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 mitochondrium (+ 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 aneupolyploids

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

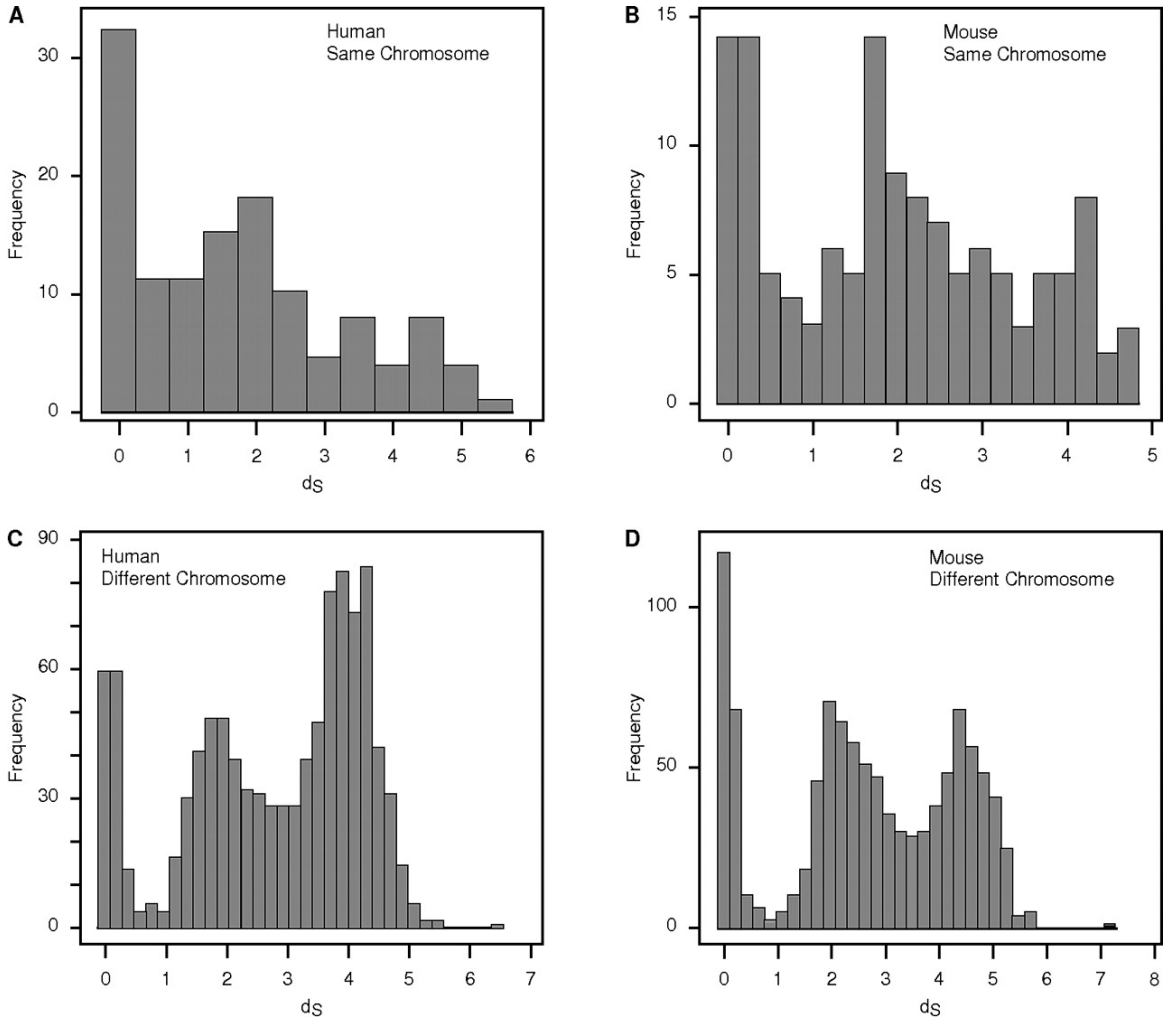
Related: Ancient genome duplication events are revealed by peaks in the divergence of paralogs.

Example: Gene versus Genome Duplications

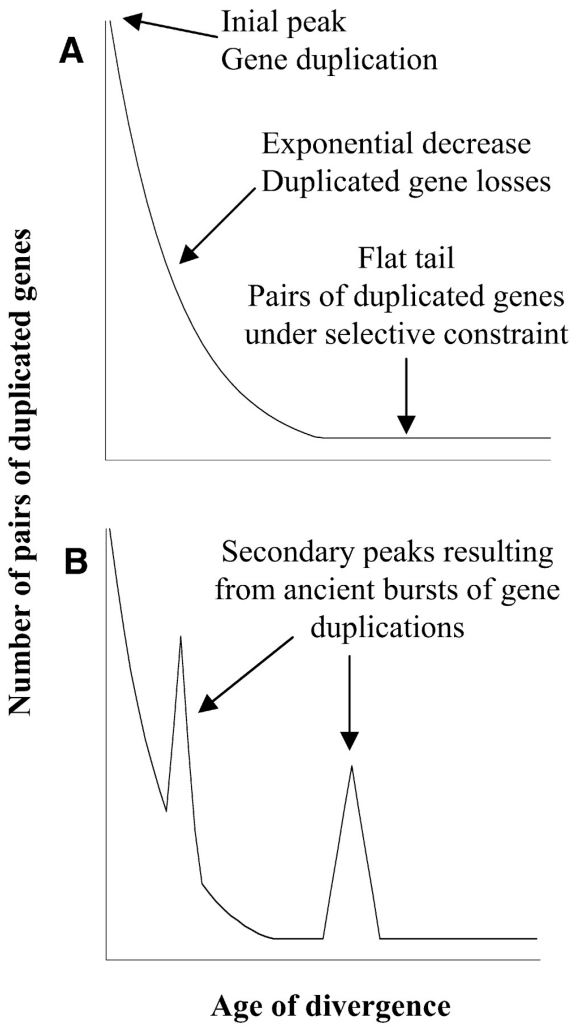
The same approach as suggested for the chimera formation can be applied to the question was the whole genome or a large segment of an organism's genome duplicated, or did the duplications occur in a piecemeal fashion?

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

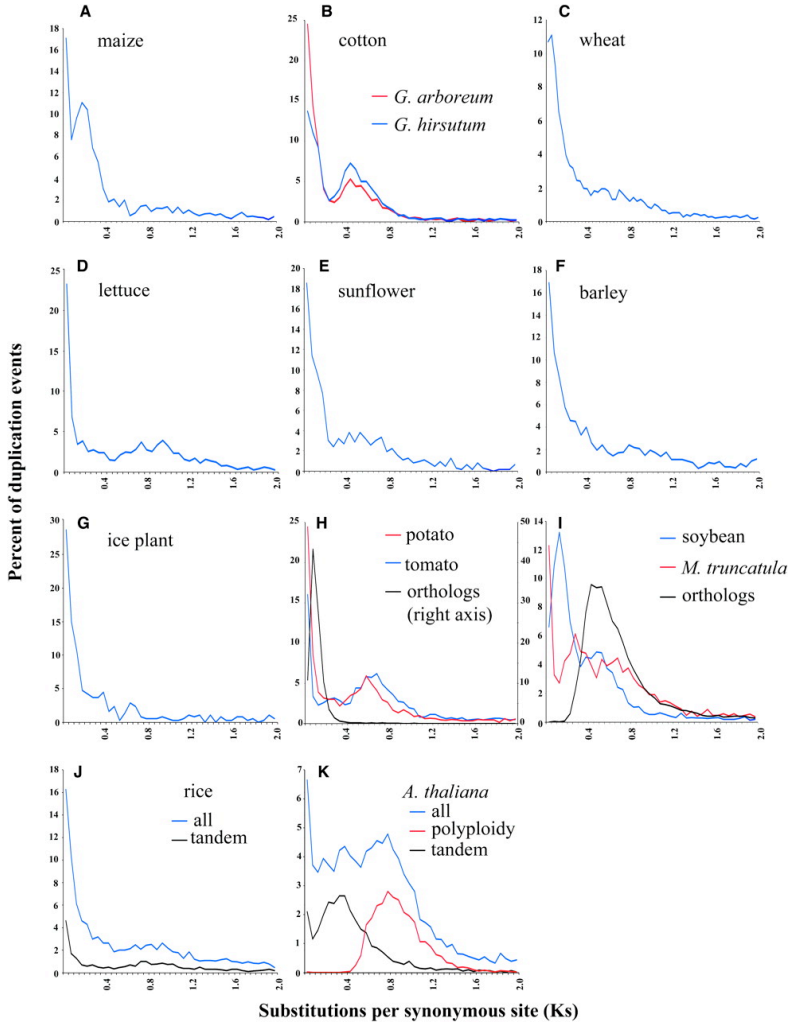


Theoretical Age Distributions of Pairs of Duplicated Genes in a Genome



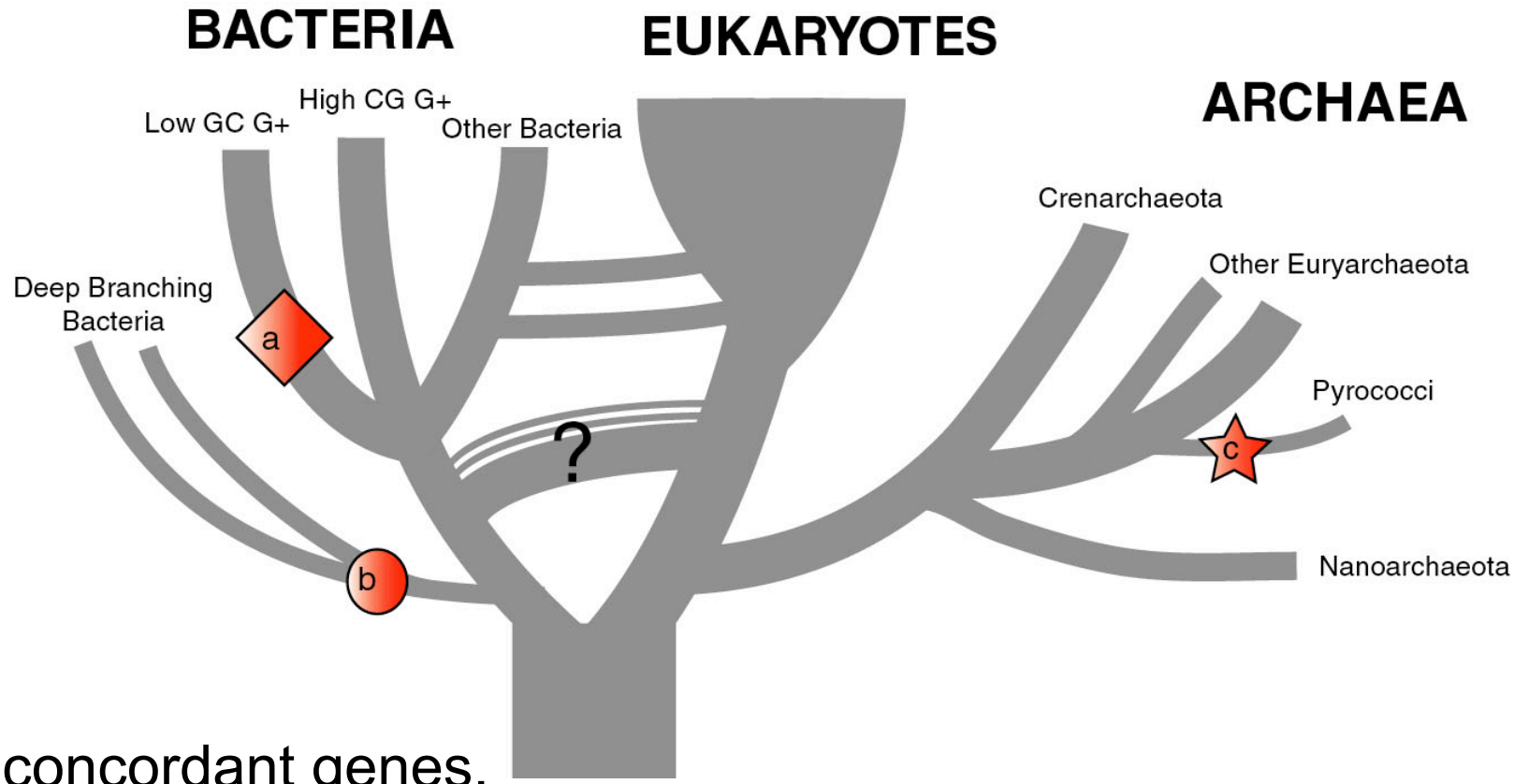
Blanc, G., et al. Plant Cell 2004;16:1667-1678

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 form 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

GTAs are propahges 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
- There molecular phylogeny should reflect 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 psudogenes).

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:

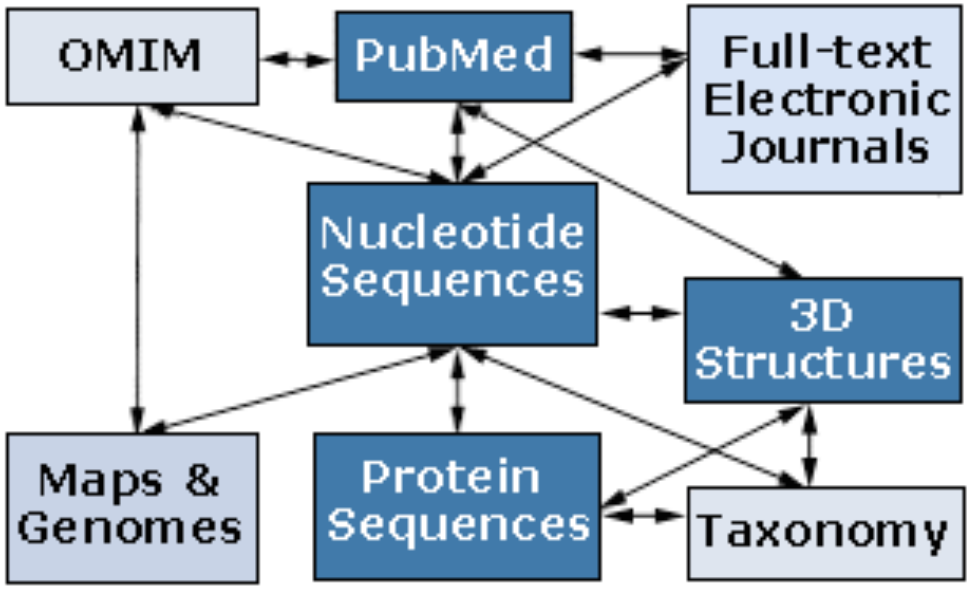
- Write a script that uses the 100+ known intein alleles each as a seed in PSI BLAST, and stores the profiles. Write a second script that uses these profiles to detect putative inteins in completely sequenced genomes.
- Same as above but use transposases, integrases, homing endonucleases, or a molecular 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)

Entrez
search and retrieval system

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 availabel
-- many duplicates
-- poor reliability of annotations

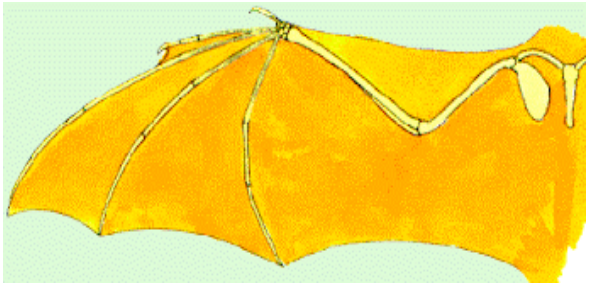
10 minute Break?

Theodosius Dobzhansky:

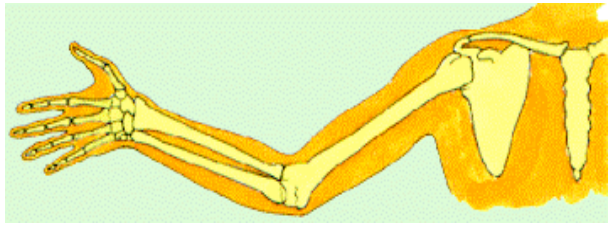
"Nothing in biology makes sense except in the light of evolution"

Homology

bird wing



bat wing



human arm

by Bob Friedman

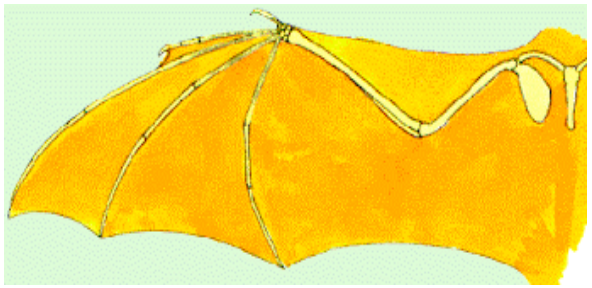
homology *vs* analogy

A priori sequences could be similar due to convergent evolution

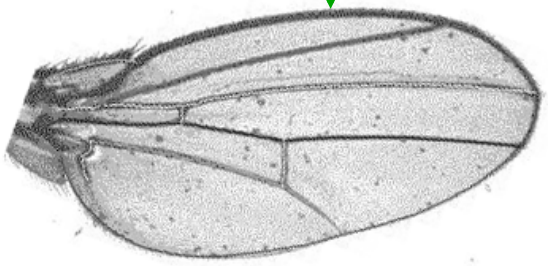
Homology (shared ancestry) *versus* Analogy (convergent evolution)

bird wing

butterfly wing



bat 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 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

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)

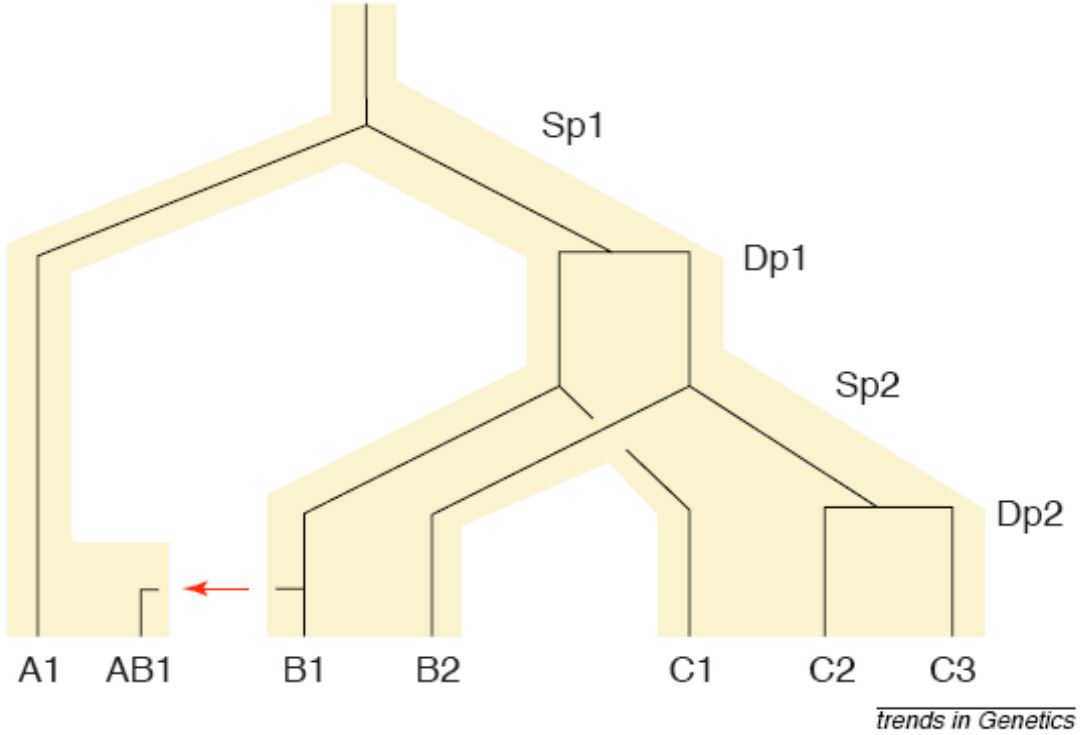
see Fitch's article in TIG 2000 for more discussion.

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* (MRCA) 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

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

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.

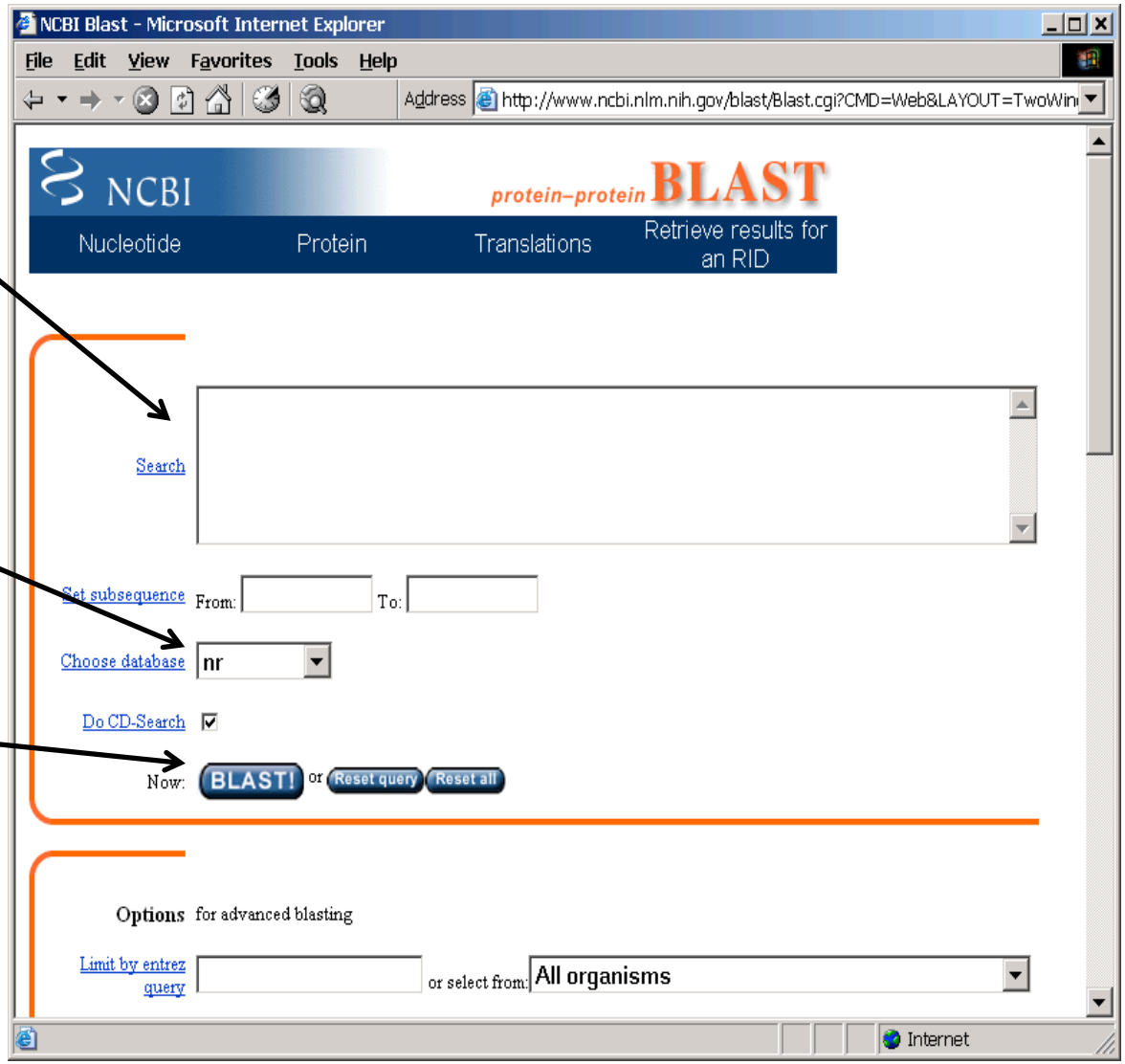
by Bob Friedman

Routine BlastP search

FASTA formatted
text
or Genbank ID#

Protein
database

Run



by Bob Friedman

BlastP parameters

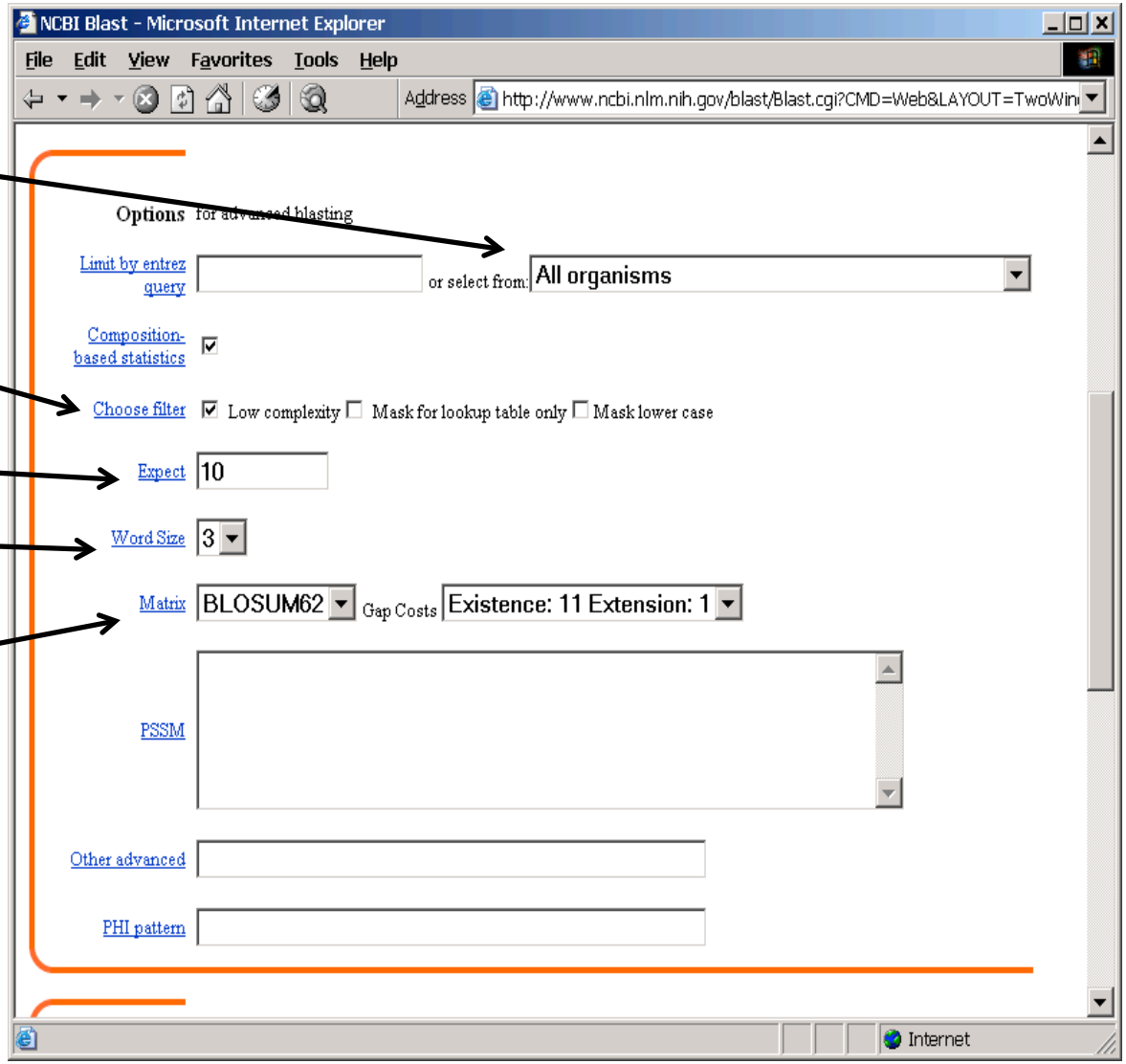
Restrict by taxonomic group

Filter repetitive regions

Statistical cut-off

Size of words in look-up table

Similarity matrix (cost of gaps)



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.

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.

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).