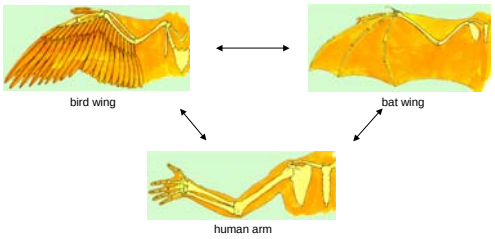
# MCB 371/372

BLAST and
PSI BLAST
3/23/05 and 3/28

*Peter Gogarten*
Office: *BSP 404*
phone: *860 486-4061*,
Email: *gogarten@uconn.edu*

## Homology




by Bob Friedman

# homology

Two sequences are homologous, if there existed an ancestral molecule in the past that is ancestral to both of the sequences

Types of Homology

***Orthologs***: "deepest" bifurcation in molecular tree reflects speciation.
These are the molecules people interested in the taxonomic classification of organisms want to study.

***Paralogs***: "deepest" bifurcation in molecular tree reflects gene duplication. The study of paralogs and their distribution in genomes provides clues on the way genomes evolved. Gen and genome duplication have emerged as the most important pathway to molecular innovation, including the evolution of developmental pathways.

***Xenologs***: gene was obtained by organism through horizontal transfer. The classic example for Xenologs are antibiotic resistance genes, but the history of many other molecules also fits into this category: inteins, selfsplicing introns, transposable elements, ion pumps, other transporters,

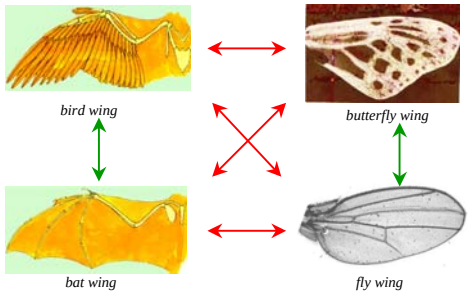
***Synologs***: genes ended up in one organism through fusion of lineages. The paradigm are genes that were transferred into the eukaryotic cell together with the endosymbionts that evolved into mitochondria and plastids
*(the -logs are often spelled with "ue" like in orthologues)*

see Fitch's article in TIG 2000 for more discussion.

# homology *vs* analogy

A priori sequences could be similar due to convergent evolution

Homology (shared ancestry) *versus* Analogy (convergent evolution)




## Two components of similarity searching

Searching method - a model of sequence evolution

Database - search only as good as sequences searching against

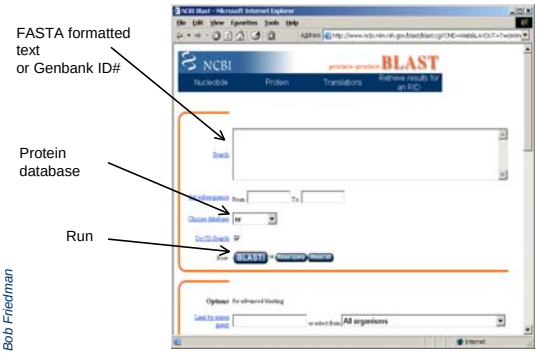
by Bob Friedman

## Types of Blast searching

- blastp compares an amino acid query sequence against a protein sequence database
- blastn compares a nucleotide query sequence against a nucleotide sequence database
- blastx compares the six-frame conceptual protein translation products of a nucleotide query sequence against a protein sequence database
- tblastn compares a protein query sequence against a nucleotide sequence database translated in six reading frames
- tblastx compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

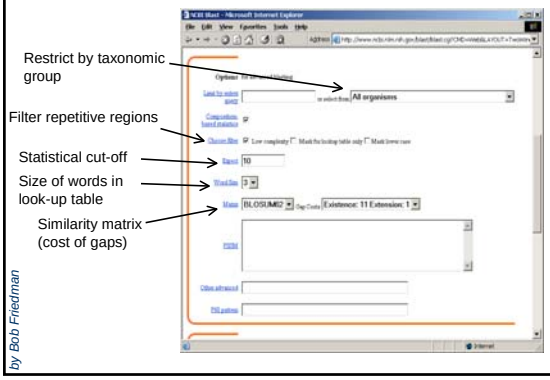
by Bob Friedman

## Routine BlastP search




by Bob Friedman

## BlastP parameters




by Bob Friedman

## Establishing a significant "hit"

Blast's E-value indicates statistical significance of a sequence match

Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. PNAS 87:2264-8

E-value is the Expected number of sequence (HSPs) matches in database of *n* number of sequences

- database size is arbitrary
- multiple testing problem
- E-value calculated from many assumptions
- so, E-value is not easily compared between searches of different databases

Examples:
E-value = 1 = expect the match to occur in the database by chance 1x

E-value = .05 = expect 5% chance of match occurring

E-value = $1x10^{-20}$ = strict match between protein domains

by Bob Friedman

## When are two sequences significantly similar? PRSS

One way to quantify the similarity between two sequences is to

1. compare the actual sequences and calculate an alignment score
2. randomize (scramble) one (or both) of the sequences and calculate the alignment score for the randomized sequences.
3. repeat step 2 at least 100 times
4. describe distribution of randomized alignment scores
5. do a statistical test to determine if the score obtained for the real sequences is significantly better than the score for the randomized sequences

**z-values give the distance between the actual alignment score and the mean of the scores for the randomized sequences expressed as multiples of the standard deviation calculated for the randomized scores.**

**For example: a z-value of 3 means that the actual alignment score is 3 standard deviations better than the average for the randomized sequences. z-values > 3 are usually considered as suggestive of homology, z-values > 5 are considered as sufficient demonstration.**

## PRSS continued

To illustrate the assessment of similarity/homology we will use a program from Pearson's FASTA package called PRSS.
This and many other programs by Bill Pearson are available from his web page at ftp://ftp.virginia.edu/pub/fasta/.

A web version is available here.

Sequences for an in class example are here (fl), here (B), here (A) and here (A2)

BLAST offers a similar service for pairwise sequence comparison bl2seq, however, the statistical evaluation is less straightforward.

To force the bl2seq program to report an alignment increase the E-value.

## E-values and significance

Usually E values larger than 0.0001 are not considered as demonstration of homology.

For small values the E value gives the probability to find a match of this quality in a search of a databank of the same size by chance alone.

**E-values give the expected number of matches with an alignment score this good or better,**
**P-values give the probability of to find a match of this quality or better.**
**P values are [0,1], E-values are [0,infinity).**
**For small values E=P**

**Problem:** If you do 1000 blast searches, you expect one match due to chance with a P-value of 0.0001

**"One should" use a correction for multiple tests, like the Bonferroni correction.**

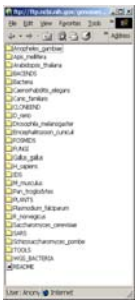
## Blast databases

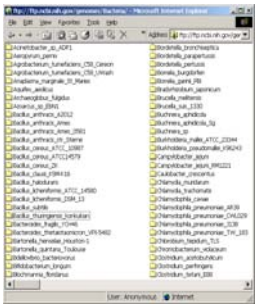
- EST - Expression Sequence Tags; cDNA
- GSS - Genome Survey Sequence; single-pass genomic sequences
- HTGS - unfinished High Throughput Genomic Sequences
- chromosome - complete chromosomes, complete genomes, contigs
- NR - non-redundant DNA or amino acid sequence database
- NT - NR database excluding EST, STS, GSS, HTGS
- PDB - DNA or amino acid sequences accompanied by 3d structures
- STS - Sequence Tagged Sites; short genomic markers for mapping
- Swissprot - well-annotated amino-acid sequences
- TaxDB - taxonomy information
- WGS_xx - whole genome shotgun assemblies
- Also, to obtain organism-specific sequence set:

  ftp://ftp.ncbi.nih.gov/genomes/

by Bob Friedman

## More databases





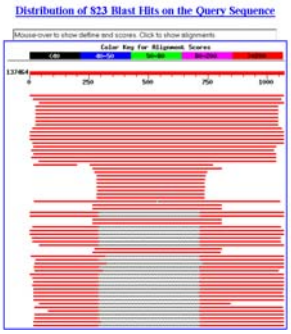
by Bob Friedman

## And more databases


by Bob Friedman

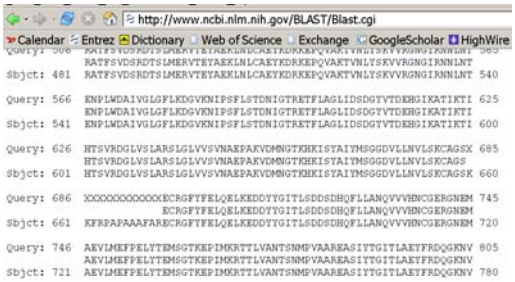
## Example of web based BLAST

**program: BLASTP**
**sequence: vma1 gi:137464**




**BLink provides similar information**

## Effect of low complexity filter




**BUT the most common sequences are simple repeats**

## Custom databases

Custom databases can include private sequence data, non-redundant gene sets based on genomic locations, merging of genetic data from specific organisms

It's also faster to search only the sequence data that is necessary

Can search against sequences with custom names

by Bob Friedman

## Uses of Blast in bioinformatics

The Blast web tool at NCBI is limited:
- custom and multiple databases are not available
- tBlastN (gene prediction) not available
- "time-out" before long searches are completed

What if researcher wants to use tBlastN to find all olfactory receptors in the mosquito?
Answer: Use Blast from command-line

The command-line allows the user to run commands repeatedly

by Bob Friedman

## Formatting a custom database

Format sequence data into Fasta format

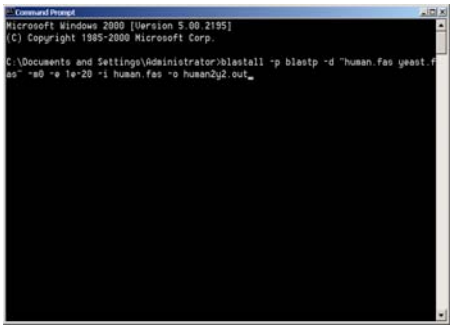
Example of Fasta format:
>sequence 1
AAATGCTTAAAAA
>sequence 2
AAATTGCTAAAAGA

Convert Fasta to Blast format by using FormatDB program from command-line:
formatdb -p F -o T -i name_of_fasta_file

(formatdb.log is a file where the results are logged from the formatting operation)

by Bob Friedman

## BlastP search of custom database




by Bob Friedman

## Psi-Blast: Detecting structural homologs

Psi-Blast was designed to detect homology for highly divergent amino acid sequences

Psi = position-specific iterated

Psi-Blast is a good technique to find "potential candidate" genes

Example: Search for Olfactory Receptor genes in Mosquito genome
Hill CA, Fox AN, Pitts RJ, Kent LB, Tan PL, Chrystal MA, Cravchik A, Collins FH, Robertson HM, Zwiebel LJ (2002) G protein-coupled receptors in Anopheles gambiae. Science 298:176-8

by Bob Friedman

## Psi-Blast Model

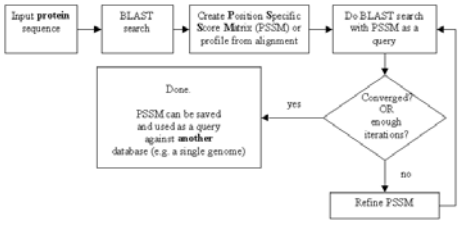
Model of Psi-Blast:
1. Use results of gapped BlastP query to construct a multiple sequence alignment
2. Construct a position-specific scoring matrix from the alignment
3. Search database with alignment instead of query sequence
4. Add matches to alignment and repeat

Similar to Blast, the E-value in Psi-Blast is important in establishing matches
E-value defaults to 0.001 & Blosom62

Psi-Blast can use existing multiple alignment - particularly powerful when the gene functions are known (prior knowledge) or use RPS-Blast database

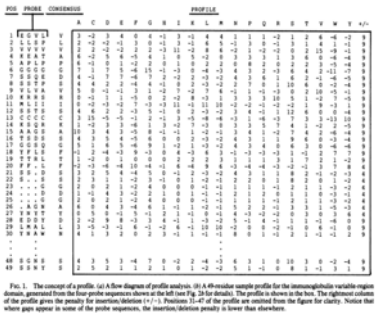
by Bob Friedman

## PSI BLAST scheme




© Olga Zhaxybayeva

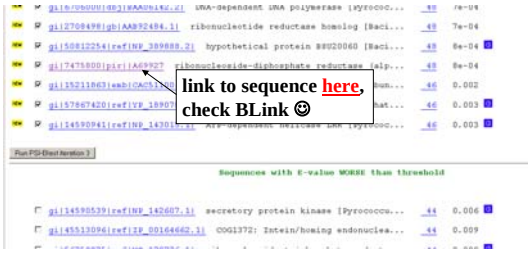
## Position-specific Matrix



M Gribskov, A D McLachlan, and D Eisenberg (1987) Profile analysis: detection of distantly related proteins. PNAS 84:4355-8.

by Bob Friedman

## Psi-Blast Results  Query: 55670331 (intein)




## Blast Summary

Blast is a fast program to find similar DNA or amino acid sequences in a database

NCBI web tool for finding sequence similarity:
http://www.ncbi.nlm.nih.gov/BLAST/

E-value is a statistic to measure the significance of a "match"

Psi-Blast is for finding matches among divergent sequences (position-specific information)
WARNING: For the nth iteration of a PSI BLAST search, the E-value gives the number of matches to the profile NOT to the initial query sequence! The danger is that the profile was corrupted in an earlier iteration.

## Summary of command-line Blast

Repetitive homology searching by use of command-line & scripting language

Another advantage is searching against custom DNA or protein database(s)

Blast results can be processed by text-processing language

by Bob Friedman

Demo using **putty** to bbcxsrv.biotech.uconn.edu
- maybe

**follow instructions of exercise one task 6 – these are the commands**

```
formatdb -i p_abyssi.faa -o T -p T

blastall -i t_maritima.faa -d p_abyssi.faa -o
blast.out -p blastp -e 10 -m 8 -a2

./extract_lines.pl blast.out
```

sftp results
load into spreadsheet
sort data, do histogram …
the extract_lines.pl script is here (you can sftp it into your account, you'll need to chmod 755 extr*.pl afterwards)

# Command Line

**The favored operating system flavor in computational biology is UNIX/LINUX.**
**The command line is similar to DOS.**
**Some of the frequently used commands are here**

| | |
|---|---|
| **pwd** | **ps** |
| **ls** | **ps aux** |
| **ls –l** | **rm** |
| **chmod** | **more** |
| **chmod a+x blastall.sh** | **cat** |
| **chmod 755 *.sh** | **vi (text editor)** |
| **cd** | **ps** |
| **cd ..** | **ps aux** |
| **cd $HOME** | **ssh** |
| **passwd** | **sftp** |

**For windows a good ssh program is putty.**
**UConn also has a site license for the ssh program from ssh.com**