# MCB 371/372

Student Projects

Databanks

3/16/05

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

# Student Projects

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

## Example 1: Evolution of a gene family

- When in the evolution of evolution of the interferon (or what ever you are interested in) gene family did gene duplications occur?
- Which of the resulting subfamilies have a 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 2: 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

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

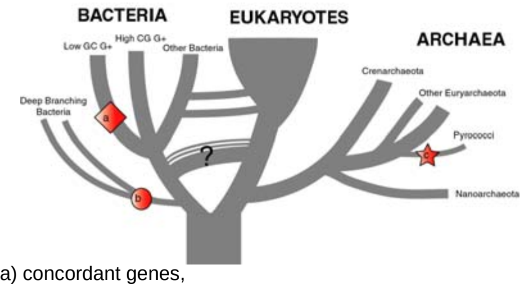
## Example 2: Chimera? continued

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

## The Phylogenetic position of *Thermotoga maritima*



(a) concordant genes,
(b) according to 16S (and other conserved genes)
(c) according to phylogenetically discordant genes

Gophna, U., Doolittle, W.F. & Charlebois, R.L.:
Weighted genome trees: refinements and applications. *J. Bacteriol.* (in press)

Example 2: Chimera? continued

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

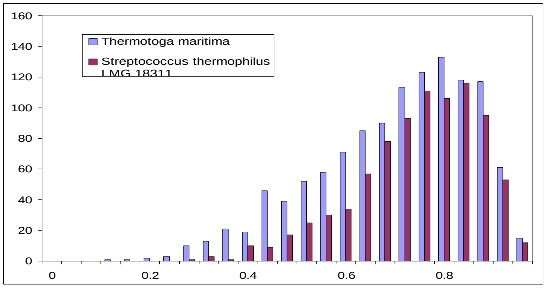
A) Two distinct types of phylogenetic affinity.

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

B) Two distinct peaks in a divergence histogram.

E.g.: If one measures the divergence Thermotoga – Archaea for all the individual genes, under the assumption of a chimera formation one should obtain a bimodal distribution in a histogram of the different genes.

Histogram of divergence to archaeal genes for two bacterial genomes



For each encoded protein BLAST searches were performed against the proteins in 5 archaeal genomes (*Pyrococcus abysii*, *P. furiosus*, *Archaeoglobus fulgidus*, *Methanocaldococcus janaschii*, and *Methanothermobacter thermautotrophicus*). The highest (bitscore divided by the alignment lengths) was utilized as a measure of sequence similarity. Relative sequence divergence between two sequences was calculates as (1-similarity(b_a)/similarity(b_b)), where similarity(b_a) is the similarity score for a bacterial sequence with the most similar archaeal one, and similarity(b_b) is the similarity score of the bacterial sequence compared with itself.

# Example 2, continued

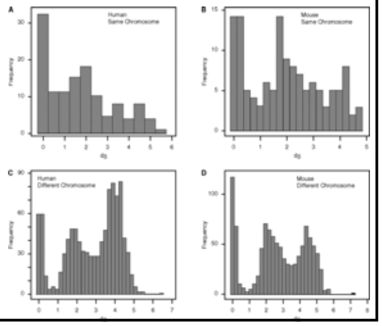
The " to-do-list" would include:

- Formulate the question you want to address
- Find a computer where you can run blastall (this might take a couple of hours)
- Download and analyze the required genomes
- 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?

## Example 3: 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?

**Bob Friedman will be the new Bioinformatics services specialist at UConn's Bioinformatics services facility !**



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

## Assignments:

Pick 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.

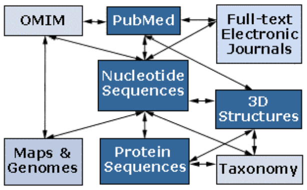
Reading for Monday:
Read through the NCBI's BLAST tutorial

## 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.
- Demonstrate Link>Books

# 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: | Repositories without gatekeeper. Examples: |
|---|---|
| Swissprot | GenBank |
| Refseq (at NCBI) | EMBL |
| | TrEMBL |
| Entries are checked for accuracy. | Everything is accepted |
| + more reliable annotations | + everything is availabel |
| -- frequently out of date | -- many duplicates |
| | -- poor reliability of annotations |

## 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*