MCB 5472

Assembly of Gene Families

Peter Gogarten Office: BSP 404 phone: 860 486-4061, Email:

gogarten@uconn.edu

PAML (codeml) the basic model

 $q_{ij} = \begin{cases} 0, & \text{if the two codons differ at more than one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa \pi_j, & \text{for synonymous transition,} \\ \omega \pi_j, & \text{for nonsynonymous transversion,} \\ \omega \kappa \pi_j, & \text{for nonsynonymous transition,} \end{cases}$

The equilibrium frequency of $\operatorname{codon} j(\pi_j)$ can be considered a free parameter, but can also be calculated from the nucleotide frequencies at the three codon positions (control variable CodonFreq). Under this model, the relationship holds that $\omega = d_N/d_S$, the ratio of nonsynonymous/synonymous substitution rates. This basic model is fitted by specifying model = 0 NSsites = 0, in the control file codeml.ctl. It forms the basis for more sophisticated models implemented in codeml.

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 any statistics

PAML – codeml – sites model (cont.)

the program is invoked by typing codeml followed by the name of a control file that tells the program what to do.

paml can be used to find the maximum likelihood tree, however, the program is rather slow. Phyml is a better choice to find the tree, which then can be used as a user tree.

An example for a codeml.ctl file is <u>codeml.hv1.sites.ctl</u> This file directs codeml to run three different models: one with an omega fixed at 1, a second where each site can be either have an omega between 0 and 1, or an omega of 1, and third a model that uses three omegas as described before for MrBayes. The output is written into a file called Hv1 sites codeml, out (as directed by

The output is written into a file called <u>Hv1.sites.codeml_out</u> (as directed by the control file).

Point out log likelihoods and estimated parameter line (kappa and omegas)

Additional useful information is in the <u>rst</u>file generated by the codeml

Discuss overall result.

PAML – codeml – branch model

For the same dataset to estimate the dN/dS ratios for individual branches, you could use this file <u>codeml.hv1.branches.ctl</u> as control file.

The output is written, as directed by the control file, into a file called <u>Hv1.branch.codeml_out</u>

A good way to check for episodes with plenty of non-synonymous substitutions is to compare the dn and ds trees.

Also, it might be a good idea to repeat the analyses on parts of the sequence (using the same tree). In this case the sequences encode a family of spider toxins that include the mature toxin, a propeptide and a signal sequence (see <u>here</u> for more information).

Bottom line: one needs plenty of sequences to detect positive selection.

PAML – codeml – branch model



where to get help

read the manuals and help files check out the discussion boards at <u>http://www.rannala.org/phpBB2/</u>

else

there is a new program on the block called <u>hy-phy</u> (=hypothesis testing using phylogenetics).

The easiest is probably to run the analyses on the authors <u>datamonkey</u>.

Discussion: Other ways to detect positive selection? Selective sweep -> fewer alleles present in population Repeated episodes of positive selection -> high dN

If time discuss http://online.itp.ucsb.edu/online/infobio01/fitch1/

hy-phy

Results of an anaylsis using the SLAC approach



86-2.54472

-1.42483 0.151309

more output might still be <u>here</u>



Fig 1. Patterns of substitutions: Bars represent dN > dS (positive) or dN < dS (negative) in random samples of 148 - 150 sequences (A) and the whole dataset of 1312 viruses (B). Included in B are regions of mapped activity and 3D structures of the RNA-binding domain (RBD, panel I) [21] and Effector domain (ED, rotated to expose the 7 β -sheets (panel II) and 2 α -helices (panel II)) [7] with residues under negative (yellow/brown), neutral (gray) or positive (red) selection highlighted. Residues 208-230 not included in the 3D structure of the ED are disordered (compare with figure 5). Note sites with dN > dS map on the helix motifs of the ED or the linkers flanking them or the disordered region.

Hy-Phy ·

Hypothesis Testing using Phylogenies.

Using Batchfiles or GUI

Information at http://www.hyphy.org/

Selected analyses also can be performed online at http://www.datamonkey.org/



HYPHY PACKAGE

lypothesis Testing Using Phylogenies

🗯 Hy	P	hy File Ec	lit Anal	lysis Dat	a Likelih	ood Wind	dows	in the second	Contraction (A Station	tin the	-9 I	
O DataSet ns1_all_nt_8_sample													
_	•	460	470	480	490	500	518	1 520	530	540	550	568	
CY005738	тс	ACAGAAGACGGT	GCTATAGT	GGCTGAAATA	TOTOCTATIO	COTCOATGOO	AGGACATICI			AATTGGAATC	CTCATCGGTG	GACTIGAATGGAAT	GATA
AB256718	тc	ACAGAAGATGGT	GCTATTGT	GGCTGAAATT	TCTCCCATCC	CCTCTATGCC	AGGACATTCI	ACAGAGGATG	TCAAAAATGC	AATTGGAATC	CTCATCGGTG	GACTTGAATGGAAT	GATA
F597385	тс	ACAGAAGATGGT	GCTATTGT	GGCTGAAATT	TOTOCOATTO	COTCOATGOD	AGGACATTCI	ACAGAGGATG	TCAA <mark>A</mark> AATGC	AATTGGAATC	стебтеббтб	GACTTGAATG GAAT	GATA
\F523503	тс	ACAGAGGATGGT	GCTATTGT	GGCTGAAATT	TOTOCOATTO	COTCOATGOO	AGGACATTCI	FACAGAGGATG	TCAAAAATGC	AATTGGAATC	CTCATCGGTG	GACTTGAATG GAAT	GATA
CY005579	тс	ACAGAAGATGGT	GCTATTGT	GGCTGAAATT	TOTOCOATTO	CCTCCATGCC	AGGACATTCI	F <mark>ACAGAGGATG</mark>	TCAA <mark>A</mark> AATGC	AATTGGAATC	CTCATCGGTG	GACTTGAATG GAAT	GATA
CY005593	тс	ACAGAAGATGGT	GCTATTGT	GGCTGAAATT	TCTCCCATTC	CCTTCATGCC	AGGACATTCI	FACAGAGGATG	TCAA <mark>A</mark> AATGC	AATTGGAATC	CTCATCGGTG	GACTTGAATGGAAT	GATA
Y724259	тс	ACAGAAGATGGT	GCTATTGT	GGCTGAAATA	TCTCCCATTC	CCTCCATGCC	AGGACATTCI	ACAGAGGATG	TCAAAAATGC	AATTGGAATC	CTCATCGGCG	GACTTGAATGGAAT	GATA
AF144307	TC	ACAGAAAGTGGT	GCTATTGT	GGCTGAAATA	TTTCCCATTC	CCTCCGTACC	AGGACATTTI	ACAGAGGATG	TCAAAAATGC	AATTGGAATC	CTCATCGGTG	GACTTGAATGGAAT	GATA
AY028445	TC	ACAGAAAGTGGT	GCTATTGT	GGCTGAAATA	TCTCCCATTC	CCTCCGTACC	AGGACATTCI	FACAGAGGATG	TCAAAAATGC	AATTGGAATC	CTCATCGGTG	GACTTGAATGGAAT	GATA
2005773	TC TC	ACAGAAGAIGGT	GUTATIGT	GGUTGAAATA	TOTOCOATTO	CUTCUATGEC	AGGACATTC	ACAGAGGATG	TCAABAAIGC	AATIGGAATC	CICATOGGTG	GACTIGAATGGAAT	GATA
)Q251454	TU TO	ACGGAAGATGGT	GCTATIGI	GGUTGAAATA	TOTOCOATTO	COTCOATGOO	AGGACATIC		TCAABAATGC	AATTGGAATC	CTCATCGGTG	GACTIGAATGGAAT	GATA
17061119	TC TC	ACAGAAGAIGGI	GCTATIGI	GGCTGAAATA	TETECCATTE	CUTCUATGOD	AGGACATIC	LACAGAGGATG	TCLAMAATGC	AATTGGAATU	CTEATEGETE	GACIIGAAIGGAAI	GATA
41303044		ACAGAGGATGGT	GULATIGI	AGCIGAAAIA	TETEETATTE	CTICLAIGUE	AGGACATIC	LACAGAGGATG	TCLAMMAATGC	AATTGGAATC		GAULIGAALGGAAL	GATA
199400		ACAGATGACGGT	CCCATGGT	AGCIGAAAIA	TETECETTE	CTTCTATGCC	AGGACATIC	LACAGAGGATG	TCLAMMAATGC	AATTGGAATC	CTCATCOGIG	UACTICAATG GAAT	GATA
JODJO7 17407474	TC	ACAGACGATOGT	CCCATTOTA	ACCTCAAATT	TETECETTTE	CTTCTATOCC	ACCACATIC	LACAGAGGATG	TCAADAATGC	AATTCCAATC	CTCATCOGTO	CACTTCAATCCAAT	CATA
V2400E7	TC	ACTOMAGATOGC	CCCATTOTA	ACCTCAAATTA	TETECCATTE	CTTCCATCCC	ACCACATIC	LACAGAGGATG	TCAALAATCC	AATTCCAATC	CTCATCOGIG	CACTICAATCCAAT	CATA
V0019907	TC	ACAGACGATGGT	BCCATCOT	CCTCAAATA	TETECCATTE	CTTCTATGCC	ACCACATICI	LACAGAGGATO	TCAALAATGC	AATTCCAATC	CTCATCOCCO	CACTTCAATCCAAT	CATA
V004270	тс	ACAGATGATGGT	CCATCOL	ACCTGAAATA	TETECCATTE	CTTCTATGCC	AGGACATICI	TACAGAGGATG	TCAAAAATGC	AATTGGAATC	CTCATCOGTO	CACTTCAATCCAAT	GATA
V633280	тс	ACAGACGATGGA	GCCATTGT	AGCTGAAATA	TOTOCOATTO	CTTCTATGCC	AGGACATTCI		TCAAAAATGC	AATTGGAATC	CTCATCOGTO	GACTTGAATG GAAT	GATA
Y021137	τc	ACAGACGATGGC	GCCATTGT	AGCTGAAATA	TOTOCOATTO	CTTCTATGCC	AGGACATTCI		TCAAAAATGC	AATTGGAATC	CTCATCGGTG		GATA
Y016399	τc	ACAGACGATGGC	GCCATTGT	AGCTGAAATA	TOTOCOATTO	COTCOATGOO	AGGACATTCI	ACAGAGGATG	TCAAAAATGC	AATTGGAATC	CTCATCGGTG	GACTTGAATGGAAT	GATA
Y021201	τc	ACAGACGATGGC	GCCATTGT	AGCTGAAATA	TOTOCOATTO	COTCOATGOO	AGGACATTCI	ACAGAGGATG	TCAAAAATGC	AATTGGAATC	CTCATCOGTO	GACTTGAATGGAAT	GATA
0021633	тс	ACAGACGATGGC	ACCATTGT	AGCTGAAATA	TCTCCCATTC	CCTCCATGCC	AGGACATTCI	ACAGAGGATG	TCAAAAATGC	AATTGGAATC	CTCATCGGTG	GACTTGAATGGAAT	GATA
CY014691	тс	ACAGATGATGC	GCCATTGT	AGCTGAAATA	TCTCCCATTC	CCTCTATGCC	AGGACATTCI	ACAGAGGATG	TCAAAAATGC	AATTGGAATC	CTCATCGGTG	GATTTGAATG GAAT	GATA
002105	тс	ACAGATGATGGC	GCCATTGT	AGCTGAAATA	TCTCCCATTC	CCTCTATGCC	AGGACATTCI	ACAGAGGATG	TCAAAAATGC	AATTGGAATC	CTCATCGGTG	GACTTGAATGGAAT	GATA
CY004773	тс	ACAGATGATGGC	GCCATTGT	AGCTGAAATA	TCTCCCATTC	CCTCTATGCC	AGGACATTCI	ACAGAGGATG	TCAAAAATGC	AATTGGAATC	CTCATCGGTG	GACTTGAATGGAAT	GATA
CY014972	тс	ACAGATGATGC	GCCATTGT	АССТБАААТА	TOTOCOATTO	COTOTATGOO	AGGACATTCI	ACAGAGGATG	TCAAAAATGC	AATTGGAATC	CTCATCGGTG	GACTTGAATG GAAT	GATA
FOR (ORO									TOUL				<u> </u>
≫ ⊕	0	Partition Name	Partition	Type Tree T	opology 💌	Substitution	wodel 🖭 P	arameters 🗵	Equilibrium	Freqs. 🖭 I	Rate Classes		
	0	END	Codon	Tree_1		MG94xTN93_	,3x4 🔽 🖸	Flobal 🔽	Partition				
		BEGINNING	Codon	Thee 1	2 🔍	MG94xTN93	3x4 🗔 o	lobal 💌	Partition				
123 🖹	-	22011110							- in the state				
<u>.</u>													
ШШ (ci ⁱⁱ													
A Nuclea	ti d -	Data 690 cite - (40	2 distinct	#app c) 150	sias Current S	Jaction 525 F2							
mucleonde Data, 650 sites (405 distinct patterns), 150 species. Lurrent selection:525-525													

Set up two partitions, define model for each, optimize likelihood

😁 🔿 Likelihood parameters for ns1_all_nt							
ቩ 🗮 🔍 🛄 🔕 🔄 🦉 🛛 Current LF							
	Parameter ID	◄	Yalue	Constra	int		
7	Tree_1						
P	Tree_12						
R	BEGINNING_Shared_AC		0.187074		-		
R	BEGINNING_Shared_CT		0.828132				
R	BEGINNING_Shared_R		0.127845	-			
R	END_Shared_AC		0.117325				
R	END_Shared_CT		0.93391				
R	END_Shared_R		0.946316	-			
θ	Tree_1.AB256718.synRate		0.12461				
θ	Tree_1.AF001672.synRate		0.016737				
θ	Tree_1.AF009898.synRate		0				
θ	Tree_1.AF055424.synRate		0.017357				
θ	Tree_1.AF074267.synRate		0				
θ	Tree_1.AF074279.synRate		0.0527182				
θ	Tree_1.AF084286.synRate		0.0176037				
θ	Tree_1.AF144307.synRate		0.0528252				
θ	Tree_1.AF256183.synRate		0				
θ	Tree_1.AF256188.synRate		0.0174124				
θ	Tree_1.AF523503.synRate		0.0527042				
θ	Tree_1.AJ344036.synRate		0				
θ	Tree_1.AJ410594.synRate		0.0350104				
θ	Tree_1.AJ410598.synRate		0.0174538				
θ	Tree_1.AM502792.synRate		0.0174516				
0	THE AMERICAN THERE		0				

Save Likelihood Function then select as alternative

The dN/dS ratios for the two partitions are different.



(to do, select both rows and then click the define as equal button on top)

Example testing for dN/dS in two partitions of the data --John's dataset DataSet ns1_all_nt_8_

	\mathcal{O}				Dat	aSet ns1_all_nt_8_
		0	Clikelihood parameter	eters for ns1	L_all_nt	
	-	ր	≒ 🔍 🕕 🐧 🤄 🖓 🖓 Cur	rent LF	A T	
			Parameter ID 💌	Yalue	Constraint	500 510
	738 T (718 T (?	Tree_1		0	TGCCAGGACATTCTAC
:	385 T	j 📍	Tree_12			TGCCAGGACATTCTA
	503 <mark>T</mark> (R	BEGINNING_Shared_AC	0.187238		TGCCAGGACATTCTAG
	579 T(R	BEGINNING_Shared_CT	0.891995		TGCCAGGACATTCTAG
	259 T	R	BEGINNING_Shared_R	0.19809		TGCCAGGACATTCTAC
	307 T	R	END_Shared_AC	0.126137		TACCAGGACATTTTAC
	145 T (R	END_Shared_CT	0.770683		TACCAGGACATTCTAC
		R	END_Shared_R	0.19809	BEGINNIN	TGCCAGGACATTCTAG
	119 T	θ	Tree_1.AB256718.synRate	0.309711		TGCCAGGACATTCTAC
(544 <mark>T</mark> (θ	Tree_1.AF001672.synRate	0.0364501		TGCCAGGACATTCTAG
	3 T(θ	Tree_1.AF009898.synRate	0		TGCC AGGACATTCT AC
	7 171 T	θ	Tree_1.AF055424.synRate	0.0414451		TGCCAGGACATTCTA
	957 T	θ	Tree_1.AF074267.synRate	0		TGCCAGGACATTCTAG
1	270 <mark>T</mark> (θ	Tree_1.AF074279.synRate	0.131262		TGCCAGGACATTCTAC
	370 T (280 T (θ	Tree_1.AF084286.synRate	0.0419524		TGCC AGGACATTCTAC
	137 T	θ	Tree_1.AF144307.synRate	0.129191		TGCCAGGACATTCTAC
(399 <mark>T</mark> (θ	Tree_1.AF256183.synRate	0		TGCCAGGACATTCTAG
	201 T	θ	Tree_1.AF256188 svnRate	0.0415364		TECCAGGACATTCTAG
	533 I 591 TI	θ	Tree_1.AF523 🔵 🔵 🔘	НҮРН	Y Console	AL
	5 T(θ	Tree_1.AJ34475:0,(CY007231	:0.00214899,		
	773 <mark>T</mark> (θ	Tree_1.AJ410	Y000085:0.0021	4801)Node85:0	, T AC
	972 T	θ	Tree_1.AJ410 Input			A(
	u 🔽	θ	Tree_1.AM502			I 1
	-	θ	Tree_1.AM502			Do 1
		θ	Tree_1.AY028			de la
	4	θ	Tree_1.AY210		CL DOWN	07.77.41
		A I	Tree_1.AY241641.svnkate	10		05:57:11 🔰 🖉





After selecting LRT (= Likelihood Ratio test), the console displays the result, i.e., the beginning and end of the sequence alignment have significantly different dN/dS ratios.

Alternatively, especially if the the two models are not nested, one can set up two different windows with the same dataset:

000	0							D	ataSet ns1	1_all_nt_8_sa	mple_finished	
	<i>.</i>											
	530	1	540	550	560	570	3 580	590	68	30 610	620	
CY005738 AB256718				CTCATCGGTG G	ACTTGAATG	GAATGATAA	CTCAATTCGAG	DataSet n	TATACAGAG s1 all nt 8	A <mark>TTCGCTT</mark> GGG 8 sample	GAATCTGTGA TGAGA	AT.
EF597385			Ē							_ ,		-
AF523503												
CY005593			Ĺ	460	470	480	490	500	51	.0 520	530	E
AY724259		CY005738	тс	ACAGAAGACG G1	гостатасто	GCTGAAATA	TCTCCTATTC	CCTCCATGCC	AGGACATTO	TACAGAGGATG	TCAAAAATGC AATTG	GAA
AF144307		AB256718	тс	ACAGAAGATG G1	гостаттот о	GCTGAAA <mark>T</mark> T	TCTCCCATCC	CCTCTATGCC	AGGACATTO	TACAGAGGATG	TCAAAAATGCAATTG	GAA
AY028445 CV005773		EF597385	TC	ACAGAAGATG G1	IGCTATTGT	GCTGAAATT	TCTCCCATTC	CCTCCATGCC	AGGACATTO	TACAGAGGATG	TCAAAAATGC AATTG	GAA
D0251454		AF523503		ACAGAGGATGGT ACAGAAGATGGT	TGCTATIGIG	GCTGAAATI	TETECCATTC	COTCOATGOD	AGGACATIC	TACAGAGGATG		GAA
EF061119		CY005593	τc	ACAGAAGATGGI	IGCTATTGT	GCTGAAATT	TCTCCCATTC	COTTCATGOO	AGGACATTC	TACAGAGGATG	TCAAAAATGCAATTG	GAA
AY303644			•	Partition Name	Partition	Type Tree	Topology 🔽	Substitution	Model	Parameters 🔻	Equilibrium Freas	▼
M55468		<u>≫</u> +		FND	Codon	Tree		MC94-TN93	3v4	Global 🔽	Partition	
AY497171		- 🕱	-	DECEMBER	Couon			MO94X11495_				– Model 1
AY619957		123 🖹	9	BEGINNING	Codon		12 🖸	MG94XIN93_	3X4 🔟		Partition	
CY004270		8. Q										
-CY004370												
A1633260 CV021137												
CY016399		🔵 Nucleo	tide	Data. 690 sites (4	03 distinct pat	terns), 150 sp	ecies. Current S	election:479-47	9			
CY021201 DQ021633		AATTGGA. AATTGGA.		CTCATCGGTG G. CTCATCGGTG G.	ACTTGAATG ACTTGAATG		C <mark>TCAATTC</mark> GAG CTCAATTCGAG	; <mark>сотст</mark> балал ; <mark>сотстбалал</mark>	TATACAGAG TATACAGAG	G <mark>A TTCGCTT</mark> GGG G <mark>A TTCGCTT</mark> GGG	G <mark>aatee</mark> gtga tgagg G <mark>aatee</mark> gtga tgagg	AT. AT
	🚺 Pa	rtition Na	ne	Partition Type	Tree Topol	logy 🔽 Su	bstitution Mo	del 🔽 Paran	neters 💌 I	Equilibrium Fre	eqs. 🔽 Rate Classes	s
	🔵 ns	1 all nt 8	saj	Codon 💌	nslallnt (8 san 🔁 M	394xTN93_3x4	- 🔽 Local	I I	artition		🗕 Model 2
- 33												
123 🖹												
÷.												
🔘 Nucleo	tide Dat	ta. 690 sites	(40)	3 distinct patterns)), 150 species.	Current Selec	tion:1-690					

Simulation under model 1, evalutation under model 2, calculate LR Compare real LR to distribution from simulated LR values. The result might look something like this or this



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 Programm and NASA Applied Information Systems Research Program

Workshop at Te Aviv University, November 29th, 2009.



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.

Detecting Errors in Genome Annotation

Analysis of 8 strains of Escherichia coli

Number of families with 1 missing gene

An ortholog from a family with 1 missing gene

Escherichia coli 536	56
Escherichia coli APEC_O1	196
Escherichia coli CFT073	45
Escherichia coli K12	4
Escherichia coli O157H7	33
Escherichia coli O157H7 EDL933	6
Escherichia coli UTI89	20
Escherichia coli W3110	8
Total:	368

Example of missed ORFs



hp - hypothetical protein

 $\epsilon, \beta, \gamma, \alpha, \delta, B, C, A, I - ATP$ synthase subunits

Analysis of 368 missing orthologs
with blastnAnalysis of 297 hits with > 90% identity
in genomes with a missing gene

Each hit was analyzed and classified as it is depicted on plates (b),(c) and (d).



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



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

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





1 gene family

0 gene family

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







Root positions



Comparison of the best BLAST hit method and BranchClust algorithm

Number of taxa -	Number of selected families:					
A: Archaea B: Bacteria	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).				



ATP-synthases: Examples of Clustering

13 gamma proteobacteria



30 taxa: 16 bacteria and 14 archaea



317 bacteria and archaea



www.bioinformatics.org/branchclust

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



www.bioinformatics.org/branchclust
BranchClust Algorithm

Data Flow



www.bioinformatics.org/branchclust

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.



- Thermotoga petrophila
- Thermotoga maritima
- Thermotoga sp. strain RQ2
- Thermotoga neapolitana
- Thermotoga naphthophila





Olga Zhaxybayeva, Kristen S. Swithers, Pascal Lapierre, Gregory P. Fournier, Derek M. Bickhart, Robert T. DeBoy, Karen E. Nelson, Camilla L. Nesbø, W. Ford Doolittle, J. Peter Gogarten, and Kenneth M. Noll-

"On the Chimeric Nature, Thermophilic Origin and Phylogenetic Placement of the Thermotogales", *Proc Natl Acad Sci U S A.*, Online Early, March 23, 2009.

to use other genomes:

- The easiest source for other genomes is via anonymous ftp from <u>ftp.ncbi.nlm.nih.gov</u> 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

to copy files and scripts into your folder

- mkdir workshop
- cd workshop
- mkdir test
- cp -R /Users/jpgogarten/workshop/test/
 - * /Users/mcb221_unnn/workshop/test/

This should be one line, and mcb221_u1nnn should be replaced with the name of your home directory.

The –R tells UNIX to copy recursively (including subdirectories) This command also copies a directory called fasta that contains 5 genomes to work on. If you want to work on different genomes, delete the 5 *.faa files that contain the genomes from the Thermotogales and replace them with the genomes of your choice. ("genomes" really means all the proteins encoded by ORFs present in the genome). 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 a four letters designating the species folloed 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 \bigcirc

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. (Best is probably ~/workshop/test.)

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.... 4568.... Enterobacteria phage T1 1757.... Enterobacteria phage T3 Enterobacteria phage T5 4640.... Enterobacteria phage T7 9627... Kluyvera phage Kvp1 2126 Lactobacillus phage phiAT3 4869.... Lactobacillus prophage Lj965 4117.... Lactococcus phage r1t 2345.... Lactococcus phage sk1 9629... 193434.. Mycobacterium phage Bxz2 | 29566...

the branchclust scripts

- are available at http://www.bioinformatics.org/branchclust/
- A copy of the tutorial is in the folder you copied into your folder: BranchClustTutorial.pdf Consult the tutorial, if you want to use branchclust on other genomes.
- The commands we use today are in a file in the test folder called commands workshop tau one script



This is a text file that you can open with any text editor.

(I use textwrangler on my mac, but you might want to use crimson)

BranchClust Article

• is available at

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

BMC Bioinformatic	We 3.78 Subject areas advanced search authors reviewers libraries about my BioMed Central	I come J Peter Gogarten (Log off) Feedback Support My details	
Тор	Methodology article Open Access	BMC Bioinformatics Volume 8	
Abstract Background	BranchClust: a phylogenetic algorithm for selecting gene families	Viewing options: Abstract	
Results Discussion	Maria S Poptsova 🔀 and J Peter Gogarten 🖂 Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269-3125, USA	 PDF (1.1MB) Additional files 	
Conclusion	author email corresponding author email BMC Bioinformatics 2007, 8:120 doi:10.1186/1471-2105-8-120	Associated material: = Readers' comments S = PubMed record	
Methods Availability and requirements Authors' contributions Acknowledgements References	The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2105/8/120 Received: 8 December 2006 Accepted: 10 April 2007 Published: 10 April 2007 © 2007 Poptsova and Gogarten; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.	 Related literature: Articles citing this article on Google Scholar on ISI Web of Science on PubMed Central Other articles by authors ⊕on Google Scholar ⊕on PubMed Related articles/pages on Google on Google Scholar on PubMed 	

Create super families, alignments and trees

vi do_blast.pl
to see what the parameters are doing type blastall or
bastall | more at the commandline.
If you move this to a different computer you might need to change a 2 to
a 1

vi parse_blast_cutoff_thermotoga.pl
change bioperl directory; change cutoff E-value
the script as written uses the bioperl library in my home directory
Note: if using closely related genomes, you can cut back on the
size of the superfamilies by using a smaller E-value
(if you genomes have normal GI numbers, use
vi parse_blast_cutoff1.pl)

check output: more parsed/all_vs_all.parsed ### type q to leave more more parsed/all_vs_all.parsed | wc -l # checks for number of lines=super families output

Super Families to Trees

- perl parse_superfamilies_singlelink.pl 1
 # 1 gives the minimum size of the superfamily
- perl prepare_fa_thermotoga.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 calcualtes trees using Kimura distances for all families in fa #trees stored in trees Check #1, 106, 1027, 111
- perl prepare_trees.pl reformats trees

Branchclust

perl branchclust_all_thermotoga.pl 2
Parameter 2 (MANY) says that a family needs to have
at least 2 members.

make_clusterlist.sh
runs perl make_fam_list_inpar.pl 5 4 0
results in test called families_inpar_5_4_0.list
5: number of genomes;
4: number of genomes in cluster;
0: number of inparalogs
(a 1 returns all the families with exactly 1 inparalog)
you could add additional lines to the shell script:
perl make fam list inpar.pl 5 4 1

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_thermotoga.pl families_inpar_5_4_0.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 Text Wrangler

00)			detailed_summary.out
×	Τ	ļ	i	Last Saved: 11/24/09 10:28:05 AM File Path + : /Volumes/Users/mcb221_u4/op/test/detailed_summary.out
< >	🗋 d	letaile	d_su	mmary.out 🗘 🛄 💭
superfo	amily	/_##	fam	nily_## nu_of_genes_in_the_family nu_of_paralogs family_name
1 1	5	0	TRQ	2_0001 Chromosomal replication initiator protein DnaA
10 1	5	0	TRQ	2_0010 Probable low-affinity inorganic phosphate transporter
100 1	5	0	TRQ	22_0106 transcriptional regulator, TetR family
1000	1	5	0	TRQ2_1263 hypothetical protein
1001	1	5	0	TRQ2_1262 RNA binding methyltransferase FtsJ like
1002	1	5	0	TRQ2_1261 hypothetical protein
1003	1	5	0	TRQ2_1260 tRNA pseudouridine synthase A (EC 4.2.1.70)
1004	1	5	0	TRQ2_1259 protein kinase
1005	1	5	0	TRQ2_1258 Signal peptidase I (EC 3.4.21.89)
1006	1	5	0	TRQ2_1257 LSU ribosomal protein L19p
1007	1	5	0	TRQ2_1256 protein of unknown function aq_054
1008	1	5	0	<pre>TRQ2_1255 tRNA (Guanine37-N1) -methyltransferase (EC 2.1.1.31)</pre>
1009	1	5	0	TRQ2_1254 16S rRNA processing protein RimM
101 1	5	0	TRQ	2_0107 Phenylalanyl-tRNA synthetase beta chain (EC 6.1.1.20)
1010	1	5	0	TRQ2_1253 KH domain RNA binding protein YlqC
1011	1	5	0	TRQ2_1252 SSU ribosomal protein S16p
1012	1	5	0	TRQ2_1250 Acylphosphate phosphohydrolase (EC 3.6.1.7), putative
1013	1	5	0	TRQ2_1249 MscS Mechanosensitive ion channel
1014	1	5	0	TRQ2_1248 hypothetical protein
1015	1	5	0	TRQ2_1247 tRNA-guanine transglycosylase (EC 2.4.2.29)
1016	1	5	0	TRQ2_1246 Formiminotetrahydrofolate cyclodeaminase (EC 4.3.1.4)
1017	1	5	0	TRQ2_1245 Deoxyribose-phosphate aldolase (EC 4.1.2.4)
1018	1	5	0	TRQ2_1244 metal dependent phosphohydrolase
1019	1	5	0	TRQ2_1243 DNA repair protein RadC

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.

		ily_	he_fa	nu_of_p						
superfamily_	##	##	mily	aralogs	ifamily_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_0766 ABC transporter related					
	129	56	4	0	Tpet_0504 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_0524 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 synt					
					•					

clusters/clusters_NNN.out.names

- Check a superfamily of your choice. Within a family, 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 ------
>lcllTnea_1049 ABC transporter related [Thermotoga neapolitana]
>lclITRQ2_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]
>lcllTnap_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

- Check the tree for a superfamily of your choice. Copy the file to your computer and open it in TreeView, NJPLOT, or FigTree (check with your neighbor on which program works).
- For at least one cluster, in the tree, check if branchclust came to the same conclusion you would have reached



prepare_bcfam_thermotoga.pl families_inpar_5_4_0.list

The script prepare_bcfam_thermotoga.pl takes a list of families (created by make_fam_list_inpar.pl) and for each family retrieves the fasta sequences from the combined genome databank and stores the sequences in the BCfam folder, one multiple sequence file per family.

One possibility for further evaluation is to take multiple sequence files, align the sequences and perform a phylogenetic reconstruction (including boostrap analysis) using programs like phyml or Raxml.

The resulting trees can be analyzed by decomposition and supertree approaches.

Decomposition of Phylogenetic Data



Phylogenetic information present in genomes Break information into small quanta of information (bipartitions or embedded quartets) Analyze spectra to detect transferred genes and plurality consensus.

TOOLS TO ANALYZE PHYLOGENETIC INFORMATION FROM MULTIPLE GENES IN GENOMES:

Bipartition Spectra (Lento Plots)

BIPARTITION OF A PHYLOGENETIC TREE

Bipartition (or split) – a division of a phylogenetic tree into two parts that are connected by a single branch. It divides a dataset into two groups, but it does not consider the relationships within each of the two groups.





Yellow vs Rest * * * . . * * compatible to illustrated bipartition

Orange vs Rest • • * • • • * incompatible to illustrated bipartition

"Lento"-plot of 34 supported bipartitions (out of 4082 possible)

13 gammaproteobacterial genomes (258 putative orthologs):

E.coli
Buchnera
Haemophilus
Pasteurella
Salmonella
Yersinia pestis (2 strains)

- •Vibrio
- •Xanthomonas (2 sp.)
- Pseudomonas
- Wigglesworthia

There are 13,749,310,575 possible unrooted tree topologies for 13 genomes





PROBLEMS WITH BIPARTITIONS (CONT.)



Decay of bipartition support with number of OTUs



Phylogenies used for simulation

Example for decay of bipartition support with number of OTUs



Only branches with better than 70% bootstrap support are shown

Decay of bipartition support with number of OTUs



Each value is the average of 10 simulations using seq-gen. Simulated sequences were evaluated using PHYML. Model for simulation and evaluation WAG + $\Gamma(\alpha=1, 4 \text{ rate categories})$

Bipartition Paradox:

- The more sequences are added, the lower the support for bipartitions that include all sequences. The more data one uses, the lower the bootstrap support values become.
- This paradox disappears when only embedded splits for 4 sequences are considered.

TOOLS TO ANALYZE PHYLOGENETIC INFORMATION FROM MULTIPLE GENES IN GENOMES:

QUARTET DECOMPOSITION

Bootstrap support values for embedded quartets



: tree calculated from one pseudosample generated by bootstraping from an alignment of one gene family present in 11 genomes

: embedded quartet for genomes 1, 4, 9, and 10. This bootstrap sample supports the

topology ((1,4),9,10).







Quartet spectral analyses of genomes iterates over three loops:

➢ Repeat for all bootstrap samples.

> Repeat for all possible embedded quartets.

➢ Repeat for all gene families.

QUARTET DECOMPOSITION METHOD



- Quartet is a smallest unit of phylogenetic information
- Each quartet is associated with only three unrooted tree topologies
- Support for different quartet topologies can be summarized for all gene families



Illustration of one component of a quartet spectral

analyses Summary of phylogenetic information for one genome quartet for all gene families





Other POSITIVE THOUGHTS ABOUT THE METHOD

- 1. No assumption that all genes in a genome have the same phylogenetic history.
- 2. The total number of quartets is much smaller than number of tree topologies, which makes it possible to evaluate all quartets.
- 3. Gene families present only in few analyzed genomes can be included in the analyses
- 4. Phylogenetic signal can be divided into consensus supported by the plurality of gene families and the conflicting signal.
- Allows us to partition analyzed genomes according to some scenario (e.g., grouping by ecology) and retrieve gene families that support or conflict it.



Quartet decomposition analysis of 19 *Prochlorococcus* and marine *Synechococcus* genomes. Quartets with a very short internal branch or very long external branches as well those resolved by less than 30% of gene families were excluded from the analyses to minimize artifacts of phylogenetic reconstruction.




Figure 8. Distribution of gene families without paralogs across functional categories. The four super-categories are defined by COG database. Notably, genes of informational storage and processing are represented in equal proportions in genes in conflict with plurality as compared to all 1812 gene families, which contradicts complexity hypothesis {Jain, 1999 #31}. Metabolic genes appear to be overrepresented in the gene family pool which conflicts with plurality.





Plurality consensus calculated as supertree (MRP) from quartets in the plurality topology.

NeighborNet (calculated with SplitsTree 4.0)



Plurality neighbor-net calculated as supertree (from the MRP matrix using SplitsTree 4.0) from all quartets significantly supported by all individual gene families (1812) without in-paralogs.

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:

			(((T
	Name		(((Tpet:0.000000476
	DCIam_245.1.5.1.pny_pnymi_bool_trees.txt		(((Tnap:0.00000000
	bcfam 244.1.5.0.phy phyml boot trees.txt		((Tmar:0.000000000
	hofam 245 1 5 0 phy phyml boot trees tyt		(((TRQ2:0.00000000
	berain_245.1.5.0.phy_phym_boot_rees.txt		((Tmar:0.0095906414
ET.	bcfam_246.1.5.0.pny_pnyml_boot_trees.txt		(((TRQ2:0.00000000
E trees ain	bcfam_247.1.5.0.phy_phyml_boot_trees.txt		(((Inap:0.00000000000000000000000000000000000
trees.zip	bcfam_248.1.5.0.phy_phyml_boot_trees.txt		(((Tnea:0.148092574
	bcfam 248.2.5.0.phy phyml boot trees.txt		((Tmar:0.0149409628
	hefam 249 1 5 0 phy phyml boot trees tyt		(((TRQ2:0.00000000
	berain_249.1.5.0.phy_phynn_boot_rees.txt		(((TRQ2:0.000000000
	bcfam_250.1.5.0.phy_phyml_boot_trees.txt		(((Thea:0.134582349
l.	bcfam_251.1.5.0.phy_phyml_boot_trees.txt		((Tmar:0.0096275515
V	bcfam 252,1,5,0 nby nbyml boot trees txt		(((Tpet:0.0000001222
	hefen 252 1 5 0 phy phymic boot trees tot		(((Tpet:0.00000057
	DCram_253.1.5.0.pny_pnymi_boot_trees.txt	*	(((Thap:0.00000000000000000000000000000000000
trees	bcfam_254.1.5.0.phy_phyml_boot_trees.txt		(((Thap:0.000000000
	bcfam_255.1.5.0.phy_phyml_boot_trees.txt		(((Tpet:0.014654976
	bcfam 256.1.5.0 pby pbyml boot trees txt		(((Tpet:0.022190558 (((Tpeg:0.18360172)
	befam_25011510.phy_phym_boot_recestor		(((Tnap:0.00000000
	DCram_257.1.5.0.pny_pnymi_boot_trees.txt		((Tmar:0.0046056775 (((Tmar:0.019346736
	bcfam_258.1.5.0.phy_phyml_boot_trees.txt		(((Tnap:0.00000000
	bcfam_259.1.5.0.phy_phyml_boot_trees.txt		(((TRQ2:0.00000000000000000000000000000000000
	bcfam_260.1.5.0.phy_phyml_boot_trees.txt		(((Tnea:0.165345744
	bcfam 261.1.5.0.phy phyml boot trees.txt		(((Tpet:0.003874468
	hcfam 262 1 5 0 phy phyml boot trees txt		(((Tmar:0.022391339
	hefam 262 1 5 0 phy phyml boot trees tyt		(((TRQ2:0.00000000000000000000000000000000000
	bciani_205.1.5.0.phy_phymi_boot_trees.txt		((Tmar:0.0095751539
	bcfam_264.1.5.0.phy_phyml_boot_trees.txt		(((TRQ2:0.000000000
	bcfam_265.1.5.0.phy_phyml_boot_trees.txt		(((Tpet:0.01109710) (((Tpap:0.00000000
	bcfam_266.1.5.0.phy_phyml_boot_trees.txt		(((TRQ2:0.000000000
	bcfam 267.1.5.0.phy phyml boot trees.txt		((Tmar:0.014215572
	hcfam 268 1 5 0 phy phyml boot trees tyt		(((Thap:0.00000000
	beram_200.1.5.0.phy_phymi_boot_trees.txt		
(Joo file por family		400
	Jne lile per lamily	\ \	()(
			101

-0	(((1pet:0.0219514318, Inea:0.1960236242):0.0145181752, Imar:0.0189973964):0.0155785587, Inap:0.0000000001, IRQ	2:0.0000000001);	
((((Tpet:0.0000004769,Tnea:0.1773430420):0.0205769649,Tmar:0.0047117206):0.0416898504,Tnap:0.0000000001,TRQ	2:0.0000000001);	
((((Tnap:0.000000001,TRQ2:0.0000000001):0.0108598711,Tmar:0.0033438192):0.0267134063,Tpet:0.0234486505,Tne	a:0.1723830952);	
Ó	(((Tpet:0.0000003788, Tnea:0.2436919374):0.0254977400, Tmar:0.0297753480):0.0149123962, TRQ2:0.0000000001, Tna	p:0.0000000001);	
(((Tmar:0.0000000001,(Tpet:0.0094293384,Thea:0.1693653990):0.0000007832):0.0092730739,TRQ2:0.0000000001,Tha	p:0.0000000001);	
Ċ	((TRQ2:0.0000000001,Tnap:0.0000000001):0.0143579956,Tmar:0.0094859692):0.0201011543,Tpet:0.0090235931,Tne	a:0.1566628790);	
ò	((Tmar:0.0095593366.(Thea:0.1477968521.Tpet:0.0139766236):0.0208932007):0.0243108915.TR02:0.0000000001.Tha	p:0.0000000001);	
ì	((Tmgr:0.0095906414.)(Theg:0.1277588182.Thet:0.0100333825):0.0043365913):0.0095754957.Thgb:0.000000001.TRU	2:0.0000000001):	
ì	((TRD2:0.000000001), Trap:0.000000001):0.0298828683, Tmar:0.0043702969):0.0354610321, Trag:0.1300784273, Trag	t:0.0157791361):	
Ż	(() Toon : 0.000000001 . TRD:: 0.000000001): 0.0152181584 . Two:: 0.0000782629): 0.0241034085 . Thet: 0.0118502181 . The	a:0.1613724681);	
2	(Tmor: 0.0185827480, (TR02:0.00000001, Tmor: 0.00000001):0.0045215011):0.0300461858, Thet: 0.00000001814, The	a:0.1049097183);	
2	///Teen+0 1480925746 Teet+0 0093957878 0 0101463099 Tear+0 0138185337 0 0191391368 Tear+0 000000001 TRU	2.0 00000000001)	
2	((Tmm:-0.094453911 (Tmm:-0.189592037 Tmm:-0.08090653):0.0247328511):0.023938188 TDD:-0.08090808091 Tmm	n 0 0000000000000000000000000000000000	
2	((Timor -0.0144406-28 (Timor -0.1340376373 Timet -0.00000184)+0.0151620300)+0.022252431 TDD2+0.00000001 Timor	n•0 00000000001);	
Σ	((mar.:e.g.2011)05220 (Thomas Beereaged TDD): Beereaged): 8 846524582 (1971) 8 826242424242 (1971)	0.1240102200\.	
Σ	((/min.s.21190725)(mip.0.000000001)(KQ.0.0000001)(0.01001007)(0.025577200),ptc.0.00721375)(mip.0.00721375)		
X	(((ThQ2.6.50000000001, ThQ1.6.500000001).6.5102.0011, ThQ1.6.511512.1077).6.5011012.1157, TPC.6.501770000, TPC	A.0.1200700000),	
S	((()RQ210.00000000000), (RD210.00000000001):0.0245/42140, (RD210.0226/15955):0.0420494142, (RE10.142000.1379, (RE		
5	((init: 1:0000022210,(i)e::0:000000200,i)e:://(0.0000000001,i)iii	5:0.0000000001);	
Ş	(((Induid:1343623747), het is .000000011); ().010023007, ().0012, (0.0097662079); ().0254001207, (Ndt: 0.000000001), ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().010023007, ().01002300		
Ş	(((Imar:0.00962/5515)()(Imap:0.0000000001)(RQ:0.00000001)(0.014521101)(0.0144699620)(pet:0.00000000207)(net:0.00000000207)	1:0.2040000733);	
Ş	((((hidp:0.00000000001),RQ2:0.000000001):0.0191004000, hidp:0.000000470/):0.01949000005, http://doi.0000000001, (((t-1)-0.000000000001), hidp:0.00000001), 0.0191004000, hidp:0.000000470/):0.01949000005, http://doi.0000000001,	1:0.1105392729);	
5	(((()pet:0.0000012224, ineu:0.10023)505):0.014022353, inut:0.004030122):0.019/363319, iRut: 0.00000000001, inu	0:0.000000001);	
S	((((be:10.0000005//6,1hed:0.161310131):0.0151309/6/,1md:0.0340401/9/):0.024529/526,1RQ2:0.0000000001,1md	5:0.000000001);	
S	(((Indp:0.0000000001, NRU2:0.00000000001):0.0200502052, Imdr:0.0029415246):0.0395338330, Ined:0.1363020193, Ipe	t:0.00/305/945);	
S	(((indp:0.00000000001, Rd2:0.0000000001):0.014446/012, Imdr:0.0240468519):0.0243429197, Ined:0.11/1008552, ipe	t:0.0000005030);	
S	((((hap:0.0000000001, RQ2:0.000000001):0.0142730043, har:0.0190927424):0.0287749906, hea:0.1403425968, jpe	t:0.0000011409);	
S	((((hap:0.0000000001,RQ2:0.000000001):0.0187720004, har:0.0140524520):0.0255993636, pet:0.000001838, har	3:0.11805/8369);	
Ş	(((1pt:0.0146549708,1hea:0.1369483644):0.0045963442,1mar:0.0048168165):0.0192544944,1hap:0.00000000001,1RQ	2:0.0000000001);	
Ş	(((1pet:0.0221905588, 1hed:0.140106/289):0.02386/1124, 1mar:0.02/8952214):0.0155495976, 1hap:0.00000000001, 1RU	2:0.000000001);	
Ş	(((lnea:0.1836017274,lpet:0.0009186098):0.0045584496,lmar:0.0096100807):0.0289491135,lRQ2:0.00000000001,lnq	p:0.0000000001);	
5	((((Inap:0.00000001, RQ2:0.000000001):0.0245340382, Imar:0.0009695788):0.0455521078, Ipet:0.0000004267, Ine	1:0.1278836141);	
9	((Tmar:0.0046056775,(TRQ2:0.0000000001,Tnap:0.0000000001):0.0338204105):0.0353863970,Tpet:0.0036576284,Tnev	1:0.1441086474);	
9	(((Tmar:0.0193467388,Thea:0.1811163023):0.0055734283,Tpet:0.0100979429):0.0412839758,Thap:0.0000000001,TRQ	2:0.0000000001);	
9	(((Tnap:0.0000000001,TRQ2:0.0000000001):0.0197191027,Tmar:0.0145678338):0.0254168418,Thea:0.1301777517,Tpe	t:0.0000005693);	
0	(((TRQ2:0.0000000001,Tnap:0.000000001):0.0046957659,Tmar:0.0095182632):0.0122300380,Tpet:0.0068760148,Tne	1:0.119415445 3);	
9	(((Tnea:0.1322456802,Tpet:0.0000007926):0.0210521767,Tmar:0.0089703119):0.0394904771,TRQ2:0.0000000001,Tnq	p:0.0000000001);	
9	(((Tnea:0.1653457442,Tpet:0.0000004733):0.0102936135,Tmar:0.0096887422):0.0299912309,TRQ2:0.0000000001,Tnq	p:0.0000000001);	
((((TRQ2:0.0000000001,Tnap:0.000000001):0.0298054517,Tmar:0.0189746383):0.0360451008,Tnea:0.1084748253,Tpe	t:0.0051247348);	
((((Tpet:0.0038744681,Thea:0.1320467813):0.0173508134,Tmar:0.0048401153):0.0252014442,Thap:0.0000000001,TRQ	2:0.0000000001);	
((((Tnap:0.0000000001,TRQ2:0.0000000001):0.0149831102,Tmar:0.0139883251):0.0414283364,Tnea:0.0926333864,Tpe	t:0.0091967570);	
((((Tmar:0.0223913390,Thea:0.1298605798):0.0023660265,Tpet:0.00000004950):0.0197163698,TRQ2:0.0000000001,Tha	p:0.0000000001);	
((((TRQ2:0.0000000001,Tnap:0.0000000001):0.0194130671,Tmar:0.0141043556):0.0359273406,Tpet:0.0000016980,Tne	a:0.1201428880);	
((((Tpet:0.0047751464, Tnea:0.0808394153):0.0148804712, Tmar:0.0194291165):0.0197516807, Tnap:0.0000000001, TRQ:	2:0.0000000001);	
(((Tmar:0.0095751539,(TRQ2:0.0000000001,Tnap:0.0000000001):0.0143794898):0.0253853377,Tnea:0.1300017287,Tpe	t:0.0089710270);	
(((Tmar:0.0046448823,(Tmap:0.0000000001,TRQ2:0.0000000001):0.0145732561):0.0444331749,Tpet:0.0066190596,Tme	a:0.1768104694);	
((((TRQ2:0.0000000001,Tnap:0.0000000001):0.0098668532,Tmar:0.0193698319):0.0200257831,Tpet:0.0156387019,Tne	a:0.1452230523);	
((((Tpet:0.0110971015,Thea:0.1049271118):0.0289623851,Tmar:0.0243199255):0.0344706147,Thap:0.0000000001,TRQ:	2:0.0000000001);	
((((Tnap:0.000000001,TRQ2:0.000000001):0.0195559849,Tmar:0.0192488717):0.0302877982,Tpet:0.0042780600,Tnet	a:0.1222578973);	
((((TRQ2:0.0000000001,Tnap:0.000000001):0.0094732999,Tmar:0.0135548356):0.0241171361,Tnea:0.1295049747,Tper	t:0.0139575049);	
((((Tpet:0.0084017360,Tnea:0.1915035333):0.0325310503,Tmar:0.0183969456):0.0297356868,Tnap:0.0000000001,TRQ	2:0.0000000001);	
(((Tmar:0.0142155727,(TRQ2:0.0000000001,Tnap:0.0000000001):0.0096548737):0.0198020813,Tnea:0.1109612966,Tpe	t:0.0000009716);	
((((Tnap:0.0000000001,TRQ2:0.0000000001):0.0247349017,Tmar:0.0087802746):0.0523256720,Tnea:0.1562773333,Tpe	t:0.0033594668);	
- 2	///		

bcfam 247.1.5.0.phy phyml boot trees.txt

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: 0.000000001,TRQ2:0.000000001); (((Tpet:0.0219514318,Tnea:0.1960236242): 0.0145181752,Tmar:0.0189973964):0.0155785587,Tnap: 0.000000001,TRQ2:0.000000001); (((Tpet:0.0000004769,Tnea:0.1773430420): 0.0205769649,Tmar:0.0047117206):0.0416898504,Tnap: 0.000000001,TRQ2:0.000000001);

The spectrum

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

Quartet Decomposition





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	<u>192</u>	((Tmar,Tnea),(Tnap,Tpet));	
4	<u>98</u>	((Tmar,Tnea),(Tnap,TRQ2));	
8	<u>190</u>	((Tmar,TRQ2),(Tnap,Tpet));	
9	<u>103</u>	((Tmar,Tnea),(Tpet,TRQ2));	
13	<u>146</u>	((Tnap,Tpet),(Tnea,TRQ2));	

Quartet Decomposition

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

Quartet ID	Gene Family Numbers	Quartet Topology
0	<u>38</u>	((Tmar,Tnap),(Tnea,Tpet));
2	<u>55</u>	((Tmar,Tpet),(Tnap,Tnea));
3	<u>64</u>	((Tmar,Tnap),(Tnea,TRQ2));
5	<u>85</u>	((Tmar,TRQ2),(Tnap,Tnea));
6	<u>46</u>	((Tmar,Tnap),(Tpet,TRQ2));
7	<u>25</u>	((Tmar,Tpet),(Tnap,TRQ2));
10	<u>57</u>	((Tmar,Tpet),(Tnea,TRQ2));
11	<u>71</u>	((Tmar,TRQ2),(Tnea,Tpet));
12	<u>66</u>	((Tnap,Tnea),(Tpet,TRQ2));
14	<u>49</u>	((Tnap,TRQ2),(Tnea,Tpet));



Quartet Decomposition

Good quartets with bootstrap support value > 0.9			
Download as newick trees			
Quartet ID	Gene Family Numbers	Quartet Topology	
1	<u>192</u>	((Tmar,Tnea),(Tnap,Tpet));	
4	<u>98</u>	((Tmar,Tnea),(Tnap,TRQ2));	
8	<u>190</u>	((Tmar,TRQ2),(Tnap,Tpet));	
9	<u>103</u>	((Tmar,Tnea),(Tpet,TRQ2));	
13	<u>146</u>	((Tnap,Tpet),(Tnea,TRQ2));	

5	2570	
TRQ2	77777777777777777777777777777777	101010101010101010
Tmar	101010101010101010101010101	???????????????????????????????????????
Tnap	010101010101010101010101010	101010101010101010
Tnea	101010101010101010101010101	01010101010101010101
Tpet	010101010101010101010101010	01010101010101010101



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



from uncorrected P distances

from uncorrected P distances