

# BLAST

May 7, 2003

Sequence Analysis & Consulting Service

Seminar Series

Susan Johns

(johns@cgl.ucsf.edu)

## BLAST

BLAST is the most heavily used of all the database searching programs. It comes in a number of different versions. It is currently being worked on by two different sites in the US, NCBI and Washington University. The software has also been incorporated into the GCG package.

With BLAST being available from NCBI via the web, it is readily available to everyone. The questions being explored here is how it is used and what do the results mean.

First, copy over to your account the necessary data. Enter the following commands. [The first one does the actual copying. The second one moves you into this area.]

```
blast_setup<rtm>  
cd myblast<rtm>
```

As a result of these commands, you are now in a sub-directory called **myblast** where the necessary files are located.

To explore the using BLAST, start by going to the main **BLAST** site at NCBI.

Notice that NCBI has completely revised the main BLAST page. This appears to be an attempt to customize the searching process based on the nature of the sequences used (protein vs. nucleotide) and to reduce the number of failed searches.

Most of the BLAST programs now accept the user selecting a portion of the query sequence to be used in the search. Most will also allow the restriction of the database to be searched to the results of an Entrez query against the chosen database. This can be used to limit searches to subsets of the BLAST databases. Any terms can be entered that would normally be allowed in an Entrez search session.

The programs have been grouped according to query sequence type (nucleotide or protein) and the searching function being sought. New features have been added to the search process in the various groupings. Help has been integrated more closely into the submission forms.

Questions to be answered.

- What programs are available?
- What databases can be searched?
- What format does data need to be in?
- What does filtered mean?
- What is the impact of changing comparison tables?
- What is the impact of changing the expect value?
- How can alignments be viewed?

Next go to the protein BLAST site at NCBI by clicking on **the standard protein-protein BLAST [blastp]** link on the main BLAST page.

Questions to be answered.

- What are the difference between the basic and advanced versions of BLAST?

You will notice that there are three regions of the page partially enclosed by an orange border. The section on the top corresponds to the former basic blast page. While the documentation (clicking on the search link) says that accession numbers are OK, it works best with gi numbers where the **:** is replaced by a **|**. There are very few options here to choose from (database, range to use and CD search).

The second region contains the advanced options for a run. Here you can select statistic analysis type, filter to be used, expect value, scoring matrix, gap and extension penalties, and include PSI matrix or PHI pattern if desired.

The third region is for the setting up of the format the of output results. The maximum number of possible hits is now 1000 in this form.

## HELP!!!!!!

Where can I find help for using the NCBI BLAST pages? I want to learn more about the program, the algorithm, and whatever else I can find out.

For this purpose go to the main BLAST page at NCBI. The links given below go to the links in the side panel of the new main BLAST page. An updated version of this help can be found by clicking on the links on the actual submission form. However, the older version of help files may provide greater detail on some aspects of the software than the new ones.

### overview

- a general overview of BLAST

### frequently asked questions

- answers to typical problems that arise using BLAST at NCBI

### tutorial

- a basic set of instructions for a first time BLAST user

### course

- a more in-depth look at understanding BLAST output results

## What is PSI-BLAST?

PSI-BLAST is a method for automatically combining statistically significant alignments produced by BLAST into a position-specific score matrix, and then searching the database using this matrix.

PSI-BLAST is much more sensitive to weak but biologically relevant sequence similarities.

## What is PHI-BLAST?

Protein families often are characterized by conserved sequence patterns or motifs. A researcher frequently wishes to evaluate the significance of a specific pattern within a protein, or to exploit knowledge of known motifs to aid the recognition of greatly diverged but homologous family members. To assist in these efforts, the pattern-hit initiated BLAST (PHI-BLAST) program takes as input both a protein sequence and a pattern of interest that it contains. PHI-BLAST searches a protein database for other instances of the input pattern, and uses those found as seeds for the

construction of local alignments to the query sequence. In many instances, the program is able to detect statistically significant similarity between homologous proteins that are not recognizably related using traditional single-pass database search methods.

## What is the difference between NCBI and WU BLAST?

The initial developers of BLAST have gone their separate ways. One is at NCBI and the other is at Washington University in St. Louis. They both have been working on BLAST at their respective institutions. The folks at Washington University have since copyrighted their version of BLAST and its latest version now requires a license to use. NCBI BLAST is still free.

For a list of sites running WU-BLAST check this url.

## What is the difference between NCBI and SACS BLAST?

The SACS BLAST page uses a slightly older version of the NCBI BLAST programs. The pages are secure, designed only to be used by SACS subscribers.

Minor differences in database naming exist between the two versions. SACS uses the term **nr** for the non-redundant protein database and **nt** for the non-redundant nucleotide database.

## How is BLAST run on socrates?

BLAST can be run on the command line on socrates. 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 and their names.

### BLAST generic command line:

```
blastall -p blastp -d swissprot -i ccho.tfa -v 25 -b 25 -o ccho-sw.blastp
```

## BLAST runs in GCG

GCG has both a version of BLAST that runs remotely at NCBI, **netblast**, and one that runs using the local databases, **blast**.

## 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 cron job will be set up in the user's account to run the process at the described times for the selected time period. 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.

## Customized BLAST searches

When you have special BLAST searching needs, check with SACS to see if scripts have already been developed that could meet your needs with little or no modifications.

Customized scripts to date:

- version 1 - Check for the names of sequence files to be searched with in a given location and convert them to the proper format. Run nr BLAST searches on each of these sequence files producing a html output file.
- version 2 - Check for sequence files and convert to proper format. Strip off primer and vector sequence from the data file using process developed with data owner. Check each sequence to see if it is identical to any sequence that has been previously run in the data set. If new, run nr, est\_human and est\_mouse BLAST searches with the sequence. Produce output in html format (hit links only), full format (html stripped out and showing alignments), plus saving the sequence searched with. Additional information is kept so that gene contig members can be identified.

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

Comparison tables or scoring matrices are a means of relating two characters in a sequence with one another.

Commonly used tables:

- Unitary tables are based on identity. Used for nucleotide sequence analysis.
- PAM (point accepted mutation) tables reflect expected evolutionary change by point mutations in similar sequences. Assumes that all residues are equally mutable. The higher the PAM number the greater the number of allowed mutations between the two residues.
- BLOSUM (block substitution matrices) tables are derived from data representing highly conserved sequence segments in divergent proteins. The number of the table refers to the maximum amount of similarity allowed between sequences used to derive the table.

References:

WU BLAST 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 developed separately at the National Center for Biotechnology Information (NCBI)

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.

## **URL's used in this seminar.**

**NCBI's main BLAST site:**

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

**NCBI's BLAST Programs PAGE (old):**

[http://www.ncbi.nlm.nih.gov/BLAST/blast\\_program.html](http://www.ncbi.nlm.nih.gov/BLAST/blast_program.html)

**NCBI's BLAST Databases PAGE (old):**

[http://www.ncbi.nlm.nih.gov/BLAST/blast\\_databases.html](http://www.ncbi.nlm.nih.gov/BLAST/blast_databases.html)

**NCBI's low complexity section:**

<http://www.ncbi.nlm.nih.gov:80/blast/html/blastcgihelp.html#filter>

**NCBI's expect section:**

<http://www.ncbi.nlm.nih.gov:80/blast/html/blastcgihelp.html#expect>

**NCBI's overview page (old):**

[http://www.ncbi.nlm.nih.gov/blast/blast\\_overview.html](http://www.ncbi.nlm.nih.gov/blast/blast_overview.html)

**NCBI's FAQ page (old):**

[http://www.ncbi.nlm.nih.gov/blast/blast\\_FAQs.html](http://www.ncbi.nlm.nih.gov/blast/blast_FAQs.html)

**NCBI's BLAST tutorial page (old):**

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

**NCBI's course page (old):**

<http://www.ncbi.nlm.nih.gov/blast/tutorial/Altschul-1.html>

**NCBI's PHI-BLAST pattern page (old):**

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

**WU-BLAST info page:**

<http://blast.wustl.edu/>

**SACS local BLAST site:**

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

**SACS Blast manual page:**

<https://www.sacs.ucsf.edu/Training/searching/blast-man.html>

**SACS Blast database list page:**

<https://www.sacs.ucsf.edu/Training/blast/searchable.html>

**SACS AutoBlast page:**

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

**Prior SACS presentation:**

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