

Triads, Dyads, and Gene Functions: When Social Network Analysis meets Phylogenetics

George G Vega Yon, Ph.D.

(with Paul D Thomas Paul Marjoram Huaiyu Mi Duncan Thomas John Morrison)

University of Southern California
Department of Population and Public Health Sciences

Networks 2021

(virtual)

July 8, 2021



The problem of genes' functions

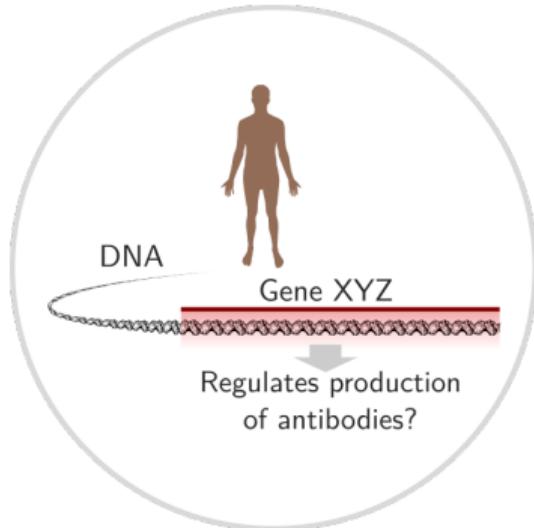
Avoiding the Curse of dimensionality

Analyzing 77 experimentally annotated trees

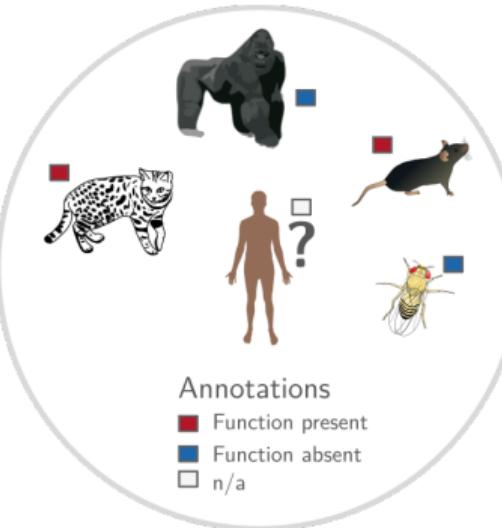
You can download the slides from ggv.cl/slides/networks2021

The problem of genes' functions

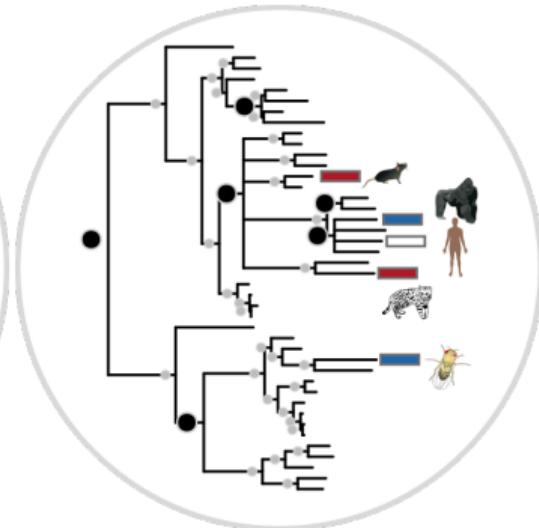
Is gene *XYZ* involved in process *ABC*?



Complex to directly assess



But we may know from other species



And we further know how these *genetically connected*

... let's rephrase the question.

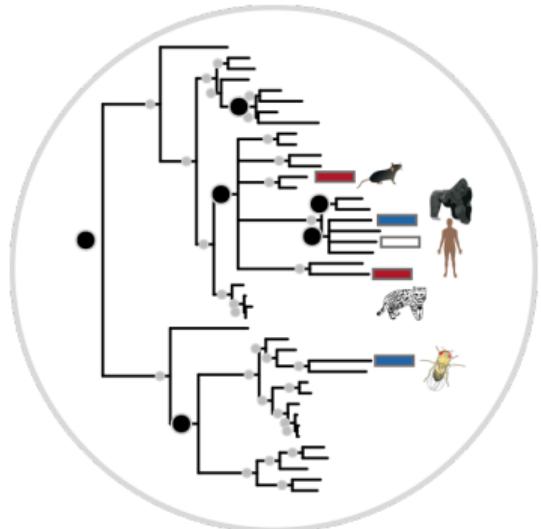
Is the human gene **XYZ** involved in process **ABC**, given what we know about that for other *related species*?



Annotations

- Function present
- Function absent
- n/a

The Gene Ontology Project

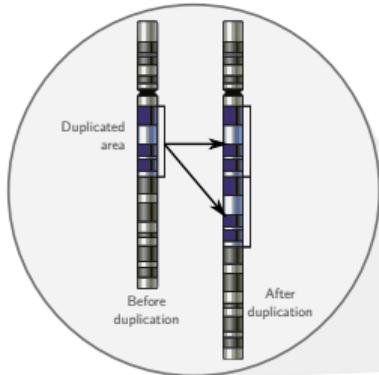


- ▶ ~ 15,000 phylogenetic trees
- ▶ ~ 8 million annotations
- ▶ ~ 600 thousand on human genes
- ▶ ~ < 10% are based on experimental evidence... Improving our knowledge on genetics is fundamental for advancing Biomedical Research

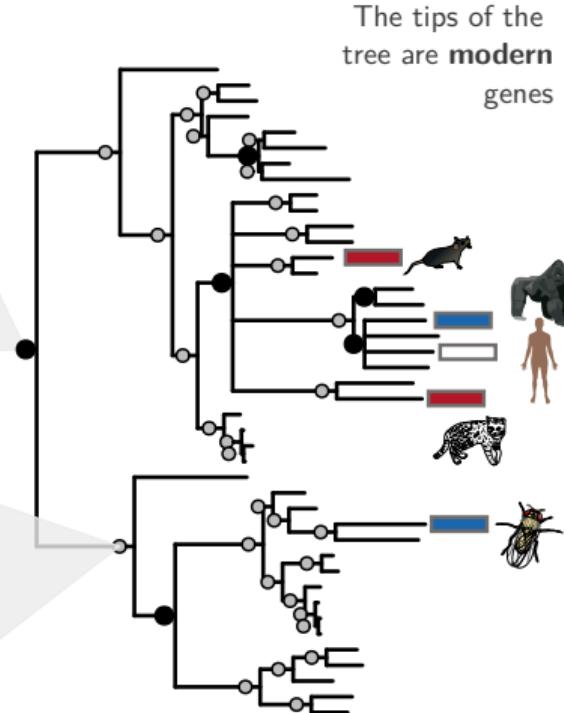
Only on 2020, 2,000+ COVID-19 papers using the GO (Google Scholar)

► more

- nodes are Duplication Events

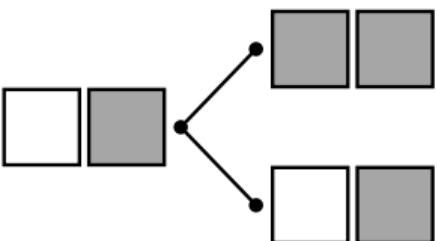
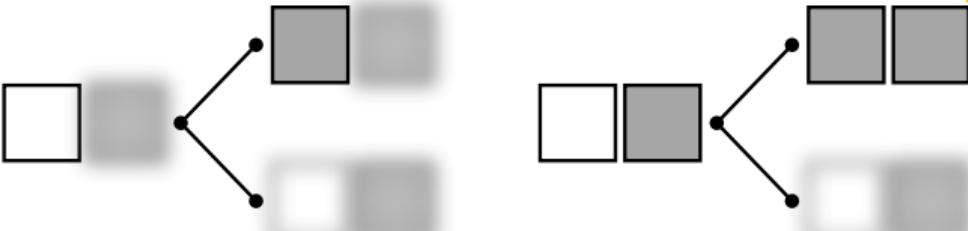
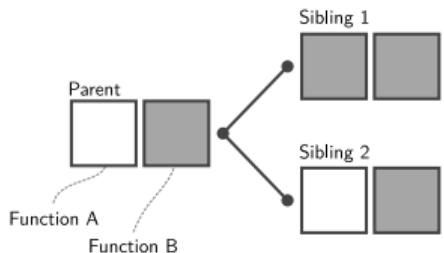


- nodes are Speciation Events



▶ example

Phylogenetics Modeling Strategies

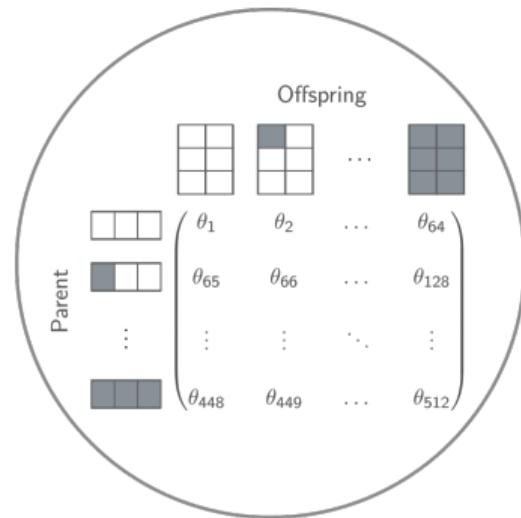


(c) No conditional
independence

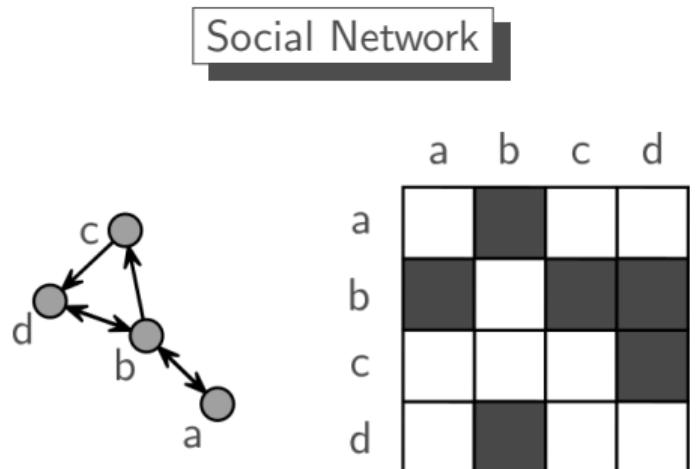
Avoiding the Curse of dimensionality

If we wanted to build a model with 3 functions, we would need to estimate...

Full Markov Transition Matrix



- ▶ 512 parameters
- ▶ Finding this many parameters not easy.
- ▶ Even if you can, interpretation is awkward.

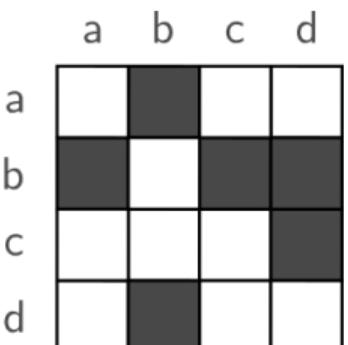
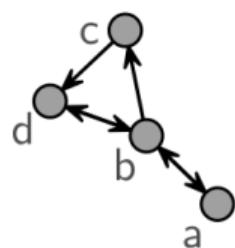


- ▶ Not about individual ties.
- ▶ Statistical inference on *motifs* (triangles, dyads, homophily, etc.)

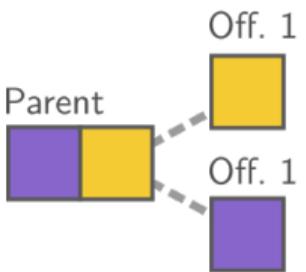
Ultimately...

ERGM ≡ Modeling binary arrays

Social Network



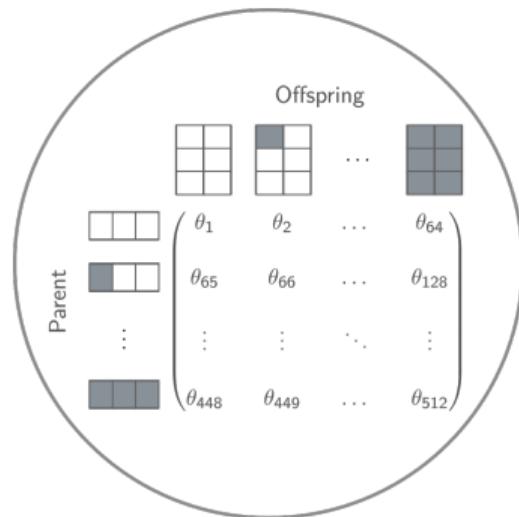
Evolutionary Event



Social Networks are usually represented as **adjacency matrices**, and so can evolutionary events!

If we wanted to build a model with 3 functions, we would need to estimate...

Full Markov Transition Matrix



512 parameters

Sufficient statistics

$$\begin{aligned} &(\theta_{\text{gain}1} \theta_{\text{gain}2} \theta_{\text{gain}3}) \\ &(\theta_{\text{loss}1} \theta_{\text{loss}2} \theta_{\text{loss}3}) \\ &\theta_{\text{neofun}} \theta_{\text{subfunc}} \\ &(\theta_{\text{root}1} \theta_{\text{root}2} \theta_{\text{root}3}) \end{aligned}$$

Easier to fit
Easier to interpret

$$\begin{aligned} &(\theta_{\text{gain}1} \theta_{\text{gain}2} \theta_{\text{gain}3}) \\ &(\theta_{\text{loss}1} \theta_{\text{loss}2} \theta_{\text{loss}3}) \\ &\theta_{\text{neofun}} \theta_{\text{subfunc}} \\ &(\theta_{\text{root}1} \theta_{\text{root}2} \theta_{\text{root}3}) \end{aligned}$$

11 parameters (for example)

◀ term examples

Tree likelihoods: Felsenstein's Pruning algorithm

Also known as *dynamic programming* or *postorder tree traversal*

$$\mathbb{P}(\tilde{D}_n \mid \mathbf{x}_n, \Theta) = \sum_{\mathbf{x}} \mathbb{P}(\mathbf{x} \mid \mathbf{x}_n) \prod_{m \in O(n)} \mathbb{P}(\tilde{D}_m \mid \mathbf{x}_m)$$

All possible transitions from \mathbf{x}_n Transition Probability (ERGM)

$$\mathbb{P}(\mathbf{x} \mid \mathbf{x}_n) = \frac{\exp \{ \Theta^t s(\mathbf{x}, \mathbf{x}_n) \}}{\sum_{\mathbf{x}'} \exp \{ \Theta^t s(\mathbf{x}', \mathbf{x}_n) \}}$$

Model Parameters Vector of Sufficient Statistics

Normalizing Constant

the *lingua franca* of SNA

Gene state given the data It's parent state given the data

$$\mathbb{P}(x^p = x \mid \tilde{D}) = \underbrace{\left\{ \prod_{m \in O(p)} \mathbb{P}(\tilde{D}_m \mid x_m) \right\}}_{\text{Everything below } x^p} \sum_{x_p} \mathbb{P}(x_p \mid \tilde{D}) \underbrace{\frac{\mathbb{P}(x^p = x \mid x_p)}{\mathbb{P}(\tilde{D}_p \mid x_p)}}_{\text{Everything above } x^p}$$

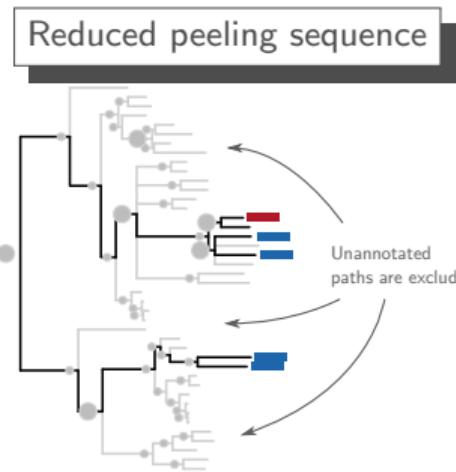
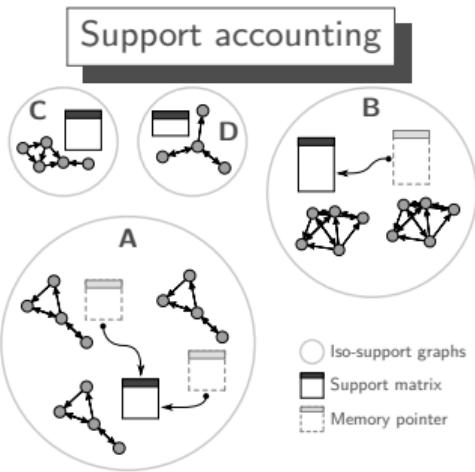
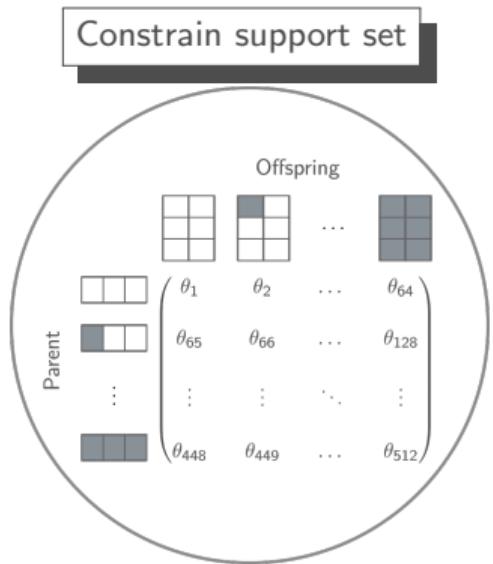
... I implemented this (and more) on **geese**



GEne functional Evolution using SufficiEncy

... as part of **barry**, your to-go motif accountant

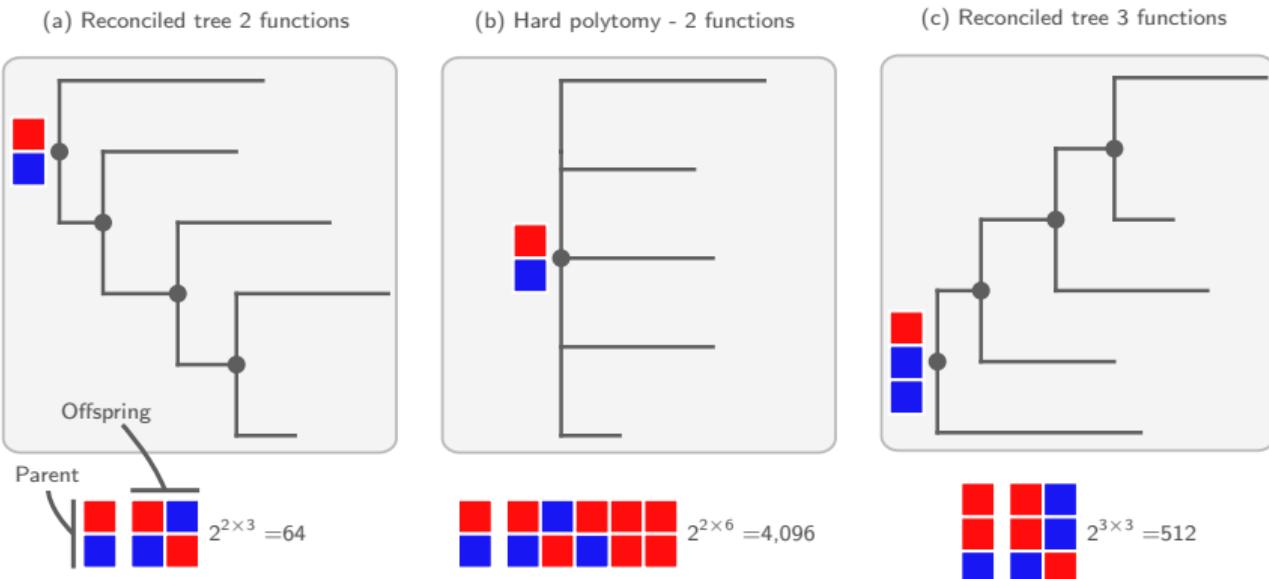


Computational Features of **geese**

... how big can we go?

Support size: Computational feasibility

Whether the likelihood is tractable or not depends on how large is the “largest event”



(in practice, arrays up to 32 cells, i.e., 4.3 billion comb., are feasible.)

Analyzing 77 experimentally annotated trees
(Vega Yon, *et al.*, WIP)

Analyzing 77 experimentally annotated trees

Using data from Vega Yon et al. (2021), we re-estimated the models using **GEESE** and compared the results to those in the **aphylo** paper:

- ▶ 77 experimentally annotated trees
- ▶ We only used trees in which

$$2^{(\max \text{Polytomy}+1) \times \text{nfun}} < 0.5 \times 10^9$$

i.e., half billion

- ▶ Both sets of models used "informative" priors.
- ▶ Both sets of were fitted using Robust Adaptive Metropolis (**fmcmc** R package).

more

Example of code (R)

After initializing a geese object named model2fit:

```
1 # For later use (see last two lines)
2 term_overall_changes(model2fit, duplication = TRUE)
3 term_overall_changes(model2fit, duplication = FALSE)
4
5 # Couting how many genes change
6 term_genes_changing(model2fit, duplication = TRUE)
7
8 # Gain at duplication
9 term_gains(model2fit, funs = 0:nfun, duplication = TRUE)
10
11 # Gain and loss at speciation
12 term_gains(model2fit, funs = 0:nfun, duplication = FALSE)
13 term_loss(model2fit, funs = 0:nfun, duplication = FALSE)
14
15 # Constraining the support set
16 rule_limit_changes(model2fit, id = 0, lb = 0, ub = 4, duplication = TRUE)
17 rule_limit_changes(model2fit, id = 1, lb = 0, ub = 4, duplication = FALSE)
```

Figure 1 Performance as Mean Absolute Error (MAE). Each set of coordinates shows the value of 1 minus the MAE of **GEESE**, x-axis, and **aphylo** (y-axis). The statistics were computed using leave-one-out based on individual parameter estimates. Overall, GEESE performs better than aphylo in most cases.

Tapping into computational scalability

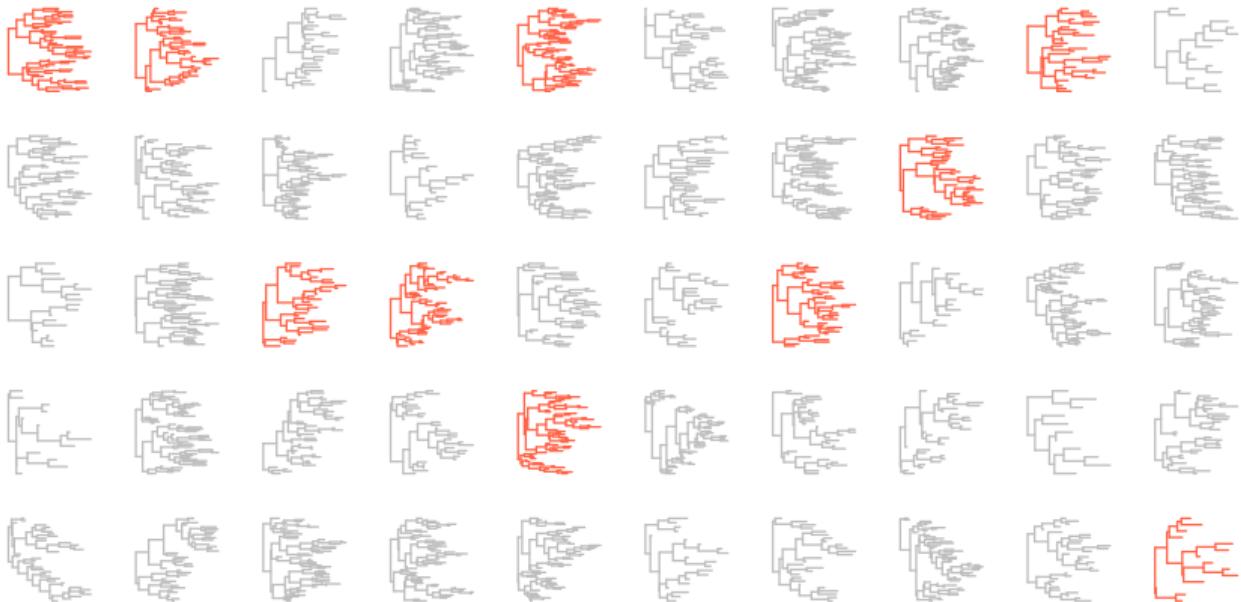


Figure 2 A dramatization of how a group of GEESE, i.e., a flock, looks like.

Applications of Statistical Computing in Complex Social and Biological Systems Modeling

George G Vega Yon

<https://ggyv.cl>

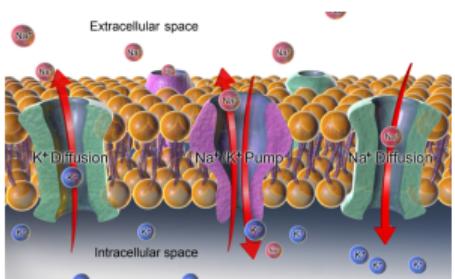
vegayon@usc.edu

Thank you!

Gene functions can be classified in three types:

Molecular function

Active transport GO:0005215



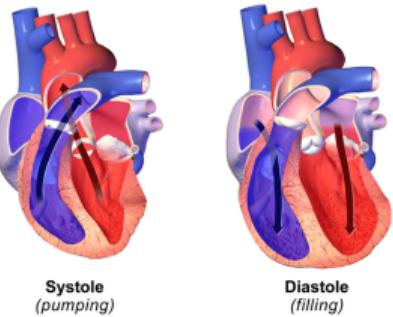
Cellular component

Mitochondria GO:0004016



Biological process

Heart contraction GO:0060047



◀ go back

The Gene Ontology Project

Example of GO term

Accession	GO:0060047
Name	heart contraction
Ontology	biological_process
Synonyms	heart beating, cardiac contraction, hemolymph circulation
Alternate	IDs None
Definition	The multicellular organismal process in which the heart decreases in volume in a characteristic way to propel blood through the body. Source: GOC:dph

Table 1 Heart Contraction Function. source: amigo.geneontology.org

◀ go back

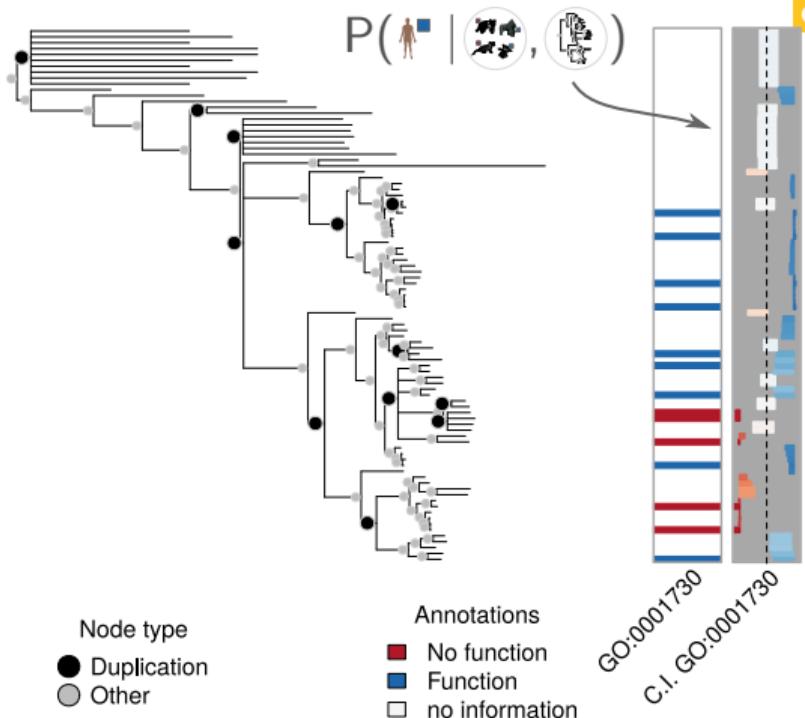
Example: Molecular function in family PTHR1128

Name: 2'-5'-oligoadenylate synthetase activity

Desc: GO:0001730 involved in the process of cellular antiviral activity (wiki on [interferon](#)).

MAE: 0.34

AUC: 0.91



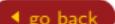
Note: Prediction made using **aphylo** (Vega Yon and *et al*, PLOS Comp. Bio 2021)

◀ go back

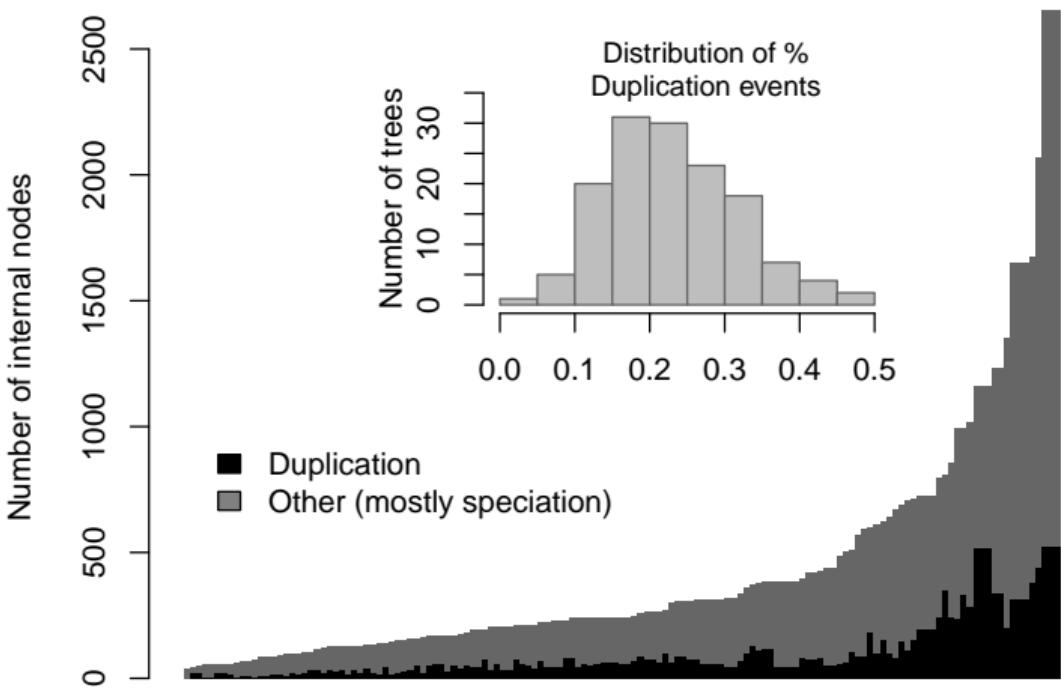
Data: Phylogenetic trees

Sample of annotations (first 10 in a single tree, Phosphoserine Phosphatase [PTHR10000])

Internal id	Branch Length	type	ancestor
AN0		S	LUCA
AN1	0.06	S	Archaea-Eukaryota
AN2	0.24	S	Eukaryota
AN3	0.44	S	Unikonts
AN4	0.42	S	Opisthokonts
AN6	0.68	D	
AN9	0.79	S	Amoebozoa
AN10	0.18	D	
AN15	0.57	S	Dictyostelium
AN18	0.52	S	Alveolata-Stramenopiles

◀ go back

Data: Node type (events)

[◀ go back](#)

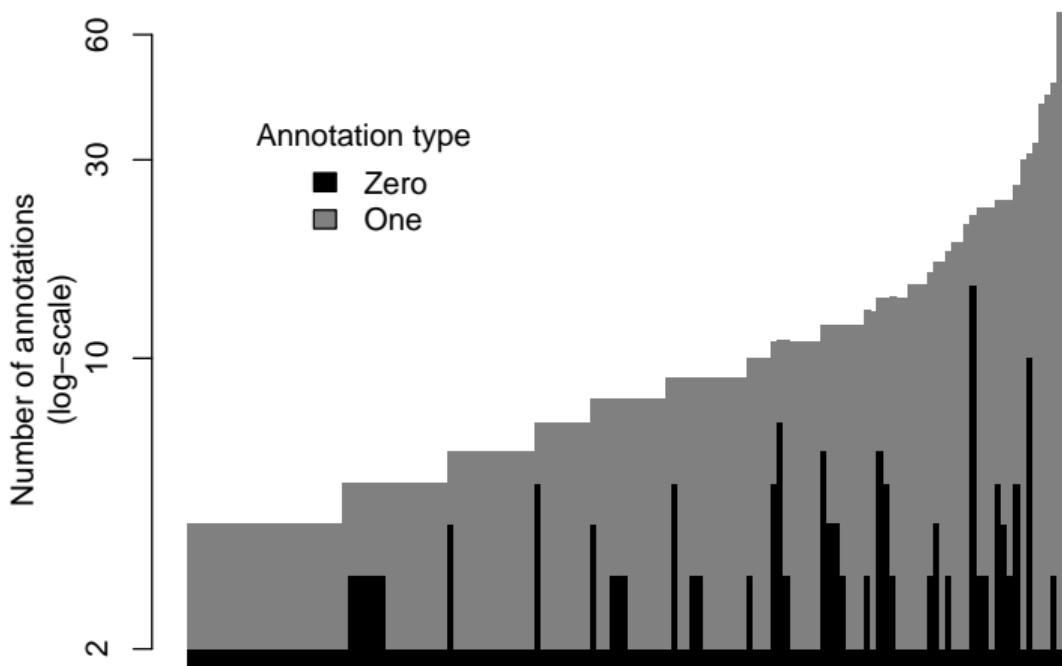
Data: Annotations (example)

This is the first 10 of ~ 400,000 experimental annotations used:

	Family	Id	GO term	Qualifier
1	PTHR12345	HUMAN HGNC=15756 UniProtKB=Q9H190	GO:0005546	
2	PTHR11361	HUMAN HGNC=7325 UniProtKB=P43246	GO:0016887	CONTRIBUTES_TO
3	PTHR10782	MOUSE MGI=MGI=3040693 UniProtKB=Q6P1E1	GO:0045582	
4	PTHR23086	ARATH TAIR=AT3G09920 UniProtKB=Q8L850	GO:0006520	
5	PTHR32061	RAT RGD=619819 UniProtKB=Q9EPI6	GO:0043197	
6	PTHR46870	ARATH TAIR=AT3G46870 UniProtKB=Q9STF9	GO:1990825	
7	PTHR15204	MOUSE MGI=MGI=1919439 UniProtKB=Q9Z1R2	GO:0045861	
8	PTHR22928	DROME FlyBase=FBgn0050085 UniProtKB=Q9XZ34	GO:0030174	
9	PTHR35972	HUMAN HGNC=34401 UniProtKB=A2RU48	GO:0005515	
10	PTHR10133	DROME FlyBase=FBgn0002905 UniProtKB=O18475	GO:0097681	

◀ go back

Data: Experimental Annotations



◀ go back

Overview of Prediction Results: aphylo

	Pooled	Type of Annotation		
		Molecular Function	Biological Process	Cellular Comp.
Mislabeling				
ψ_{01}	0.23	0.18	0.09	
ψ_{10}	0.01	0.01	0.01	
Duplication Events				
μ_{d01}	0.97	0.97	0.10	
μ_{d10}	0.52	0.51	0.03	
Speciation Events				
μ_{s01}	0.05	0.05	0.05	
μ_{s10}	0.01	0.01	0.02	
Root node				
π	0.79	0.71	0.88	
Trees	141	74	45	22
Accuracy under the by-aspect model				
AUC	-	0.77	0.83	
MAE	-	0.34	0.26	
Accuracy under the pooled-data model				
AUC	-	0.77	0.75	
MAE	-	0.35	0.34	

Previously, joint estimates out-performed one-at-a-time

- ▶ **Molecular Function** No change.
- ▶ **Biological Process** Significantly better.
- ▶ **Cellular Component** Does not converge.

Molecular Function \neq Biological Process ? Cellular Component

▶ data

▶ go back

Barry: your go-to *motif* accountant

- ▶ Sparse matrix represented using double hashmaps (fast row/column access).
- ▶ Template implementation for flexible weights and metadata.
- ▶ Fast counting using change statistics (Ch. 4).
- ▶ Calculation of support for sufficient stats.

[https://USCbiostats.github.io/
binaryarrays](https://USCbiostats.github.io/binaryarrays)

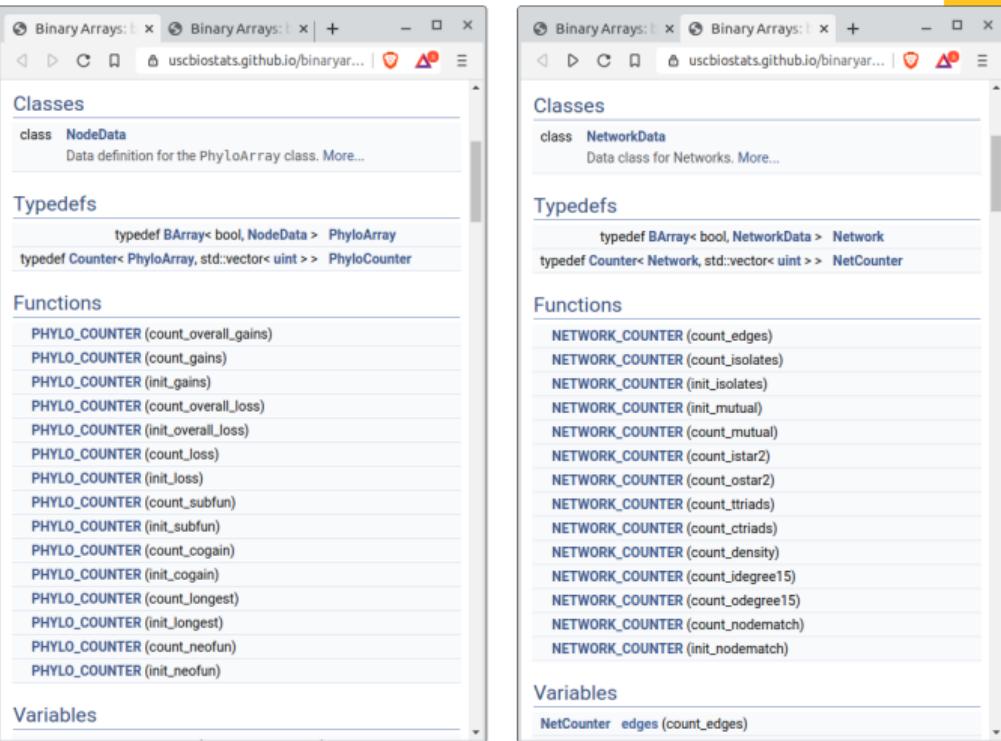


Figure 3 Screenshots from the project's website on GitHub.

What Drives Evolution

Imagine that we have 3 functions (rows) and that each node has 2 siblings (columns)

		Transitions to	
		Case 1	Case 2
Parent	A	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
	B	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$
	C	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$

Sufficient statistics

# Gains	1	1
Only one offspring changes (yes/no)	1	0
# Changes (gain+loss)	2	3
Subfunctionalizations (yes/no)	0	1

▶ return

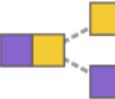
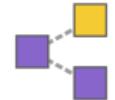
Representation	Description	Definition
	Gain of function	$(1 - x_p) \sum_{n:n \in Off} x_n$
	Loss of function	$x_p \sum_{n:n \in Off} (1 - x_n)$
	Subfunctionalization	$x_p^k x_p^j \sum_{n \neq m} x_n^k (1 - x_n^j) (1 - x_m^k) x_m^j$
	Neofunctionalization	$x_p^k (1 - x_p^j) \sum_{n \neq m} x_n^k (1 - x_n^j) (1 - x_m^k) x_m^j$
	Longest branch gains	$(1 - x_p^k) \mathbf{1} (x_m^k : m = \text{argmax}_n \text{blength}_n)$

Table 2 Example of sufficient statistics for evolutionary transitions.

Example: Simple model with two functions

To illustrate, we will **simulate** and then **estimate** the parameters for the following process:

1. 100 genes on a simulated phylogenetic tree.
2. Two functions, **0** and **1**,
3. **Function 0** is likely to be gain gained at a dupl. event,
4. **Function 1** is gained as neofunctionalization (**from 0**) at a dupl. event,
5. There is a higher chance of changes at duplication (explicit).
6. Root node starts off without either function (i.e. prob $\rightarrow 0$).

We will fit the model using Robust Adaptive Metropolis with a logistic prior centered at 0 with scale 2.

◀ go back

Example: Simple model with two functions

posterior distributions

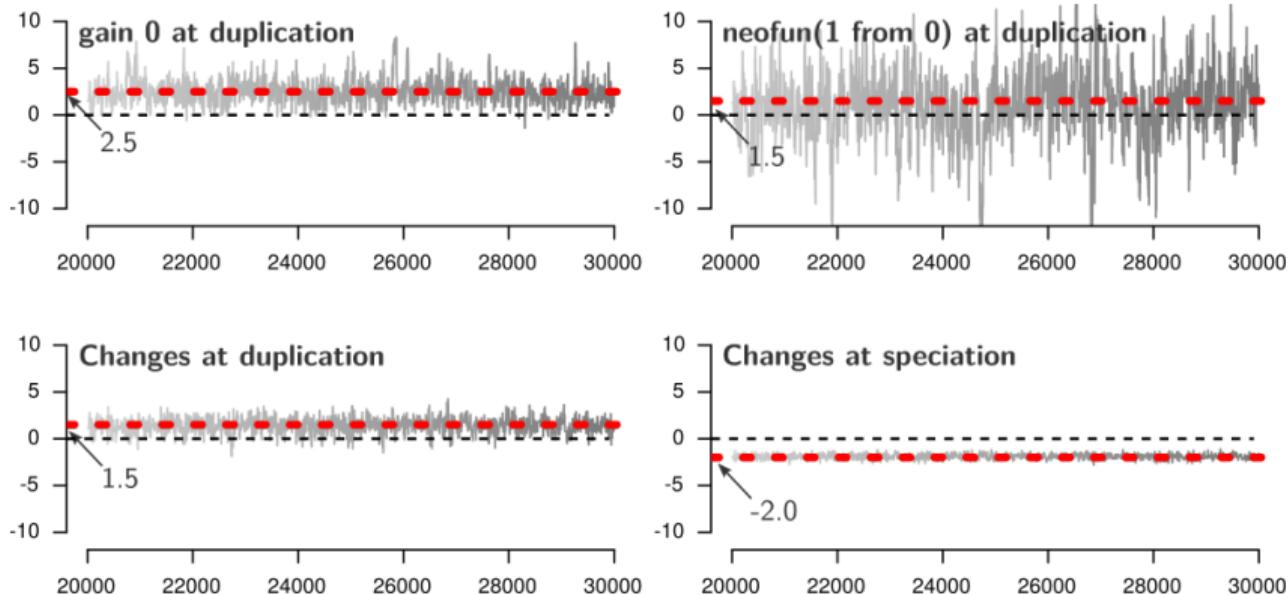


Figure 4 MCMC Trace of the functional gain of 0, neofunctionalization (1 from 0), and change rate (by event type).

Example: Simple model with two functions posterior distributions (contd')

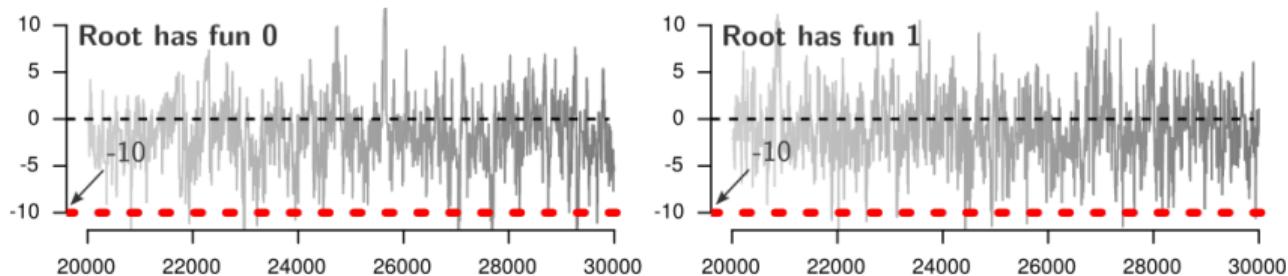


Figure 5 MCMC Trace of root parameters. The true population parameters are $(\theta_{root0}, \theta_{root1}) = (-10.0, -10.0)$.
Root node probabilities are always hard to get.

Figure 6 Distribution of parameter estimates from 5,000 phylo trees
w/ 100 leafs.

Repeated this experiment 5,000 times:

- ▶ MCMC for fitting.
- ▶ RAM kernel.
- ▶ Logistic prior at zero with scale two.
- ▶ Each tree took < 1min estimation.

