concerns. It is not always easy to determine just what

sort of environment you are working in when you are on the web.

Some of these concerns are:

security  
finding data  
input formats  
database age

If you are concerned about your data being exposed on a wide open web server, only use a secure one. Find out where results are kept and go there and remove your results once the task is completed.

On socrates, gcg results are kept in [/usr/tmp/webgcg](#). The user's account name is given in a directory listing to aid in locating your data.

Finding data depends on the type you want. If searching for a topic use a resource such as ENTREZ old site or the new one or SACS's stringsearch. If searching for a similar sequence use a database searching tool such as BLAST or FASTA.

Web sites use a number of different input formats. Some want you to paste in just the sequence. Even then they might not handle blank lines or numbers in the text pasted into the input box. Others require specific input formats such as fasta, genbank, or gcg.

It is hard to tell at times just how old the databases are at a given server. You want the latest and greatest versions of the data to work with. Check any page update messages to get a feel for how old a site is, if it is being maintained and what the possible age of its data is.

Now go through and carry out the following tasks. Some will require the use of the files in your socrates account to complete.

Use the socrates [cytc](#) file to do a MultAlin run on the SACS server. Check the format of this file before hand so you know what sort of data you are using. Examine both the colored output on the screen and the actual alignment file.

Do a SACS basic motifs run and then a profilescan on the [sh17](#) sequence in your socrates account. [GCG's motifs program can be found starting at the SACS Resources page and the profilescan at the expasy site from the SACS Links page]

Search for the complete genomic sequence for Bacteriophage lambda which has the accession code [LAMCG](#) at the entrez site.

Go to the pfam site and use the [ccca](#) file on socrates that is in the most appropriate format to run a pfam search.

There are other means to do sequence analysis tasks on the web with a browser.

Establish your own list of useful sites and bookmark them on your machine. Organize the list in the bookmarks to have similar sites under a general topic. Do periodic searches to keep your list up-to-date.

Keep track of another group's list of useful tools and use them as the updating agent. One such list is Baygenomics' resources page.

Use an academic sequence analysis service such as the San Diego SuperComputer Center's Biology WorkBench. You need to register to use the service (fill out a form and create an account name and password). You can then upload sequences to the site or search their databases for sequences to use with their tools.

Use a commercial sequence analysis service such as Entigen's BioNavigator. Here you pay to work with the tools developed by the company.

URL's used in this seminar:

ENTREZ sites:

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>

pfam site:

<http://pfam.wustl.edu/hmmsearch.shtml>

profile scan:

<http://hits.isb-sib.ch/cgi-bin/PFSCAN>

SACS pages:

links page

<http://www.sacs.ucsf.edu/Links/>

miscellaneous page

<http://www.sacs.ucsf.edu/Links/misc.html>

sequence analysis on the web

<http://www.sacs.ucsf.edu/Resources/sequenceweb.html>

gcg web interface

<http://www.sacs.ucsf.edu/Resources/webgcg/>

motifs - basic

<https://www.sacs.ucsf.edu/secure/cgi-bin/analyze.pl?anatype=motifs>

multalin

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

stringsearch

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

secure blast

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

secure fasta

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

search engines:

<http://google.com>

<http://www.hotbot.com>

<http://www.altavista.com>

<http://www.yahoo.com>

sequence search sites:

NCBI blast

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

UVA fasta

<http://fasta.bioch.virginia.edu/fasta/cgi/searchx.cgi>

Baygenomics resources page

<http://baygenomics.ucsf.edu/education/workshop1/resources-full.html>

BioNavigator page:

<http://www.bionavigator.com/>

SDSC Biology Workbench

<http://workbench.sdsc.edu/>