MCB 371/372

Databanks
Databank Searches
3/21/05

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

Other web pages besides the NCBI

- *Nucleic Acid Research Database Issue* Every year, the first issue of *Nucleic Acid Research* is devoted to updates on biological databases.
- http://www.ebi.ac.uk/ The European homolog/analog to NCBI.
- http://rdp.cme.msu.edu/ The US ribosomal databank project
- http://www.psb.ugent.be/rRNA/ The European ribosomal databank project
- http://genome.jgi-psf.org/mic_home.html The Joint Genome Institute
  A recent (well hidden) addition is the integrated microbial genomes site at http://img.jgi.doe.gov/v1.0/main.cgi, the coolest feature is the selected gene neighborhoods.
- http://www.genomesonline.org/ Most up to date information on ongoing and completed genome projects – free for academic users.

Several organism specific resources:
- *http://genome-www.stanford.edu/ Yeast and Arabidopsis genome projects*
- *http://www.flybase.org/ Database of Drosophila Genome*
- *http://www.arabidopsis.org/ TAIR - The Arabidopsis Information Resource*

NR, GenBank, and EMBL

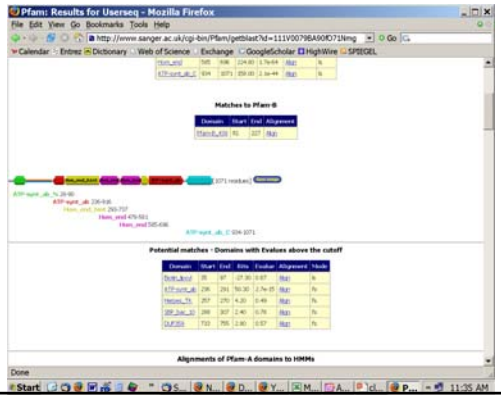
The European, Japanese and US sequence repository have agreed on the information that needs to be contained in a databank submission – the layout of the forms is different, but the content is the same. They share the information that one of the cooperating databanks receives - See Fig. 1.1.

An example of a nucleotide sequence entry (GenBank format – note that the NCBI switched from a flatfile databank to an object oriented one that uses the asn format) is here.

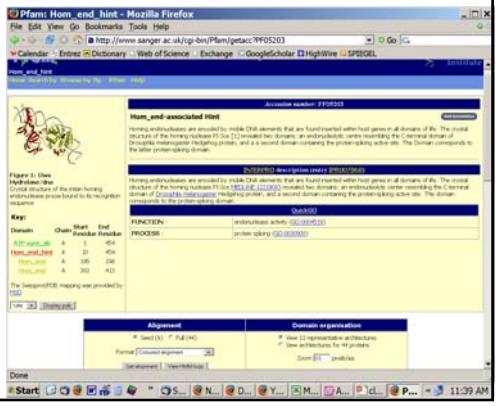
Other frequently used formats are described here.

else: bionet, Intelligenetics, genbank, a trip down memory lane ncbi, PIR, uniprot, genpept, pdb, Structural Classification Of Proteins (SCOP), PFAM and CDD (try with gi 67951)

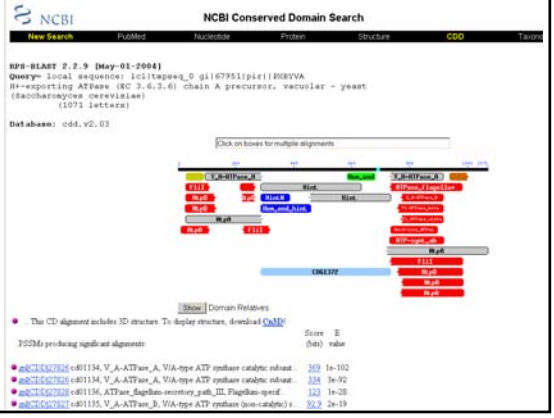
Search of PFAM with vma1



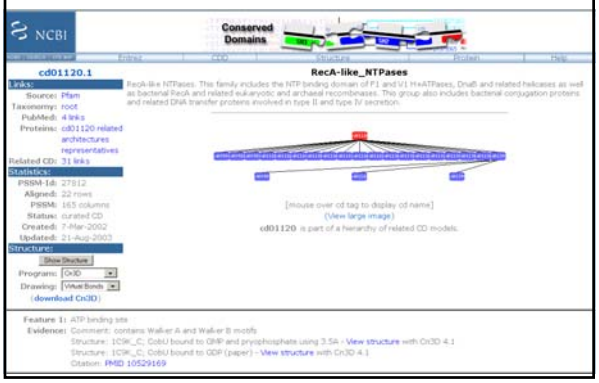
Search of PFAM with vma1, continued



CDD searched with vma1



CDD searched with vma1(cont.)



Links

A computer sciences oriented intro to biological databanks, their design and management is at https://gcrcweba.lacusc.co.la.ca.us/ bifresources/BuildingUsingBiolDB.pdf

Theodosius Dobzhansky

"Nothing in biology makes sense except in the light of evolution"

Related proteins

Present day proteins evolved through substitution and selection from ancestral proteins.
Related proteins have similar sequence AND similar structure AND similar function.

In the above mantra "similar function" can refer to:

- identical function,
- similar function, e.g.:
  - identical reactions catalyzed in different organisms; or
  - same catalytic mechanism but different substrate (malic and lactic acid dehydrogenases);
  - similar subunits and domains that are brought together through a (hypothetical) process called domain shuffling, e.g. nucleotide binding domains in hexokinse, myosin, HSP70, and ATPsynthases.

homology

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

Homology is a "yes" or "no" character (don't know is also possible). Either sequences (or characters share ancestry or they don't (like pregnancy). Molecular biologist often use homology as synonymous with similarity of percent identity. One often reads: sequence A and B are 70% homologous. To an evolutionary biologist this sounds as wrong as 70% pregnant.

Types of Homology
Orthology: bifurcation in molecular tree reflects speciation
Paralogy: bifurcation in molecular tree reflects gene duplication

Sequence Similarity *vs* Homology

The following is based on observation and not on an *a priori* truth:

If two sequences show significant similarity in their primary sequence, they have shared ancestry, and probably similar function.
(although some proteins acquired radically new functional assignments, lysozyme -> lense crystalline).

The Size of Protein Sequence Space

(back of the envelope calculation)

Consider a protein of 600 amino acids.
Assume that for every position there could be any of the twenty possible amino acid.
Then the total number of possibilities is 20 choices for the first position times 20 for the second position times 20 to the third .... = 20 to the 600 = $4*10^{780}$ different proteins possible with lengths of 600 amino acids.

For comparison the universe contains only about 10^{89} protons and has an age of about $5*10^{17}$ seconds or $5*10^{29}$ picoseconds.

If every proton in the universe were a super computer that explored one possible protein sequence per picosecond, we only would have explored $5*10^{118}$ sequences, i.e. a negligible fraction of the possible sequences with length 600 (one in about 10^{662}).

no similarity *vs* no homology

THE REVERSE IS NOT TRUE:

PROTEINS WITH THE **SAME OR SIMILAR FUNCTION DO NOT ALWAYS SHOW SIGNIFICANT SEQUENCE SIMILARITY**
for one of two reasons:

a) they evolved independently
(e.g. different types of nucleotide binding sites);
or
b) they underwent so many substitution events that there is no readily detectable similarity remaining.

Correllar: PROTEINS WITH SHARED ANCESTRY DO NOT ALWAYS SHOW SIGNIFICANT SIMILARITY.

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