MCB5472
Computer methods in molecular evolution

Lecture 4/21/2014

Signup sheet for presentations

https://docs.google.com/spreadsheet/ccc?key=0AjJsMXOC-NEVdEhuenJ3bFEwdUJRUno1alg0aVlvcEE&usp=sharing

sites versus branches

You can determine omega for the whole dataset; however, usually not all sites in a sequence are under selection all the time.

PAML (and other programs) allow to either determine omega for each site over the whole tree, *Branch Models* **, or determine omega for each branch for the whole sequence,** *Site Models* **.**

It would be great to do both, i.e., conclude codon 176 in the vacuolar ATPases was under positive selection during the evolution of modern humans – alas, a single site does not provide much statistics ….

ML Ratio test

In case of two nested models that differ by n parameters, one can test if the increase in likelihood (=probability of the data) is significant, i.e., more than expected from having an additional parameter.

E.g., phylogenetic tree with and without clock. In this case, the model without clock is the more complex model, that has n-1 more parameters (the branch lengths leading to the leaves) than the clock model.

Using the atp_all.phy example from tree-puzzle

ML Ratio test

atp_all.phy example from tree-puzzle

go through outfile_atp_all_puzzle_clock.out (at http://gogarten.uconn.edu/mcb5472_2014/outfile_atp_all_puzzle_clock.out

```
MOLECULAR CLOCK LIKELIHOOD RATIO TEST FOR USER TREE # 1

log L without clock: -35429.62 (independent branch parameters: 89)
log L with clock:    -35650.78 (independent branch parameters: 45)

Likelihood ratio test statistic delta: 442.31
Degrees of freedom of chi-square distribution: 44
Critical significance level: 0.00%

The simpler (clocklike) tree is rejected on a significance level
of 5%. The log-likelihood of the more complex (no clock) tree is
significantly increased.

Please take care that the correct root is used!
```

ML Ratio test

codeml example hv1.phy

Control file

```
model = 0   * 0:one, 1:b, 2:2 or more dN/dS ratios for branches

NSsites = 0 1 2    * 0:one w;1:neutral;2:selection; 3:discrete;4:freqs;
                   * 5:gamma;6:2gamma;7:beta;8:beta&w;9:beta&gamma;
                   * 10:beta&gamma+1; 11:beta&normal>1; 12:0&2normal>1;
                   * 13:3normal>0
```

Output file:

http://gogarten.uconn.edu/mcb5472_2014/Hv1.sites.codeml.out

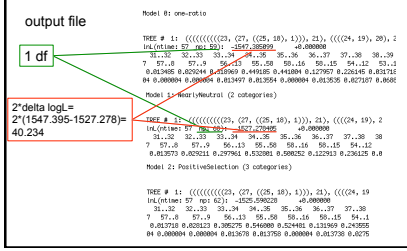
ML Ratio test

codeml example hv1.phy

output file

1 df

2*delta logL= 2*(1547.395-1527.278)= 40.234



ML Ratio test

40 is way off the table

Degrees of freedom (df)	χ^2 value[17]									
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

The improvement in fit, when sites under neutral **and** under purifying selection are allowed, is very significant (P<<0.0001)

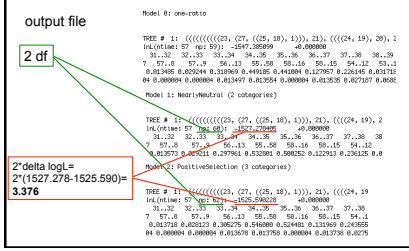
ML Ratio test

codeml example hv1.phy

output file

2 df

2*delta logL= 2*(1527.278-1525.590)= **3.376**



3.37 with 2 df

Degrees of freedom (df)	χ^2 value[17]									
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

The improvement in fit, when sites under positive (diversifying selection are added is not significant (P>0.05)

H_0: variable selective pressure but NO positive selection (M1a)

H_1: variable selective pressure with positive selection (M2a)

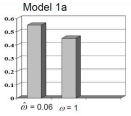
Likelihood ratio test (LRT)

$2\Delta l = 2\,(l_1 - l_0)$ χ^2 distribution

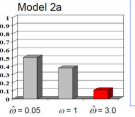
Parameters in Site Models

Model	NSsites	p	Parameters
M1a (neutral)	1	2	p_0 ($p_1 = 1 - p_0$), $\omega_0 < 1$, $\omega_1 = 1$
M2a (selection)	2	4	p_0, p_1 ($p_2 = 1 - p_0 - p_1$), $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 1$

Model 1a



Model 2a



SE:

- IL MODELLO M2a SI ADATTA MEGLIO AI DATI
- IL VALORE DI ω STIMATO E' >1 PER LA CLASSE DI SITI p_2

UNA PROPORZIONE DI SITI PARI A p2 E' SOTTOPOSTA A SELEZIONE POSITIVA

From: http://m.docente.unife.it/silvia.fuselli/dispense-corsi/EM_Lezione%208.pdf

H_0: Beta distributed variable selective pressure (M7)

H_1: Beta plus positive selection(M8)

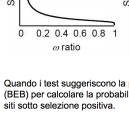
Likelihood ratio test (LRT)

$2\Delta l = 2\,(l_1 - l_0)$ χ^2 distribution

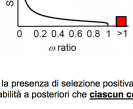
Parameters in Site Models

Model	NSsites	p	Parameters
M7 (beta)	7	2	p, q
M8 (beta&ω)	8	4	p_0 ($p_1 = 1 - p_0$), p, q, $\omega_s > 1$

M7: beta



M8: beta&ω



SE:

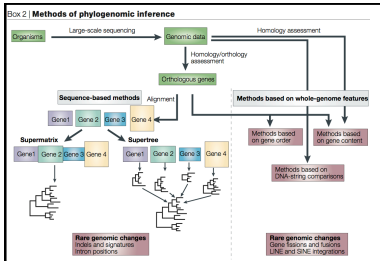
- IL MODELLO M8 SI ADATTA MEGLIO AI DATI
- IL VALORE DI ω_s STIMATO E' >1 PER LA CLASSE DI SITI p1

UNA PROPORZIONE DI SITI PARI A p1 E' SOTTOPOSTA A SELEZIONE POSITIVA

Quando i test suggeriscono la presenza di selezione positiva si utilizzano dei metodi Bayesiani (BEB) per calcolare la probabilità a posteriori che **ciascun codone** provenga dalla classe di siti sotto selezione positiva.

From: http://m.docente.unife.it/silvia.fuselli/dispense-corsi/EM_Lezione%208.pdf

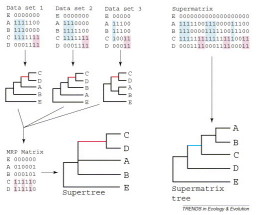
Box 2 | **Methods of phylogenomic inference**



The flowchart shows steps in the inference of evolutionary trees from genomic data. Genomic information is obtained by large-scale DNA sequencing. In general, sets of orthologous genes are then assembled from specific sets of species for phylogenetic analysis. This homology or orthology assessment is a crucial step that is almost always based on simple similarity comparisons (for example, BLAST[18] searches). Most methods used for the subsequent reconstruction of phylogenetic trees are either sequence-based or are based on whole-genome features.

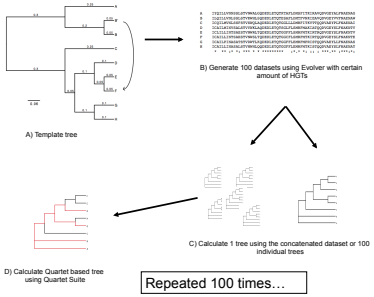
From: Daubin C, Binkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 2005 May;6(5):361-75

Supertree *vs.* Supermatrix

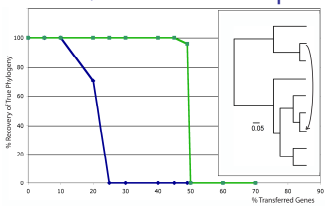


Schematic of MRP supertree (left) and parsimony supermatrix (right) approaches to the analysis of three data sets. Clade C+D is supported by all three separate data sets, but not by the supermatrix. Synapomorphies for clade C+D are highlighted in pink. Clade A+B+C is not supported by separate analyses of the three data sets, but is supported by the supermatrix. Synapomorphies for clade A+B +C are highlighted in blue. E is the outgroup used to root the tree.

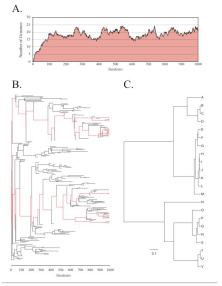
From: Alan de Queiroz John Gatesy: The supermatrix approach to systematics Trends Ecol Evol. 2007 Jan;22(1):34-41



Supermatrix versus Quartet based Supertree

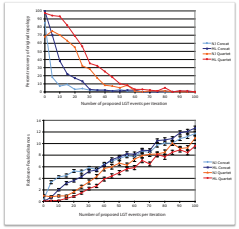


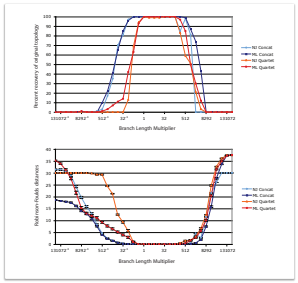
inset: simulated phylogeny



Note : Using same genome seed random number will reproduce same genome history

HGT EvolSimulator Results

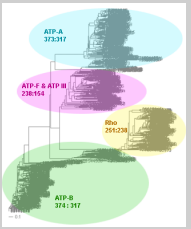




- See http://bib.oxfordjournals.org/content/15/1/79.full for more information
- What is the bottom line?

Automated Assembly of Gene Families Using BranchClust

J. Peter Gogarten
University of Connecticut
Dept. of Molecular and Cell Biol.



Collaborators:
Maria Poptsova (UConn)
Fenglou Mao (UGA)

Funded through the
Edmond J. Safra Bioinformatics Program.
Fulbright Fellowship,
NASA Exobiology Program,
NSF Assembling the Tree of Life Program and
NASA Applied Information Systems Research Program

Why do we need gene families?

Which genes are common between different species?

Which genes were duplicated in which species?
(Lineage specific gene family expansions)

Do all the common genes share a common history?
Reconstruct (parts of) the tree/net of life /
Detect horizontally transferred genes.

Why do we need gene families?

Help in genome annotation.

A) Genes in a family should have same annotation across species (usually).

B) Genes present in almost all genomes of a group of closely related organisms, but absent in one or tow members, might represent genome annotation artifacts.

Selection of Orthologous Gene Families

All automated methods for assembling sets of orthologous genes are based on sequence similarities.

BLAST hits

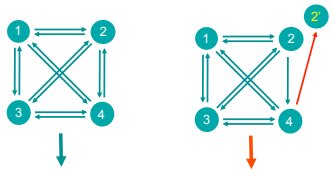
Triangular circular BLAST significant hits
(COG, or Cluster of Orthologous Groups)

Sequence identity of 30% and greater
(SCOP database)

Similarity complemented by HMM-profile analysis
Pfam database

Reciprocal BLAST hit method

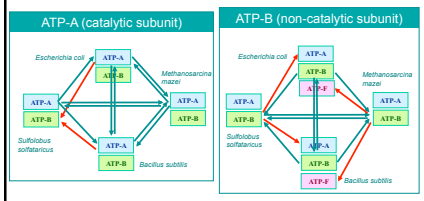
Strict Reciprocal BLAST Hit Method



often fails in the presence of paralogs

Families of ATP-synthases

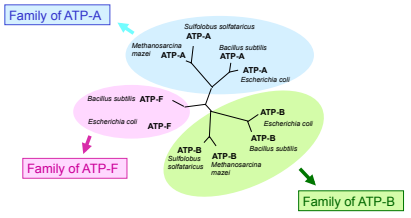
Case of 2 bacteria and 2 archaea species



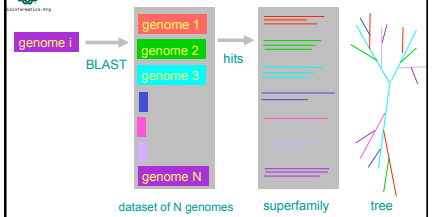
Neither ATP-A nor ATB-B is selected by RBH method

Families of ATP-synthases

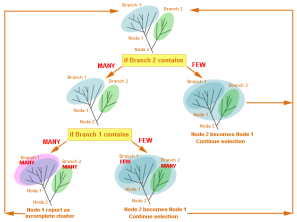
Phylogenetic Tree



BranchClust Algorithm



BranchClust Algorithm



BranchClust Algorithm

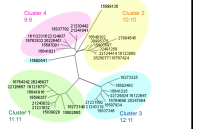
Root positions

Superfamily of penicillin-binding protein

13 gamma proteo bacteria

Superfamily of DNA-binding protein

13 gamma proteo bacteria



BranchClust Algorithm

Comparison of the best BLAST hit method and BranchClust algorithm

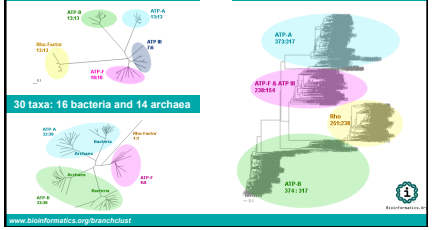
Number of taxa - A: Archaea B: Bacteria	Number of selected families:	
	Reciprocal best BLAST hit	BranchClust
2A 2B	80	414 (all complete)
13B	236	409 (263 complete, 409 with n≥8)
16B 14A	12	126 (60 complete, 126 with n≥24).

BranchClust Algorithm

ATP-synthases: Examples of Clustering

13 gamma proteobacteria

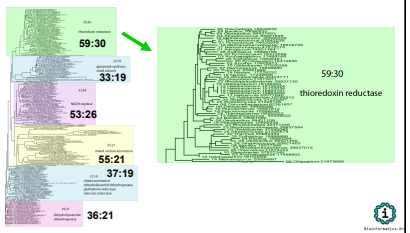
317 bacteria and archaea



30 taxa: 16 bacteria and 14 archaea

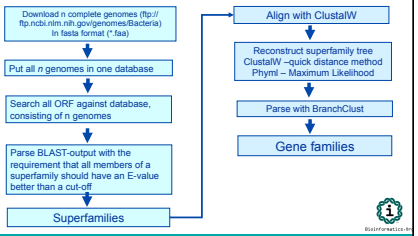
BranchClust Algorithm

Typical Superfamily for 30 taxa (16 bacteria and 14 archaea)



BranchClust Algorithm

Data Flow



BranchClust Algorithm

Implementation and Usage

The BranchClust algorithm is implemented in Perl with the use of the BioPerl module for parsing trees and is freely available at http://bioinformatics.org/branchclust

Required

1. Bioperl module for parsing trees Bio::TreeIO
2. Taxa recognition file gi_numbers.out must be present in the current directory. For information on how to create this file, read the Taxa recognition file section on the web-site.
3. Blastall from NCB needs to be installed.

to use genomes from NCBI:

- The easiest source for other genomes is via anonymous ftp from ftp.ncbi.nlm.nih.gov
  Genomes are in the subfolder genomes.
  Bacterial and Archaeal genomes are in the subfolder Bacteria
- For use with BranchClust you want to retrieve the .faa files from the folders of the individual organisms (in case there are multiple .faa files, download them all and copy them into a single file).
- Copy the genomes into the fasta folder in directory where the branchclust scripts are.
- To create a table that links GI numbers to genomes run perl extract_gi_numbers.pl or `qsub extract_gi_numbers.sh`

If you use other genomes you will need to generate a file that contains assignments between name of the ORF and the name of the genome. This file should be called gi_numbers.out

If your genomes follow the JGI convention, every ORF starts with four letters designating the species followed by 4 numbers identifying the particular ORF. In this case the file gi_numbers.out should look as follows. It should be straight forward to create this file by hand ☺

```
Thermotoga maritima |   Tmar.....
Thermotoga naphthophila |    Tnap.....
Thermotoga neapolitana |    Tnea.....
Thermotoga petrophila | Tpet.....
Thermotoga sp. RQ2 |   TRQ2.....
```

If your genomes conform to the NCBI *.faa convention, put the genomes into a subdirectory called fasta, and run the script extract_gi_numbers.pl in the parent directory.

The script should generate a log file and an output file called gi_numbers.out

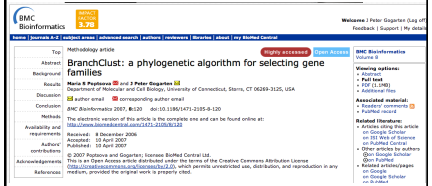
```
Burkholderia phage Bcep781 | 2375....      4783....      1179.....
Enterobacteria phage K1F |   7711....
Enterobacteria phage N4 |    1199.....
Enterobacteria phage P22 |   5123....      9635... 1271....
      193433..
Enterobacteria phage RB43 |  6639....
Enterobacteria phage T1 |    4568....
Enterobacteria phage T3 |    1757....
Enterobacteria phage T5 |    4640....
Enterobacteria phage T7 |    9627...
Kluyvera phage Kvp1 | 2126.....
Lactobacillus phage phiAT3 |  4869....
Lactobacillus prophage Lj965 |      4117....
Lactococcus phage c1t |     2345....
Lactococcus phage sk1 |     9629... 193434..
Mycobacterium phage Bxz2 |  29566...
```

the branchclust scripts

- are available at http://www.bioinformatics.org/branchclust/
- Consult the tutorial http://www.bioinformatics.org/branchclust/BranchClustTutorial.pdf

BranchClust Article

- is available at http://www.biomedcentral.com/1471-2105/8/120



Create super families, alignments and trees

```
perl do_blast.pl
```

Super Families to Trees

- `perl parse_superfamilies_singlelink.pl 4`
  # 4 gives the minimum size of the superfamily
- `perl prepare_fa.pl parsed/all_vs_all.fam`
  Creates a multiple fasta file for each superfamily
- `perl do_clustalw_aln.pl`
  aligns sequences using clustalw
- `perl do_clustalw_dist_kimura.pl`
  calculates trees using Kimura distances for all families in fa #trees stored in trees
- `perl prepare_trees.pl`
  reformats trees

Branchclust

```
perl branchclust_all.pl 10
 # Parameter 10 (MANY) says that a family needs to have
 # at least 10 members.

perl make_fam_list.pl 15 10
# results in file called families.list
15 gives the number of genomes, 10 the minimum size of the gene family
```

Process Branchclust output

```
perl names_for_cluster_all.pl
#  (Parses clusters and attaches names.
#  Results in sub directory clusters. List in test)

perl summary.pl
# (makes list of number of complete and incomplete families
# file is stored in test)

perl detailed_summary_dashes.pl
# (result in test: detailed_summary.out - can be used in Excel)

perl prepare_bcfam.pl families.list
#(writes multiple fasta files into bcfam subdirectory.
# Can be used for alignment and phylogenetic reconstruction)
```

Summary Output

- `complete: 1564`
- `incomplete: 248`
- `total: 1812`
- `------ details -------`
- `incomplete 4: 87`
- `incomplete 3: 53`
- `incomplete 2: 66`
- `incomplete 1: 42`

done with many = 3 and E-value cut-off of 10^{-25}

Detailed Summary in Excel

- copy detailed summary out onto your computer
- In EXEL Menu: Data -> get external data -> import text file -> in English version use defaults for other options.
- In EXEL Menu: Data -> sort -> sort by "superfamily number"-> if asked, check expand selection
- Scrolling down the list, search for a superfamily that was broken down into many families.

Do the families that were part of a superfamily have similar annotation lines?
How many of the families were complete?
Do any have inparalogs? Take note of a few super families.

superfamily_##	ny_ ##	he_fa mily	nu_of_p aralogs	family_name
129	51	2	0	Tnea_0520 Inositol transport system ATP-binding protein
129	52	2	0	TRQ2_1091 oligopeptide ABC transporter, ATP-binding protein
129	53	1	0	Tnea_0642 ABC transporter related
129	54	1	0	Tnap_0004 oligopeptide/dipeptide ABC transporter, ATPase subunit
129	55	5	0	TRQ2_0768 ABC transporter related
129	56	4	0	Tpet_0304 sugar ABC transporter, ATP-binding protein
129	57	5	0	TRQ2_0228 ABC transporter related
129	58	5	0	TRQ2_0461 ABC transporter related
129	59	5	0	TRQ2_0594 ABC transporter related
129	60	1	0	Tnap_0003 oligopeptide/dipeptide ABC transporter, ATPase subunit
129	61	5	0	TRQ2_1593 Phosphate transport ATP-binding protein PstB (TC 3.A.1.7.1)
129	62	1	0	Tnea_0924 ABC transporter related
130	1	5	0	TRQ2_0139 Putative preQ0 transporter
131	1	5	0	TRQ2_0140 NADPH dependent preQ0 reductase
132	1	5	0	TRQ2_0141 Phosphomethylpyrimidine kinase (EC 2.7.4.7) / Thiamin-phosphate syn

clusters/clusters_NNN.out.names

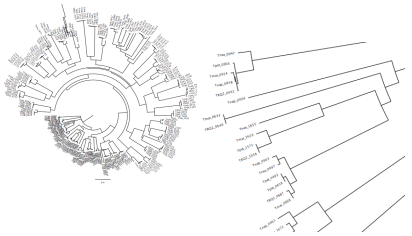
- *Are all the annotation lines uniform?*
- Within this report, if there are inparalogs, one is listed as a family member, the other one as inparalog. This is an arbitrary choice, both inparalogs from the same genome should be considered as being part of of the family.
- Out of cluster paralogs are paralogs that did not make it into a cluster with "many" genomes.

```
COMPLETE: 5
---------- CLUSTER ----------
>lcl|Tnea_1049 ABC transporter related [Thermotoga neapolitana]
>lcl|TRQ2_0990 ABC transporter related [Thermotoga sp. RQ2]
>lcl|Tnea_1896 Ribose ABC transport system, ATP-binding protein RbsA (TC 3.A
>lcl|Tmar_1872 Ribose ABC transport system, ATP-binding protein RbsA (TC 3.A
>lcl|Tpet_1811 ABC transporter related [Thermotoga petrophila]
>lcl|Tnap_1536 ABC transporter related [Thermotoga naphthophila]

---------- FAMILY ----------
>lcl|Tmar_1872 Ribose ABC transport system, ATP-binding protein RbsA (TC 3.A
>lcl|Tnap_1536 ABC transporter related [Thermotoga naphthophila]
>lcl|Tnea_1049 ABC transporter related [Thermotoga neapolitana]
>lcl|Tpet_1811 ABC transporter related [Thermotoga petrophila]
>lcl|TRQ2_0990 ABC transporter related [Thermotoga sp. RQ2]

COMPLETE: 5
>>>>> IN-PARALOGS ----------
>lcl|Tnea_1896 Ribose ABC transport system, ATP-binding protein RbsA (TC 3.A
```

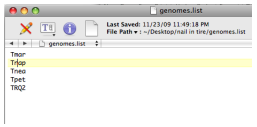
trees/fam_XYZ.tre



The Quartet Decomposition Server

http://csbl1.bmb.uga.edu/QD/phytree.php

Input A):
a file listing the names of genomes: E.g.:



The Quartet Decomposition Server

http://csbl1.bmb.uga.edu/QD/phytree.php

Input B):
An Archive of files where every file contains all the trees that resulted from a bootstrap analysis of one gene family:

One file per family

100 trees per file

The Quartet Decomposition Server

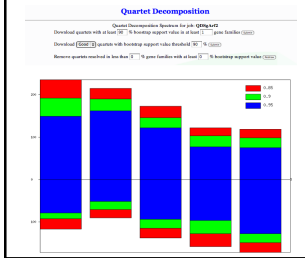
http://csbl1.bmb.uga.edu/QD/phytree.php

Trees from the bootstrap samples should contain branch lengths, but the name for each sequence should be translated to the genome name, using the names in the genome list.
See the following three trees in Newick notation for an example:

(((Tnea:0.1559823230,Tpet:0.0072068797):0.0287486818,Tmar:0.0046676053):0.0407339037,Tnap:0.0000000001,TRQ2:0.0000000001);
(((Tpet:0.0219514318,Tnea:0.1960236242):0.0145181752,Tmar:0.0189973964):0.0155785587,Tnap:0.0000000001,TRQ2:0.0000000001);
(((Tpet:0.0000004769,Tnea:0.1773430420):0.0205769649,Tmar:0.0047117206):0.0416898504,Tnap:0.0000000001,TRQ2:0.0000000001);

The spectrum

http://csbl1.bmb.uga.edu/QD/jobstatus.php?jobid=QDSgArf2&source=0&resolve=0&support=0



good and bad quartets

Quartet Decomposition

Good quartets with bootstrap support value > 0.9
Download as newick trees

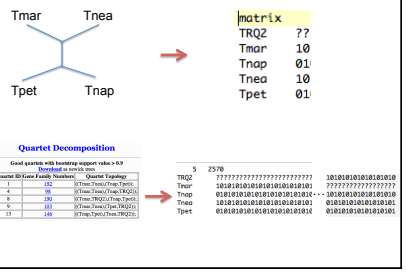
Quartet ID	Gene Family Numbers	Quartet Topology
1	192	((Tmar,Tnea),(Tnap,Tpet));
4	98	((Tmar,Tnea),(Tnap,TRQ2));
8	190	((Tmar,TRQ2),(Tnap,Tpet));
9	103	((Tmar,Tnea),(Tpet,TRQ2));
13	146	((Tnap,Tpet),(Tnea,TRQ2));

Quartet Decomposition

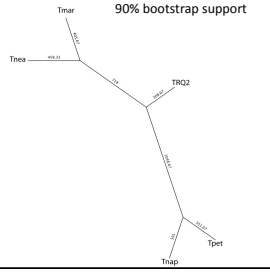
Bad quartets with bootstrap support value > 0.9
Download as newick trees

Quartet ID	Gene Family Numbers	Quartet Topology
0	38	((Tmar,Tnap),(Tnea,Tpet));
2	55	((Tmar,Tpet),(Tnap,Tnea));
3	64	((Tmar,Tnap),(Tnea,TRQ2));
5	85	((Tmar,TRQ2),(Tnap,Tnea));
6	46	((Tmar,Tnap),(Tpet,TRQ2));
7	25	((Tmar,Tpet),(Tnap,TRQ2));
10	37	((Tmar,Tpet),(Tnea,TRQ2));
11	71	((Tmar,TRQ2),(Tnea,Tpet));
12	66	((Tnap,Tnea),(Tpet,TRQ2));
14	49	((Tnap,TRQ2),(Tnea,Tpet));

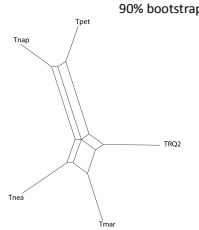
Quartets -> Matrix Representation Using Parsimony



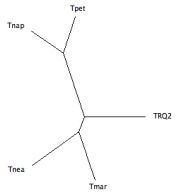
Most Parsimonious Tree (MRP)
Using all Quartets from all Gene Families that have more than 90% bootstrap support



Splits Tree Representation
Using all Quartets from all Gene Families that have more than 90% bootstrap support



Split Decomposition tree
from uncorrected P distances



NJ tree
from uncorrected P distances