

MCB 5472

Student Projects
Databanks, Blast
possibly unix Perl

J. Peter Gogarten

Office: *BPB 404*

phone: *860 486-4061,*

Email: *gogarten@uconn.edu*

Student Projects

- Should be related to your interests !!!
- Examples for possible projects:

Example: Evolution of a gene family

- When in the evolution of the interferon (or whatever you are interested in) gene family did gene duplications occur?
- Which of the resulting subfamilies (if any) have acquired a new function?
- What is the phylogenetic distribution of this subfamily? (Would you expect members of this subfamily to be present in insects, fish, chicken, fungi, archaea?)
- Can you detect episodes of positive selection?
- Is there anything that would suggest gene conversion events?

The “to-do-list” would include:

- gather data (note for some of the questions mentioned above you’ll need aa **and** nucleotide sequences),
- align sequences
- build phylogenies
- analyze sequences
- assess reliability of branches
- INTERPRET WHAT YOU GOT!

Example: Can one detect a distinct second peak in the divergence of putatively chimeric genomes?

Genome fusions are the latest rage in evolutionary biology:

For example:

- Koonin EV, Mushegian AR, Galperin MY, Walker DR. *Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea.* Mol Microbiol. 1997 Aug;25(4):619-37.
- The Eukaryotes are a chimera of at least an archaeal like host cell and a bacterium that evolved into a mitochondrion (+ in some cases a cyanobacterium that evolved into a plastid)
- The Haloarchaea contain many bacterial genes
- The Thermotogales contain many archaeal genes
- Most plants and many fungi (likely including bakers yeast) are aneuployploids

In most of these instances it is not clear that the transfer (duplication) really occurred in a single massive event, or if the transfers (duplications) occurred on a gene by gene basis.

(in yeast the type of genes that were duplicated suggest distinct selection pressures, see Benner et al [here](#))

Example: Chimera? continued

In case of a chimera formed in a single historic event one would expect

A) Two distinct types of phylogenetic affinity.

E.g.: Genes in *Thermotoga maritima* should either group with the sistergroup of the bacterial partner, or with the sistergroup of the archaeal donor

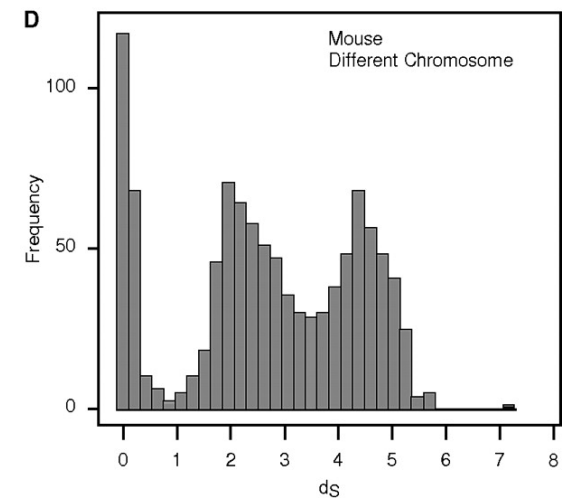
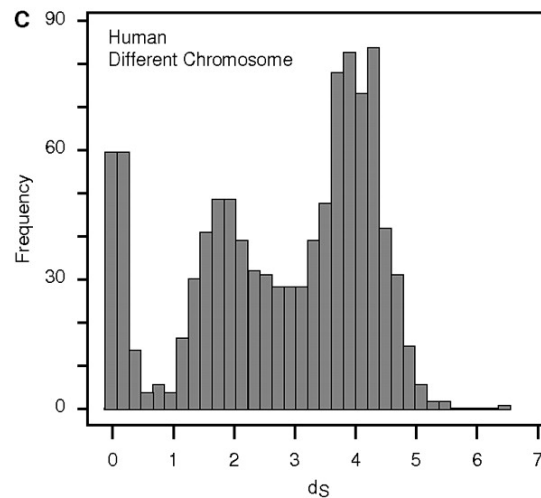
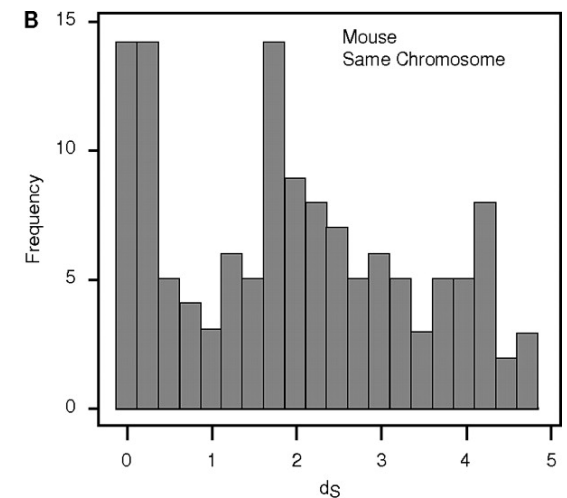
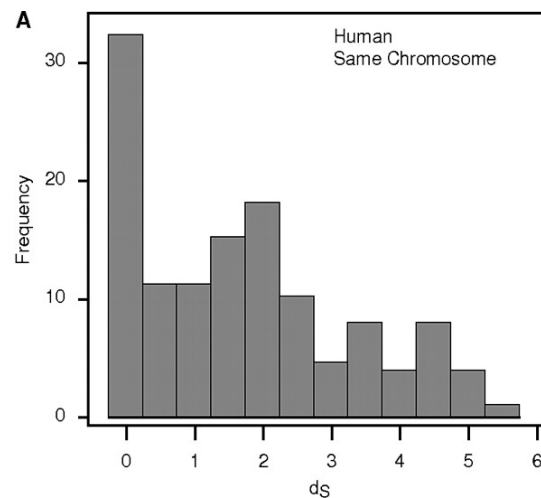
Related: Ancient genome duplication events are revealed by peaks in the divergence of paralogs.

Example: Gene versus Genome Duplications

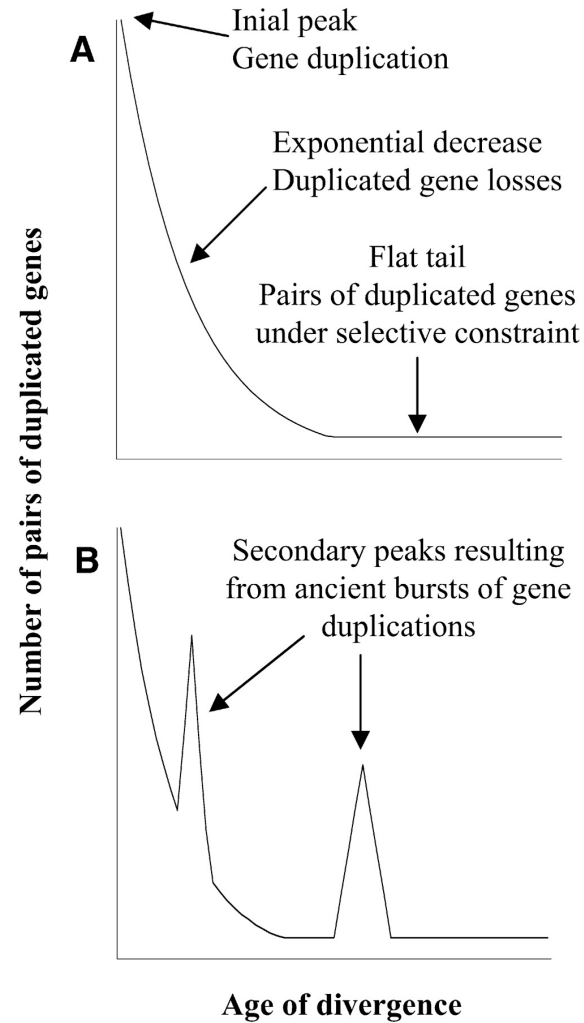
The same approach as suggested for the chimera formation can be applied to the question was the whole genome or a large segment of an organism's genome duplicated, or did the duplications occur in a piecemeal fashion?

Frequency distributions of d_s in human and mouse between the members of two-member gene families located on the same and different chromosomes

From: Robert Friedman and Austin L. Hughes: *Two Patterns of Genome Organization in Mammals: the Chromosomal Distribution of Duplicate Genes in Human and Mouse*. *Mol. Biol. Evol.* 21(6):1008–1013. 2004



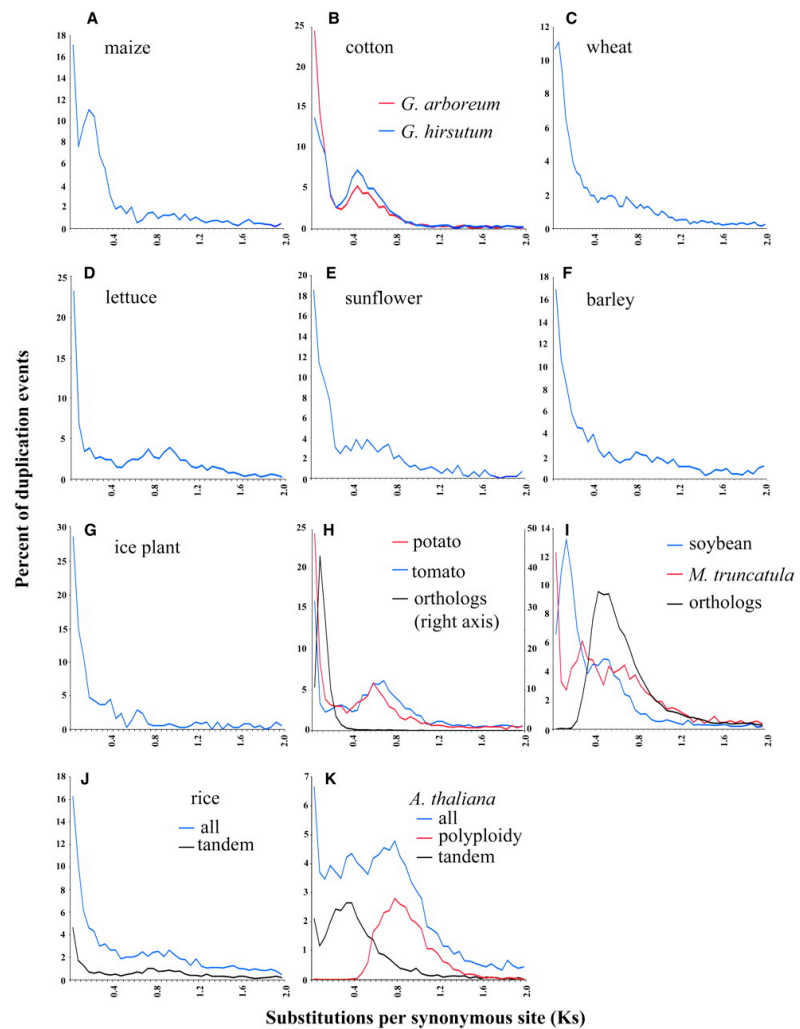
Theoretical Age Distributions of Pairs of Duplicated Genes in a Genome



Blanc, G., et al. *Plant Cell* 2004;16:1667-1678



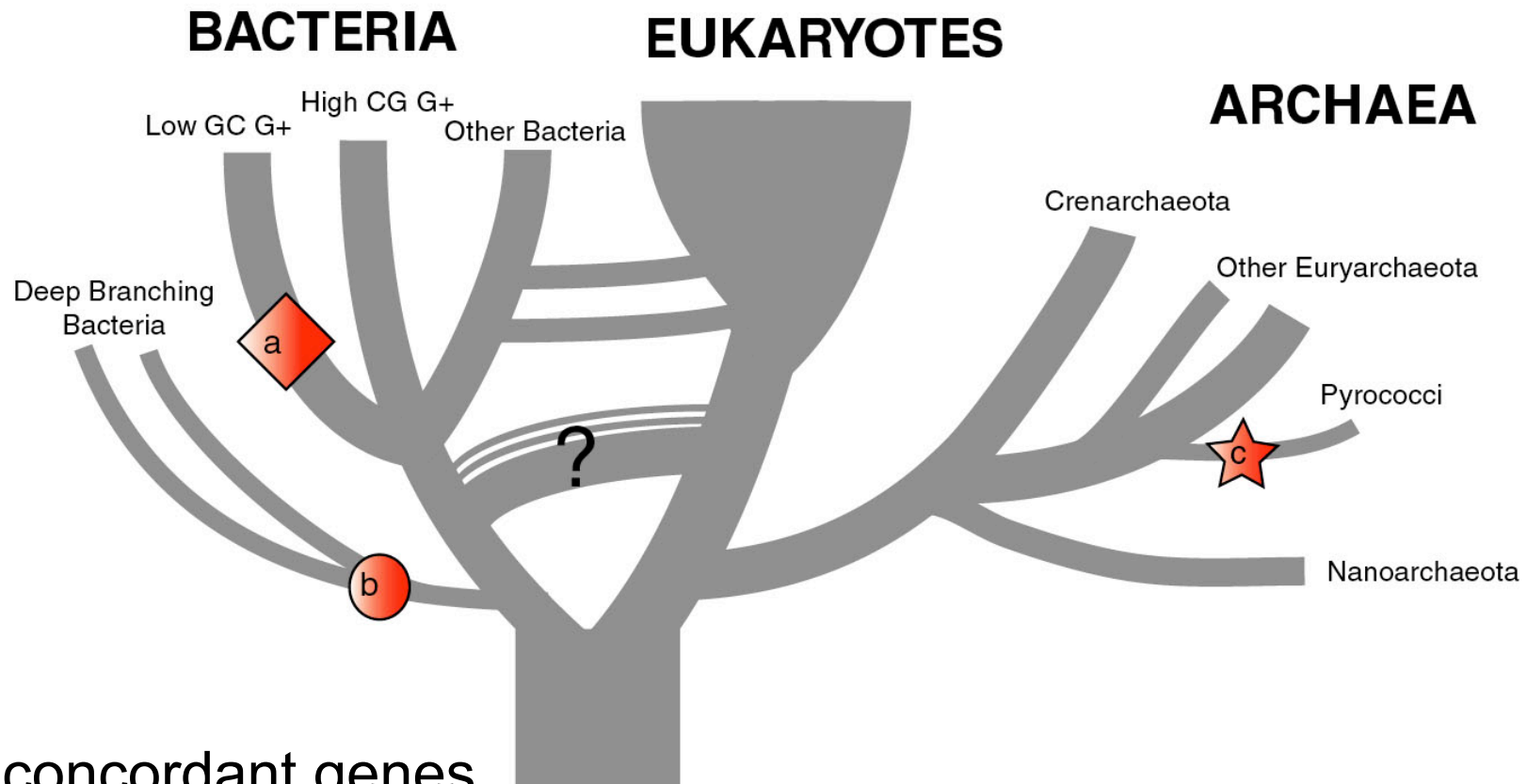
Distributions of the Fraction of Duplication Events as a Function of Their Levels of Synonymous Substitution for 14 Model Plant Species



Blanc, G., et al. Plant Cell 2004;16:1667-1678



The Phylogenetic position of *Thermotoga maritima*



(a) concordant genes,

(b) according to 16S (and other conserved genes)

(c) according to phylogenetically discordant genes

Gophna, Doolittle & Charlebois: Weighted genome trees: refinements and applications. *J. Bacteriol.* [here](#)

Gogarten & Townsend: Horizontal gene transfer, genome innovation, and evolution
Nature Reviews in Microbiology 3(9) 679-687 ([pdf](#))

Chimera Example, continued

The “to-do-list” would include:

- Formulate the question you want to address
- Download and analyze the required genomes
- Run blastall (this might take a couple of hours)
- Analyze the results in an Excel spreadsheet
- Selected some genes (e.g., the ones that are most archaeal), assemble gene families and reconstruct their phylogenies.
- **INTERPRET YOUR RESULTS!** What does it all mean?

Background for group selection Example:

Selection acts on

- **genes** (as in the selfish gene theory, the genes are the replicators that build the body of the organism). According to this all genes are selfish, most are cooperating with one another, a few are not. To distinguish the latter from the former, I call them parasitic genes (or molecular parasites).
- **individuals** in a population (the survival of the fittest).
- **groups** of organisms (group selection). The group that has properties that allows it to adapt better, or to evolve faster, or to make better use of resources will be selected. In this case the group (community, *not necessarily all belonging to the same species*) is the unit of selection. (see group selection entry at [wikipedia](#))

Note: in general this is controversial. To what extent is group selection reflecting kin-selection: the organism acting to guarantee survival of genes that are related to its own genes (bees in a beehive are all closely related).

Examples for “group selection” in microbes: (a) *Agrobacteria*

Agrobacteria that carry a Ti plasmid can transform plant cells with a T DNA. As result of a successful transformation the plant cell has integrated the T DNA into its genome and expresses the encoded genes. This results in the transformed cells forming a tumor, and, in addition, the transformed plant cells also produce a strange amino acid that cannot be utilized by the plant cells, but that serves as a carbon and nitrogen source for the *Agrobacteria*. The genes responsible for transferring the Ti plasmid between different *Agrobacteria* (*tra* genes) are under the control of quorum sensing. The effect is that if one *Agrobacterium* strain has successfully transformed a plant, and now lives from the plant produced strange amino acid, other *Agrobacteria* can receive the Ti plasmid, which contains the T DNA transferred into the plant and in addition encodes enzymes that allow the metabolism of the strange amino acids. The *Agrobacteria*, which receive the Ti-plasmid thus participate in the utilization of the plant produced carbon and nitrogen source. This observation **might be described as group selection**: the population of *Agrobacteria* avoids a selective sweep and carries larger genetic diversity into the population living on the transformed plant. The increased diversity will facilitate future adaptations to a changing environment, and will avoid the fixation of slightly deleterious mutations that might have been carried by the *Agrobacterium* that transformed the plant cell. On the other hand, **one can consider this process the outcome of the "selfishness" of the *tra*-genes and of the Ti plasmid**. These genes manage to move themselves into the growing part of the population, and they will benefit from a more diverse group of host organisms.

Examples for “group selection” in microbes (b):
*Metal resistance genes in microbial communities
inside rocks in the dry valleys of Antarctica*

These rocks have high concentrations of toxic heavy metals. The endolithic microbial community readily shares heavy metal resistant genes with microbes that might be able to become part of the community. At the community level the outcome is a higher diversity, and a richer network of metabolic reactions. Presumably the more diverse communities are more stable towards perturbations, and provided the community can propagate as a whole, this would provide a **selective advantage to the community**. However from the **selfish gene point of view**, the resistance gene increases its chances of long term survival by invading as many additional species as possible.

Examples for “group selection” in microbes (c): Gene Transfer Agents (GTA) in alpha proteobacteria

GTA are prophages that do not specifically pack their own DNA, but that unselectively pack host DNA into the phage head (see [here](#)).

- Are these just defective prophages that lost their sequence specificity in DNA packaging?
- Is this an illustration that HGT is beneficial and under group selection?

(Aside: In general, HGT might reflect uptake of DNA for food, recombination might be a negligible side effect (Rosi Redfield, e.g. [here](#)), or HGT might reflect the selfishness of the transferred DNA.

Testing GTAs as agents selected by group selection.

Possible hypotheses:

- GTAs are defective prophages that lost their sequence specificity in DNA packaging?
- GTAs evolved from phages but now benefit the group and are under group selection?

Under #2:

- The GTA should be more related to one another than to functioning phage
- Their molecular phylogeny should reflect the phylogeny of the organism (as measured by rRNA and ribosomal proteins)
- The genes encoding the GTA should be under strong purifying selection (under #1 they should be pseudogenes).

GTAs: to do list

- ❖ Identify GTAs in genomes of closely related organisms.
- ❖ Align the major conserved genes from these GTAs.
- ❖ Include an appropriate outgroup
From the same genome select genes from the translation machinery, whose phylogeny likely reflects the main current of the organismal history.
- ❖ Calculate and compare the phylogenies.
- ❖ Test the GTA genes for positive/purifying selection

other ideas:

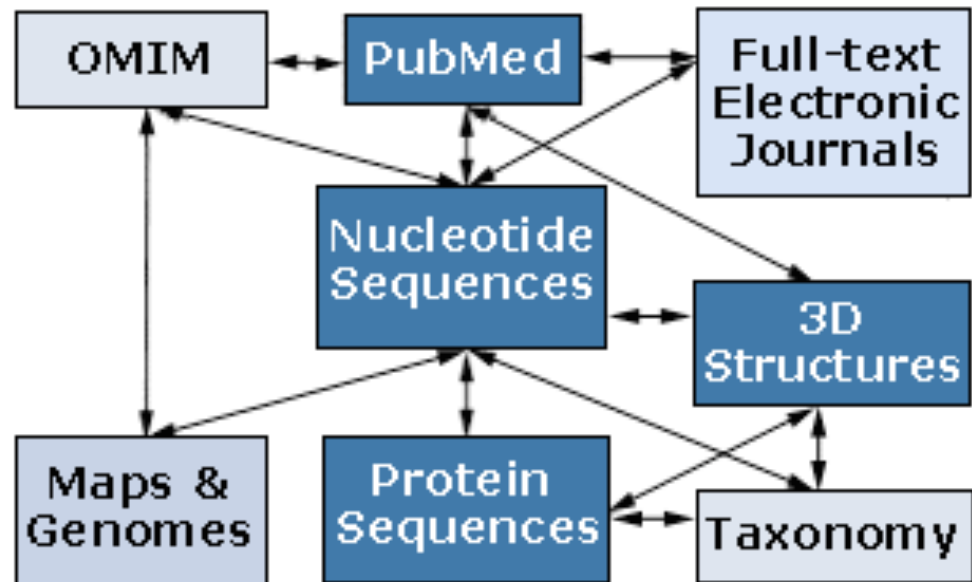
- Write a script that uses the 100+ known intein alleles each as a seed in PSI BLAST, and stores the profiles. Write a second script that uses these profiles to detect putative inteins in completely sequenced genomes.
- Same as above but use transposases, integrases, homing endonucleases, or a molecular parasite of your choice as a seed.
- Determine the impact of HGT on reconstruction of organismal evolution. Use one of the several available programs to simulate sequence evolution for several genes along a tree. Reconstruct the phylogeny using either the concatenated genes, or the individual data sets (in the latter case use a super tree approach to calculate the organismal tree as consensus.
Which approach (supertree versus concatenation) recovers the correct tree?
Use different approaches to identify the transferred genes.
- Search the different versions of the Mosquito genome for genes from *Aeromonas*.
- Form families for all genes from Thermotogales, add the fifteen most similar sequences from reference genomes, calculate phylogenies, screen for polyphyly of Thermotogales, screen for conflict with consensus.

Databanks (A)



NCBI (National Center for Biotechnology Information) is a home for many public biological databases (see an older diagram below). All of the databases are **interlinked**, and they all have common search and retrieval system - **Entrez**.

Another more complete representation with an interactive display of the number of the connections between the different databases in ENTRZ is [here](#).



Entrez / Pubmed, continued

- An interactive Pubmed tutorial click [here](#).
- An Entrez tutorial (non interactive) is [here](#)
- Use Boolean operators (**AND**, **OR**, **NOT**) to perform advanced searches.
[Here](#) is an explanation of the Boolean operators from the Library of Congress Help Page.
- Explore features of [Entrez](#) interface:
Limits, Index, History, and Clipboard.
- Search Field Tags- Listed [here](#).

Other Literature databanks and Services

While Pubmed is incorporating more and more non-medical literature, there might still be gaps in the coverage.

Alternatives are local services offered at the UConn libraries. Especially Current Contents and Agricola nicely complement PubMed. The best way to access them is the use of "SilverPlatter" database.

Also, the "Web of Science" database gives access to the Science Citation Index: a database that tracks cited references in journals. Note that these resources are restricted to UConn domain, so you either need to access it from a campus computer or through the proxy account.

Search Robots



[PubCrawler](#) allows to run predefined literature searches. Results are written into a database and you are send an email, if there were new results. NCBI now offers a similar service (see My NCBI (Chubby), check the tutorial).



[Swiss-Shop](#) is offering the same service for proteins

Sequence and structure databanks

can be divided into many different categories.

One of the most important is

Supervised databanks with gatekeeper. Examples:

Swissprot

Refseq (at NCBI)

Entries are checked for accuracy.

+ more reliable annotations

-- frequently out of date

Repositories without gatekeeper. Examples:

GenBank

EMBL

TrEMBL

Everything is accepted

+ everything is available

-- many duplicates

-- poor reliability of annotations

10 minute Break?

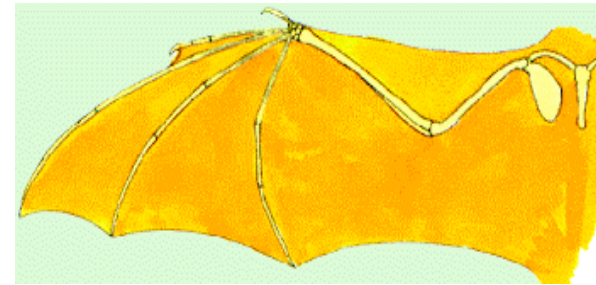
Theodosius Dobzhansky:

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

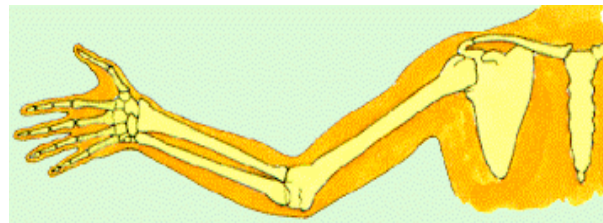
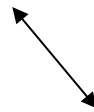
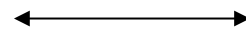
Homology



bird wing



bat wing



human arm

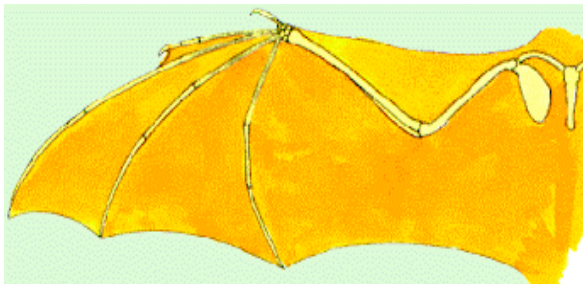
homology vs analogy

A priori sequences could be similar due to convergent evolution

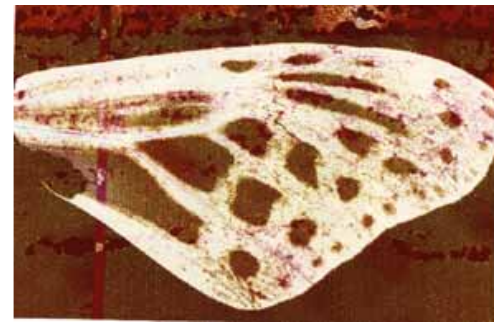
Homology (shared ancestry) *versus* **Analogy** (convergent evolution)



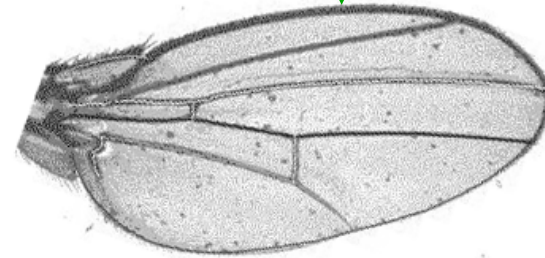
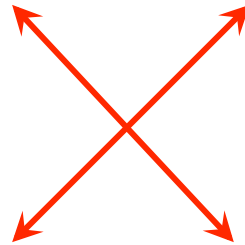
bird wing



bat wing



butterfly wing



fly wing



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 hexokinase, 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 biologists 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 (complex) 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 = $4 \cdot 10^{780}$ 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 $5 \cdot 10^{17}$ seconds or $5 \cdot 10^{29}$ picoseconds.

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

no similarity vs no homology

If two (complex) sequences show significant similarity in their primary sequence, they have shared ancestry, and probably similar function.

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.

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

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)

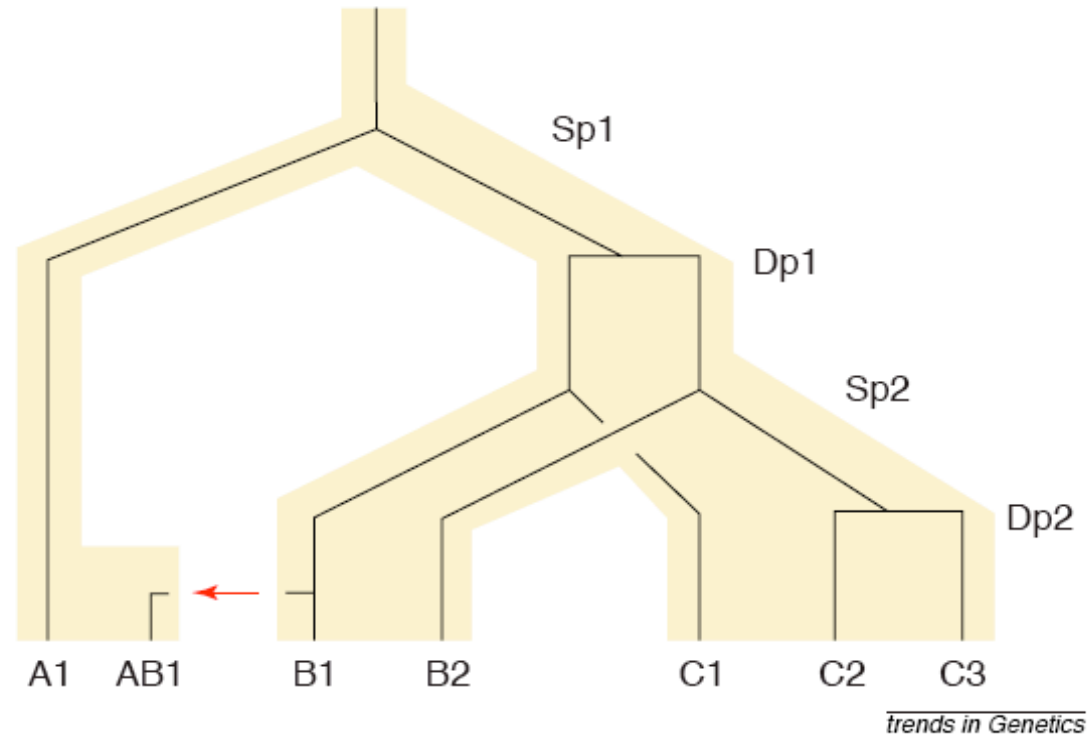
see Fitch's article in [TIG 2000](#) for more discussion.

Homologs, orthologs, and paralogs

- **Homologous** structures or characters evolved from the same ancestral structure or character that *existed in some organism in the past*.
- **Orthologous** characters present in two organism (A and B) are homologs that are derived from a structure *that existed in the most recent common ancestor* (MRCAs) of A and B (orthologs often have the same function, but this is NOT part of the definition; e.g. human arms, wings or birds and bats).
- **Paralogous** characters in the same or in two different organisms are homologs that are not derived from the same character in the MRCA, rather they are *related* (at their deepest node) *by a gene duplication event*.

Examples

FIGURE 1. Orthology, paralogy and xenology



B1 is an ortholog to C1 and to A1

C2 is a paralog to C3 and to B1;

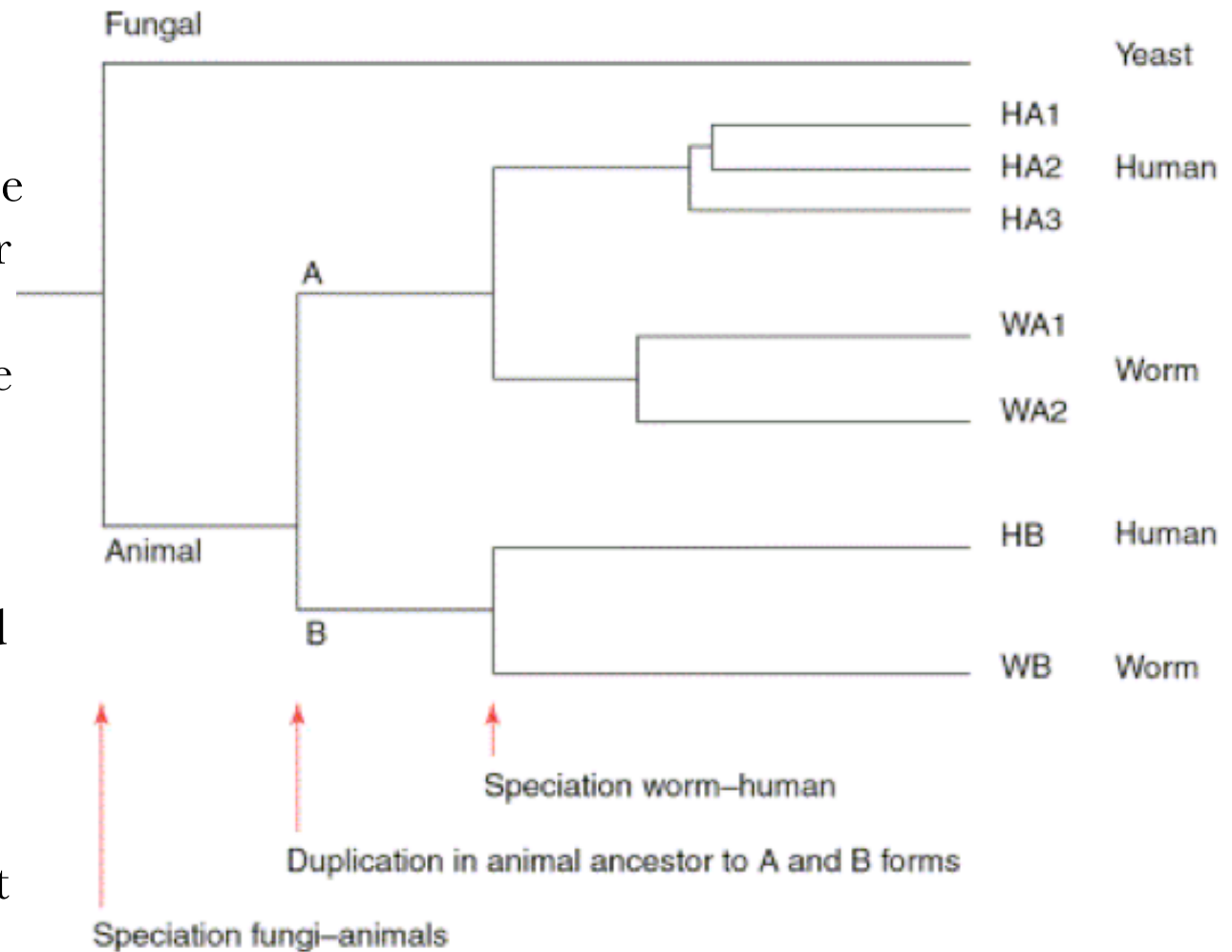
BUT

A1 is an ortholog to both B1, B2, and to C1, C2, and C3

From: Walter Fitch (2000): *Homology: a personal view on some of the problems*, TIG 16 (5) 227-231

Types of Paralogs: In- and Outparalogs

.... all genes in the HA* set are co-orthologous to all genes in the WA* set. The genes HA* are hence 'inparalogs' to each other when comparing human to worm. By contrast, the genes HB and HA* are 'outparalogs' when comparing human with worm. However, HB and HA*, and WB and WA* are inparalogs when comparing with yeast, because the animal–yeast split pre-dates the HA*–HB duplication.



From: Sonnhammer and Koonin: Orthology, paralogy and proposed classification for paralog TIG 18 (12) 2002, 619-620

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? Or, if you want to check the presence of a (pseudo)gene in a preliminary genome assembly?

Answer: Use Blast from command-line

Also: The command-line allows the user to run commands repeatedly

Types of Blast searching

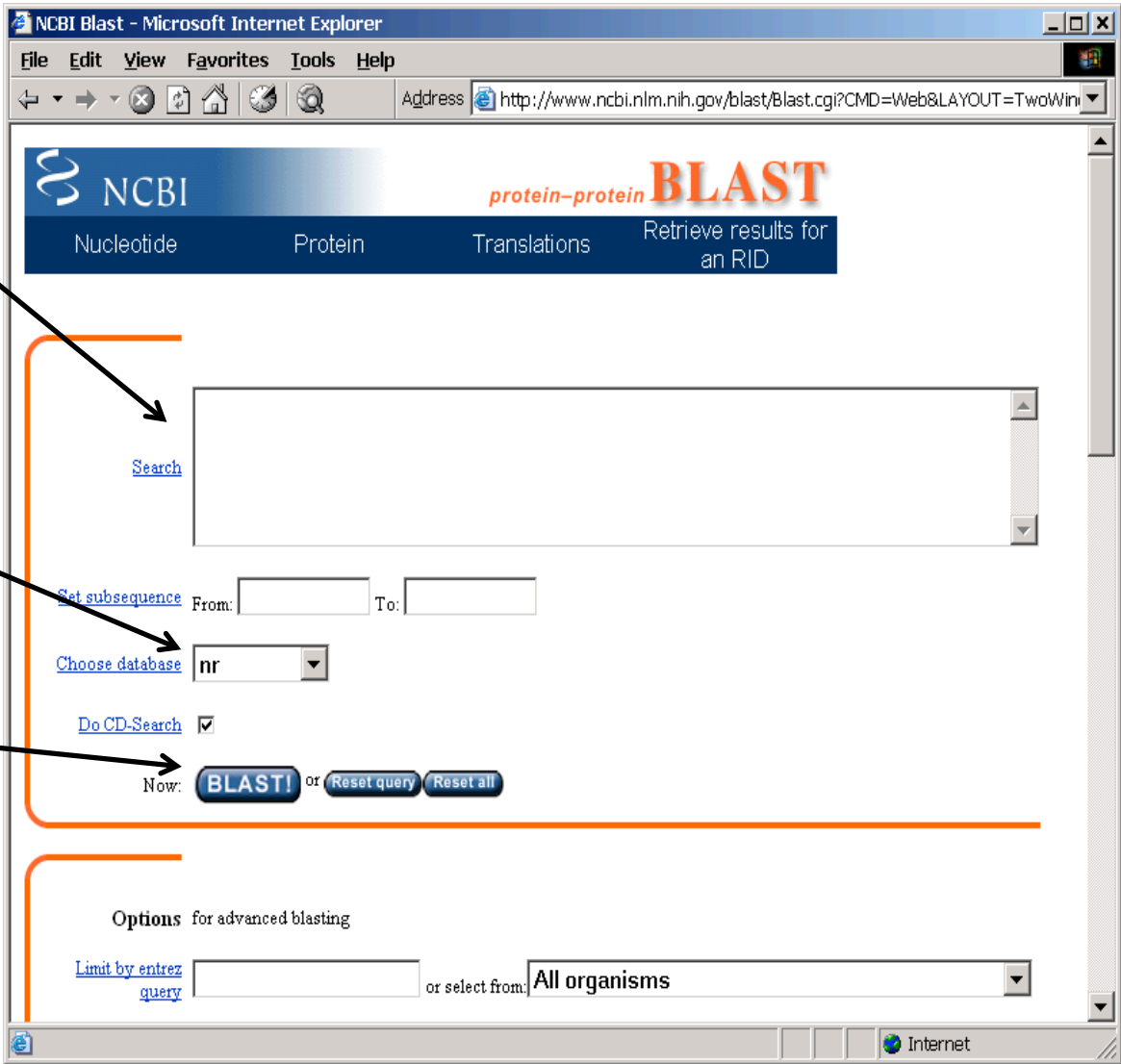
- `blastp` compares an amino acid query sequence against a protein sequence database
- `blastn` compares a nucleotide query sequence against a nucleotide sequence database
- `blastx` compares the six-frame conceptual protein translation products of a nucleotide query sequence against a protein sequence 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.

Routine BlastP search

FASTA formatted text
or Genbank ID#

Protein database

Run



BlastP parameters

Restrict by taxonomic group

Filter repetitive regions

Statistical cut-off

Size of words in look-up table

Similarity matrix (cost of gaps)

Options for advanced blasting

[Limit by entrez query](#) or select from: **All organisms**

[Composition-based statistics](#)

[Choose filter](#) Low complexity Mask for lookup table only Mask lower case

[Expect](#)

[Word Size](#)

[Matrix](#) **BLOSUM62** Gap Costs **Existence: 11 Extension: 1**

[PSSM](#)

[Other advanced](#)

[PHI pattern](#)

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 = 1×10^{-20} = strict match between protein domains

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.

A [web version](#) is available [here](#).

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

Results are [here](#)

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 $[0,\text{infinity})$.

For small values $E=P$

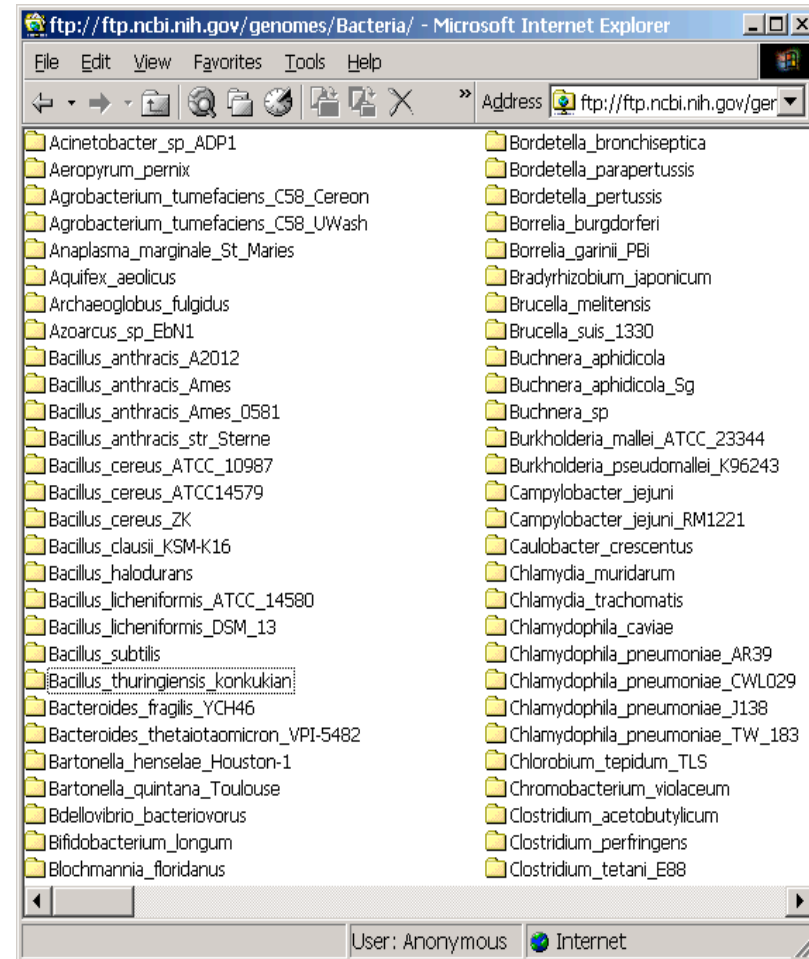
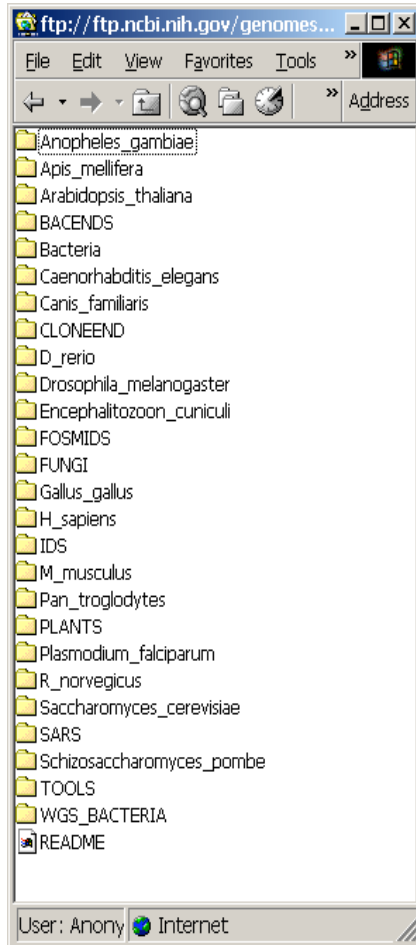
Problem: If you do 1000 blast searches, you expect one match due to chance with a P-value of 0.0001

“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/`

More databases



And more databases

BLAST the Environmental Samples data

The data sources:

Sargasso Sea Environmental Samples: The Institute for Biological Energy Alternatives.

Venter, J.C., et al., *Environmental Genome Shotgun Sequencing of the Sargasso Sea*, [Science](#), 2004 Apr 2,304(5667):66-74.

Mine Drainage Environmental Samples: DOE Joint Genome Institute.

Tyson, G.W., et al., *Community structure and metabolism through reconstruction of microbial genomes from the environment*, [Nature](#), 2004 Mar 4;428(6978):37-43.

Database:

Nucleotide:
Sargasso Sea
Mine Drainage
Both Sample Sets
Proteins:
Sargasso Sea

Program: MegaBlast

Enter an accession, GI, or a sequence in FASTA format:

Optional parameters:

[Expect](#) [Filter](#) [Descriptions](#) [Alignments](#)

[Advanced options:](#)

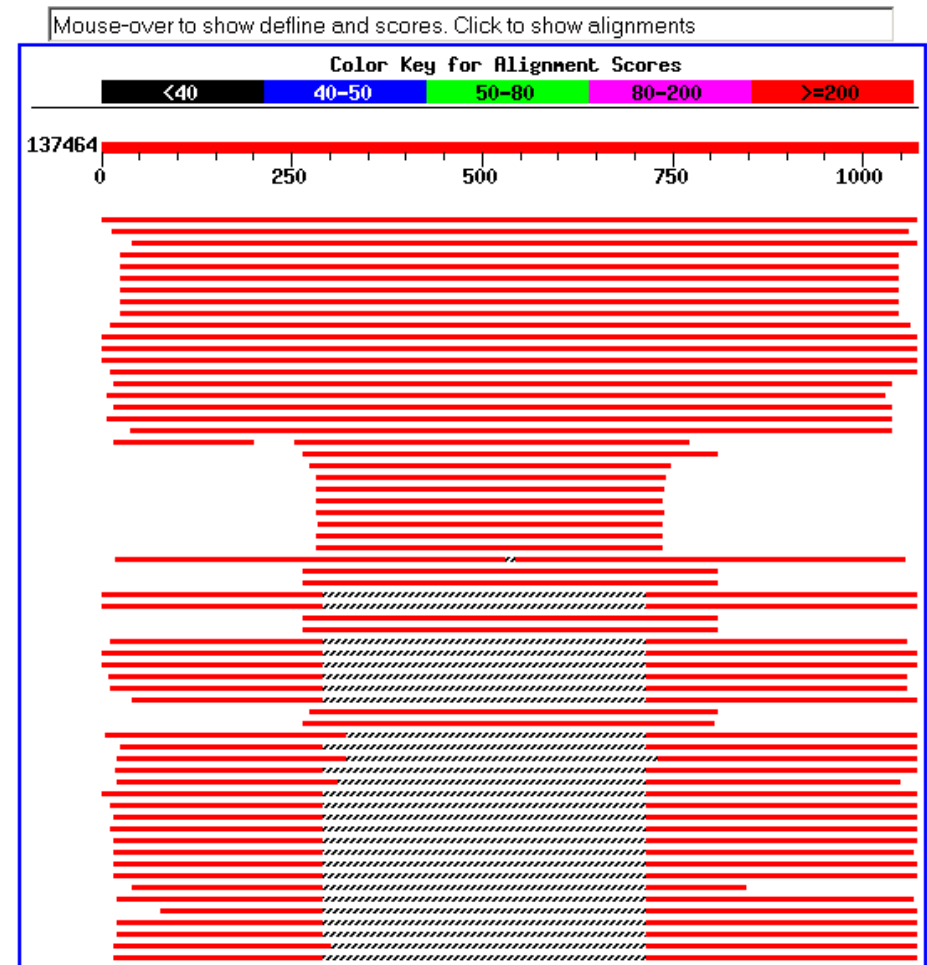
Example of web based BLAST

program: **BLASTP**

sequence: vma1 gi:
137464

BLink provides similar
information

Distribution of 823 Blast Hits on the Query Sequence



Effect of low complexity filter

```
Query: 506  RATFSVDSRDTSLMERVTEYAERKLNLCAEYKDRKEPQVAKT VNLVYSKVVRGNGIRNNLNT 565
          RATFSVDSRDTSLMERVTEYAERKLNLCAEYKDRKEPQVAKT VNLVYSKVVRGNGIRNNLNT
Sbjct: 481  RATFSVDSRDTSLMERVTEYAERKLNLCAEYKDRKEPQVAKT VNLVYSKVVRGNGIRNNLNT 540

Query: 566  ENPLWDAIVGLGFLKDGVKNIPSFLSTDNIGTRETFLAGLIDSDGYVTDEHGKATIKTI 625
          ENPLWDAIVGLGFLKDGVKNIPSFLSTDNIGTRETFLAGLIDSDGYVTDEHGKATIKTI
Sbjct: 541  ENPLWDAIVGLGFLKDGVKNIPSFLSTDNIGTRETFLAGLIDSDGYVTDEHGKATIKTI 600

Query: 626  HTSVRDGLVSLARSLGLVSVNAEPAKVDMMNGTKHKISYAIYMSGGDVLLNVLSKCAGSX 685
          HTSVRDGLVSLARSLGLVSVNAEPAKVDMMNGTKHKISYAIYMSGGDVLLNVLSKCAGS
Sbjct: 601  HTSVRDGLVSLARSLGLVSVNAEPAKVDMMNGTKHKISYAIYMSGGDVLLNVLSKCAGSK 660

Query: 686  XXXXXXXXXXXXXEGRGFYFELQELKEDDYYGITLSDDSDHQFLLANQVVVHNCGERGNEM 745
          ECRGFYFELQELKEDDYYGITLSDDSDHQFLLANQVVVHNCGERGNEM
Sbjct: 661  KFRPAPAAAFARECRGFYFELQELKEDDYYGITLSDDSDHQFLLANQVVVHNCGERGNEM 720

Query: 746  AEVLMEFPPELYTEMSGTKPEIMKRTTLVANTSNNMPVAAREASITYGITLAEYFRDQGKNV 805
          AEVLMEFPPELYTEMSGTKPEIMKRTTLVANTSNNMPVAAREASITYGITLAEYFRDQGKNV
Sbjct: 721  AEVLMEFPPELYTEMSGTKPEIMKRTTLVANTSNNMPVAAREASITYGITLAEYFRDQGKNV 780
```

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

Formatting a custom database

Format sequence data into Fasta format

Example of Fasta format:

```
>sequence 1
```

```
AAATGCTTAAAAA
```

```
>sequence 2
```

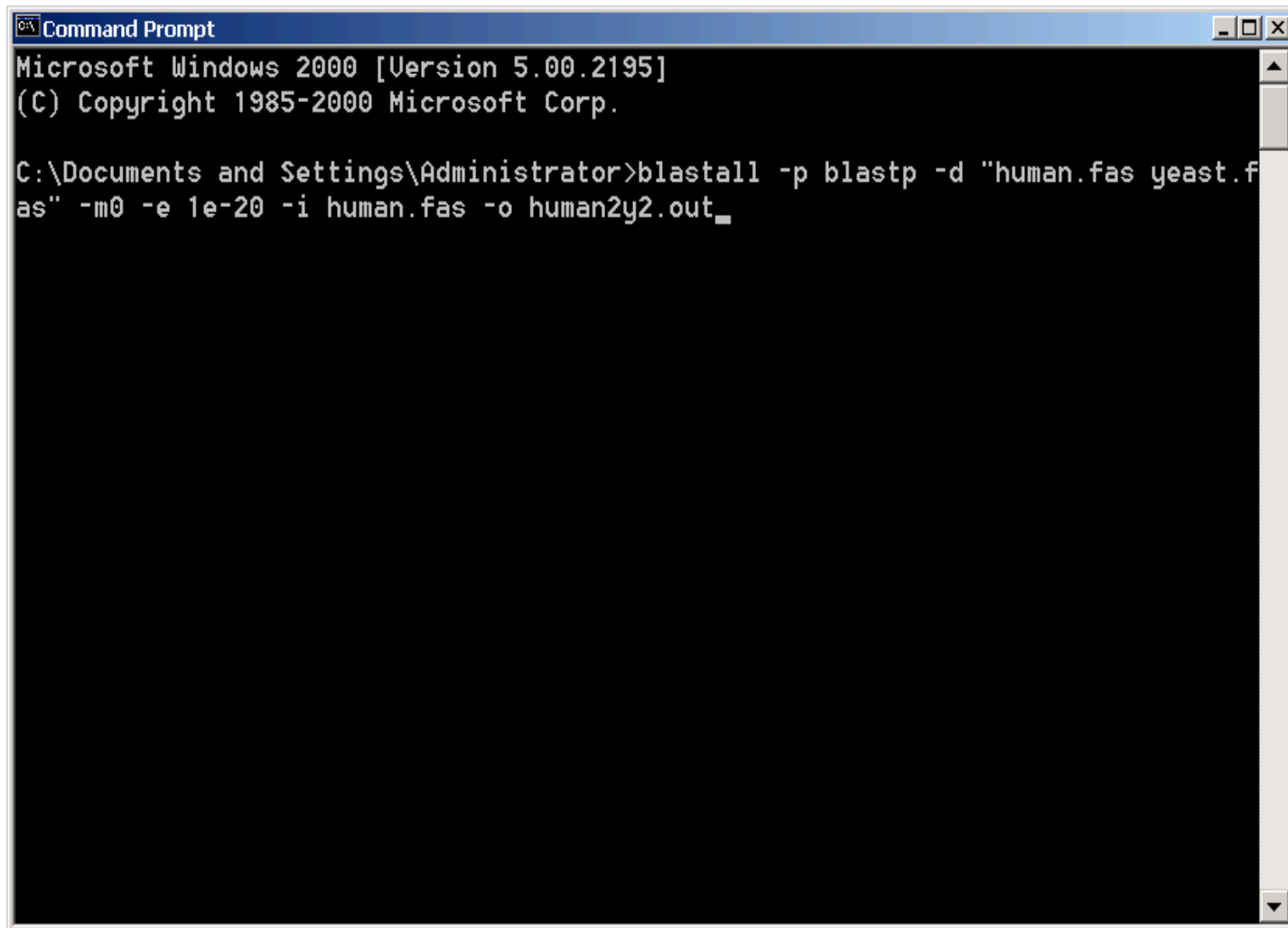
```
AAATTGCTAAAAGA
```

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)

BlastP search of custom database



```
Command Prompt
Microsoft Windows 2000 [Version 5.00.2195]
(C) Copyright 1985-2000 Microsoft Corp.

C:\Documents and Settings\Administrator>blastall -p blastp -d "human.fas yeast.f
as" -m0 -e 1e-20 -i human.fas -o human2y2.out_
```

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.

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

Search of PFAM with vma1

Pfam: Results for Userseq - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.sanger.ac.uk/cgi-bin/Pfam/getblast?id=111V0079BA90f071Nmg

Calendar Entrez Dictionary Web of Science Exchange GoogleScholar HighWire SPIEGEL

Hom_end	585	696	224.80	1.7e-64	Align	ls
ATP-synt_ab_C	934	1071	158.00	2.1e-44	Align	ls

Matches to Pfam-B

Domain	Start	End	Alignment
Pfam-B_439	92	227	Align

ATP-synt_ab_N 28-90
 ATP-synt_ab 236-916
 Hom_end_hint 293-737
 Hom_end 478-581
 Hom_end 585-696
 ATP-synt_ab_C 934-1071

Potential matches - Domains with Evalues above the cutoff

Domain	Start	End	Bits	Evalue	Alignment	Mode
Biotin_lipoyl	35	97	-27.30	0.67	Align	ls
ATP-synt_ab	236	291	50.30	2.7e-15	Align	fs
Herpes_TK	257	270	4.20	0.49	Align	fs
SBP_bac_10	288	307	2.40	0.78	Align	fs
DUF359	733	755	2.80	0.57	Align	fs

Alignments of Pfam-A domains to HMMs

Done

Start S... N... D... Y... M... A... cl... P... 11:35 AM

Search of PFAM with vma1, continued

Pfam: Hom_end_hint - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF05203

Calendar Entrez Dictionary Web of Science Exchange GoogleScholar HighWire SPIEGEL

Home Search by Browse by ftp iPfam Help




Figure 1: 1lws
Hydrolase/dna
Crystal structure of the intein homing endonuclease pi-scei bound to its recognition sequence

Key:

Domain	Chain	Start Residue	End Residue
ATP-synt_ab	A	1	454
Hom_end_hint	A	10	454
Hom_end	A	195	298
Hom_end	A	302	413

The Swissprot/PDB mapping was provided by [MSD](#)

1dfa

Accession number: PF05203

Hom_end-associated Hint [Add Annotation](#)

Homing endonucleases are encoded by mobile DNA elements that are found inserted within host genes in all domains of life. The crystal structure of the homing nuclease P1-Sce [1] revealed two domains: an endonucleolytic centre resembling the C-terminal domain of *Drosophila melanogaster* Hedgehog protein, and a second domain containing the protein-splicing active site. This Domain corresponds to the latter protein-splicing domain.

INTERPRO description (entry IPR007868)

Homing endonucleases are encoded by mobile DNA elements that are found inserted within host genes in all domains of life. The crystal structure of the homing nuclease P1-Sce [MEDLINE:12219083](#) revealed two domains: an endonucleolytic center resembling the C-terminal domain of *Drosophila melanogaster* Hedgehog protein, and a second domain containing the protein-splicing active site. This domain corresponds to the protein-splicing domain.

QuickGO

FUNCTION :	endonuclease activity (GO:0004519)
PROCESS :	protein splicing (GO:0030908)

Alignment

Seed (9) Full (44)

Format:

Domain organisation

View 11 representative architectures

View architectures for 44 proteins

Zoom pixels/aa.

Done

Start [Taskbar icons] 11:39 AM

CDD searched with vma1



NCBI Conserved Domain Search

[New Search](#)
[PubMed](#)
[Nucleotide](#)
[Protein](#)
[Structure](#)
[CDD](#)
[Taxonomy](#)

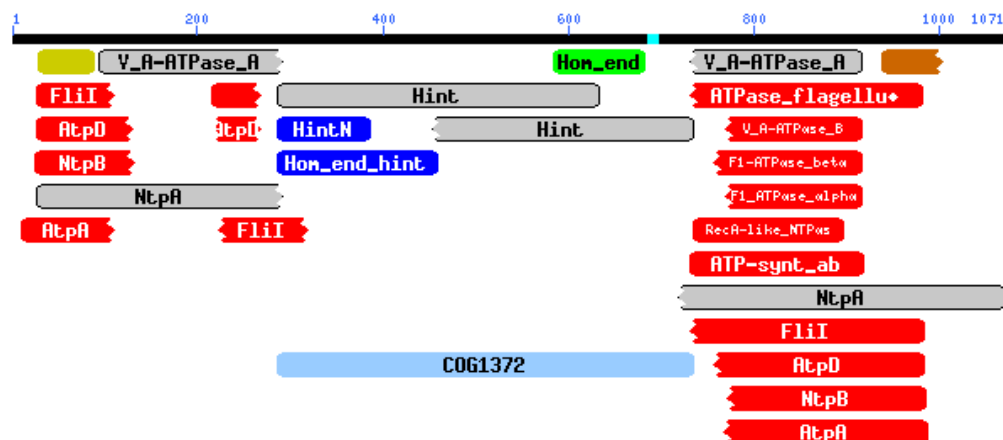
RPS-BLAST 2.2.9 [May-01-2004]

Query= local sequence: lcl|tmpseq_0 gi|67951|pir||PXBYVA

H+-exporting ATPase (EC 3.6.3.6) chain A precursor, vacuolar - yeast
(*Saccharomyces cerevisiae*)
(1071 letters)

Database: cdd.v2.03

Click on boxes for multiple alignments



Show Domain Relatives

... This CD alignment includes 3D structure. To display structure, download [Cn3D!](#)

PSSMs producing significant alignments:	Score	E value
gnl CDD 27826 cd01134, V_A-ATPase_A, V/A-type ATP synthase catalytic subunit...	369	1e-102
gnl CDD 27826 cd01134, V_A-ATPase_A, V/A-type ATP synthase catalytic subunit...	334	3e-92
gnl CDD 27828 cd01136, ATPase_flagellum-secretory_path_III, Flagellum-specif...	123	1e-28
gnl CDD 27827 cd01135, V_A-ATPase_B, V/A-type ATP synthase (non-catalytic) s...	92.9	2e-19

CDD searched with vma1(cont.)

NCBI

Conserved Domains

HOME SEARCH SITE MAP Entrez CDD Structure Protein Help

cd01120.1 RecA-like_NTPases

Links: RecA-like NTPases. This family includes the NTP binding domain of F1 and V1 H+ATPases, DnaB and related helicases as well as bacterial RecA and related eukaryotic and archaeal recombinases. This group also includes bacterial conjugation proteins and related DNA transfer proteins involved in type II and type IV secretion.

Source: Pfam
Taxonomy: root
PubMed: 4 links
Proteins: cd01120 related architectures representatives

Related CD: 31 links

Statistics:
PSSM-Id: 27812
Aligned: 22 rows
PSSM: 165 columns
Status: curated CD
Created: 7-Mar-2002
Updated: 21-Aug-2003

Structure:

Program:
Drawing:
(download Cn3D)

[mouse over cd tag to display cd name]
(View large image)

cd01120 is part of a hierarchy of related CD models.

Feature 1: ATP binding site
Evidence: Comment: contains Walker A and Walker B motifs
Structure: 1C9K_C; CobU bound to GMP and pryophosphate using 3.5A - View structure with Cn3D 4.1
Structure: 1C9K_C; CobU bound to GDP (paper) - View structure with Cn3D 4.1
Citation: PMID 10529169

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.jgi.doe.gov/> The Joint Genome Institute
A recent addition is the integrated microbial genomes site at <http://img.jgi.doe.gov/>, 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 more 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*
- <http://www.ensembl.org/>

Ensembl Genome Browser (Eukaryotic genomes, including Human and Mouse genomes)

UNIX

Basic UNIX commands

ls, cd, chmod, cp, rm, mkdir, more (or) less, vi, ps, kill -9, man
A brief listing is [here](#)

chmod is a particular pain in the

Under unix every file has an owner and the owner, his group and everyone else have permissions to read, write and/or execute the file (or they don't). If you want to see which permissions are currently assigned to your files, type ls -l at the command prompt.

chmod a+x *.pl gives everyone execute permission for all files that end with .pl the * is a wildcard. (warning don't ever use rm in conjunction with *)

For more on chmod type "man chmod" or see [here](#).

(In the OSX GUI you can control click at a file, and change permissions in the info box). Most ssh clients (FUGU and SSH) allow you to use a GUI to change file permissions (in FUGU ctrl click).

Unix - command line interface

If you tried to execute a command, and you made a mistake, for example, you mistyped a file name, you can recall the last command using the up arrow (down arrow for more recent).

**If you are tired typing long filenames, you can use the tab key to complete the line, provided there is only one way to complete the line. E.g: `cd /Desktop` could be replaced by `cd /D<tab>`
If there are two or more choices you hear a boing, if you hit `<tab>` again, you get a list of choices.**

writing Perl scripts

Use unix/ linux /OsX if possible (talk with Tim if you want to use windows).

A) open a terminal window ; type "which perl <return>"

B) SSH to a unix machine (cluster OsX), log in, type "which perl <return>"

C) to check the version type perl -v <return>The response of the system should tell you, where Perl is installed on your machine (you need to know this for the first line of your perl program, which tells the operating system how to interpret what follows. On most installations this is `#!/usr/bin/perl`).

WINDOWS: If you use a windows machine, you can use an ssh program to connect to the biotech cluster. A good ssh client is available at <ftp://ftp.ssh.com/pub/ssh/>- highly recommended. I am sure that there are editors available that are more useful than notepad, but I don't know of them. :(

MAC OsX: If you use a Mac under OS X, and you do not want to (only) use the PERL locally, you want to install both jellyfish (ssh terminal) and fugu (a secure file transfer program). Both are available at <ftp://ftp.uconn.edu/pub/packages/ssh/mac/> or through the people who wrote the software - GOOGLE)

Also, the bbcxsr1 is available as a server using ssh or afp. You can connect to it from the finder menu (-> GO -> Connect to Server) pasting the following into the menu box `afp://bbcxsr1.biotech.uconn.edu` (select your account).

LINUX: Most editors on linux systems recognize Perl programs and provide context dependent coloring. Ssh and Konquerer work well for file transfer.

characters at the end of lines

File transfers from Windows to UNIX and return:

End of Line characters are a problem. Under Windows DO NOT use notepad, it does not understand UNIX newline symbols '\n'.

Best write your programs under UNIX using vi or vim (or any other editor you are comfortable with)

2nd best is to use a text editor like [textwrangler](#) (very nice and free program for UNIX). Like vi and vim it provides context dependent coloring.

3rd best is to remove end of line symbols in a UNIX editor or use sed (Stream EDitor) after you transferred the file:

```
sed s/.$// name_of_WINDOWS_infile > name_of_UNIX_outfile
```

(This replaces the last non letter character before the eol (\$) with nothing)

Some versions of office allow to change files as UNIX textfiles, but ...

A related problem is encountered by Mac users. Most text editors will use MAC carriage returns at the end of the line. Most unix programs will not be able to handle these. In a terminal window you could use the following command to convert your file:

```
tr '\r' '\n' < name_of_the_Mac_file > name_of_the_unix_file
```

If you are working in a GUI environment, you also could use the convertNewLines.app program (install it in your application folder, drag the file you want to convert into the icon). The program is available [here](#). This is very inconvenient, but there really is no easy solution, tough luck; and you better know about this incase something goes wrong.

vi

A short introduction to vi is at <http://goforit.unk.edu/unix/unix11.htm> -- however, if you run into problems google usually helps (e.g. google: vi replace unix gives you many pages of info on how to replace one string with another under vi)

```
vi myprogram.pl #starts the editor and loads the file myprogram.pl into the editor
```

The following should get you started:

The arrow keys move the cursor in the text

(if you have a really dumb terminal you can use the letter hjkl to move the cursor)

x deletes the character under the cursor
esc leaves the edit mode
i enters the edit mode and inserts before the cursor
a enters the edit mode and appends

esc : opens a command line (here you can start searches, and replacements)

:w #saves the file

:w new_name_of_file #writes the file into a new file.

:wq #saves the file and exits vi

:q! #exits vi without saving

customizing vi

One of the beauties of vi is that usually it provides context dependent coloring.

You need to tell vi which terminal you use.

One way to do so is to add a file called `.vimrc` to your home directory.

The following works under both, MAS OSX and using ssh via the secure shell program under windows:

```
vi .vimrc #opens vi to edit .vimrc (Files that start with a dot are not listed if you list a directory. List with ls -a)
```

```
set term=xterm-color #tells the editor that you use a terminal that conforms to some standard
```

```
syn on # tells the editor program that you want to use syntax dependent coloring.
```

```
esc:wq
```

This might seem a little inconvenient, but it really comes in handy to trouble shoot the program in the same environment where you want to run it.

(comment on textwrangler alternative, ssh is included inside the program)

PERL conventions and rules

Basic Perl Punctuation:

line ends with “;”

empty lines in program are ignored

comments start with #

first line points to path to interpreter:

```
#! /usr/bin/perl
```

“#!” is known as “shebang”;

keep one command per line for readability

For shell scripts the first line would refer to the type of shell to be used, e.g.:

```
#!/bin/bash
```

use indentation do show program blocks.

Variables start with **\$**scalars, **@**rarrays, or **%**ashes

Scalars: floating point numbers, integers,
non decimal integers, strings

Scalar variables are placeholders that can be assigned a scalar value (either number or string).

Scalar variables begin with \$

```
$n=3; #assigns the numerical value 3 to the variable $n.  
#Variables are interpolated, for example if you print text
```

```
$b = 4 + ($a = 3); # assign 3 to $a, then add 4 to that  
# resulting in $b getting 7  
$d = ($c = 5); # copy 5 into $c, and then also into $d  
$d = $c = 5; # the same thing without parentheses
```

```
$a = $a + 5; # without the binary assignment operator  
$a += 5; # with the binary assignment operator
```

```
$str = $str . " "; # append a space to $str  
$str .= " "; # same thing with assignment operator
```

```
"hello" . "world" # same as "helloworld"  
'hello world' . "\n" # same as "hello world\n"  
"fred" . " " . "barney" # same as "fred barney"  
"fred" x 3 # is "fredfredfred"  
"barney" x (4+1) # is "barney" x 5, or # "barneybarney....."  
(3+2) x 4 # is 5 x 4, or really "5" x 4, which is "5555"
```

Note: these are not mathematical equations but assignments!

Numbers can be manipulated using the typical symbols:

$2 + 3$ # 2 plus 3, or 5

$5.1 - 2.4$ # 5.1 minus 2.4, or approximately 2.7;

$3 * 12$ # 3 times 12 = 36;

$2^{**}3$ # 2 taken to the third power = $2*2*2 = 8$

$14 / 2$ # 14 divided by 2, or 7;

$10.2 / 0.3$ # 10.2 divided by 0.3, or approximately 34;

$10 / 3$ # always floating point divide, so approximately 3.3333333...

Special characters:

```
\n #newline  
\t #tab
```

Double quoted strings are interpolated by the Perl interpreter:

```
"hello world\n" # hello world, and a newline  
"coke\tsprite" # a coke, a tab, and a sprite
```

The backslash can precede many different characters to mean different things (typically called a backslash escape).

Variable interpolation - single quoted strings are not interpolated:

```
'hello' # five characters: h, e, l, l, o
'don\'t' # five characters: d, o, n, single-quote, t
'' # the null string (no characters)
'silly\\me' # silly, followed by backslash, followed by me
'hello\n' # hello followed by backslash followed by n
'hello
there' # hello, newline, there (11 characters total)
```

Assignments for next week:

Think about a topic for your student project!
Please, don't hesitate to send me an email in case you have a question.

Let me know what you are interested in (email).
What we will do in this course will in part depend on your interests.

Assignment for next Monday

1) On the computer that you plan to use for your project set up a connection (or connections) to `bbcxsrv1` that allows you

- (a) ssh to the server using a command line interface
- (b) allows you to drop and drag files from your computer to the server.

2) check that your vi editor on `bbcxsrv1` is set up to have context dependent coloring (do this, even if you don't plan to use vi on the server!).

3) if you do not want to use vi, install an editor on your computer that provides context dependent coloring.

4) Read through pages 53-61 of the [Unix and Perl Primer for Biologists](#)

4) Create first Perl Program- "Hello, world!" [make file executable using `chmod`]

```
#!/usr/bin/perl -w  
print ("Hello, world! \n");
```

What happens if you leave out the new line character?

You can run the program by typing `./program_name.pl`, if the file containing the program is made executable (using `chmod u+x *.pl`).