

Triads, Dyads, and Gene Functions

When Social Network Analysis Meets Phylogenetics

George G Vega Yon, Ph.D.

University of Southern California, Department of Preventive Medicine

NY Genome Center

March 9, 2021

Parsimonious modeling of gene functional evolution

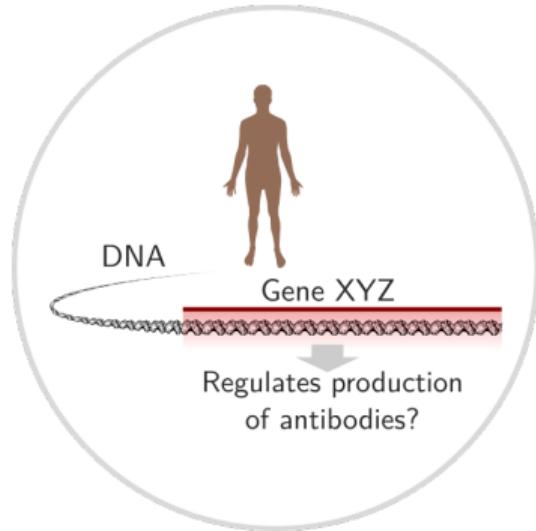
A general framework for modeling functional evolution

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

Parsimonious modeling of gene functional evolution

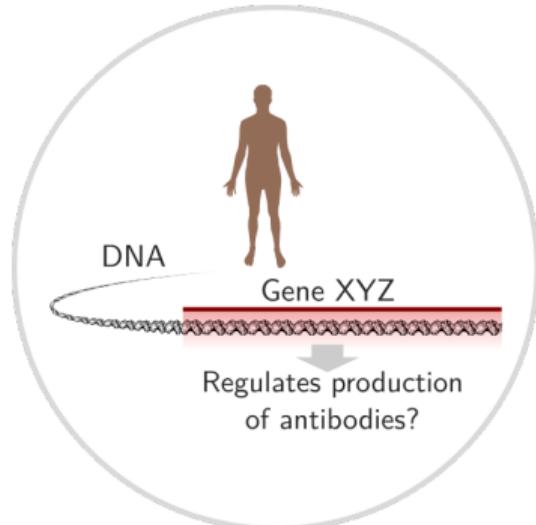
Joint with: Paul D Thomas, Paul Marjoram, Huaiyu Mi, Duncan Thomas, and John Morrison
(Published at *PLOS Computational Biology*)

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

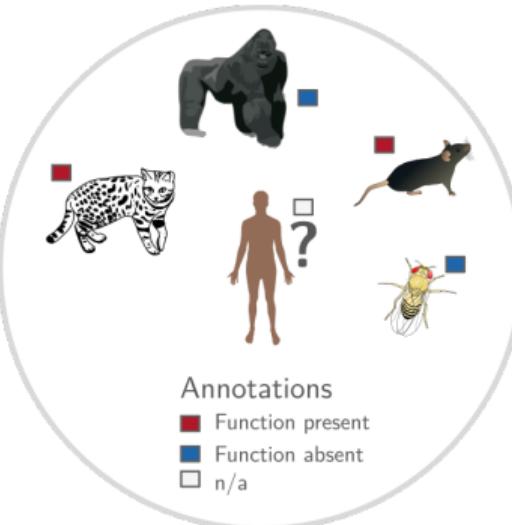


Complex to directly assess

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

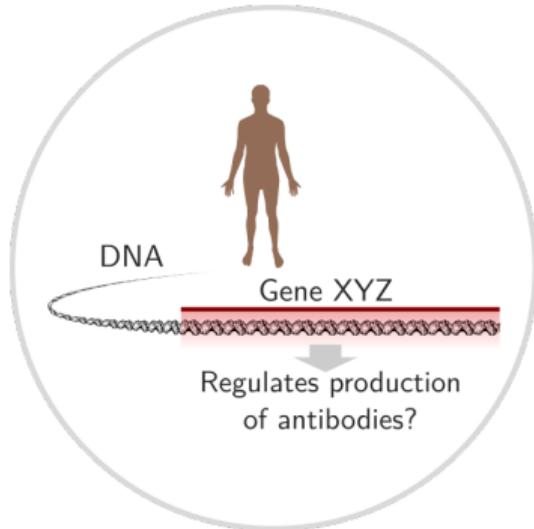


Complex to directly assess

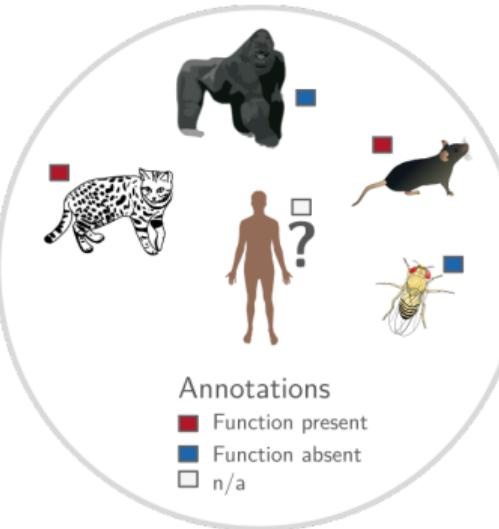


But we may know from other
species

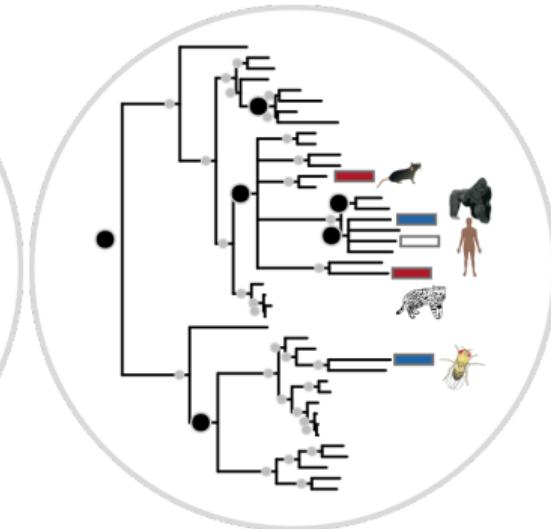
Is gene XYZ involved in process ABC?



Complex to directly assess

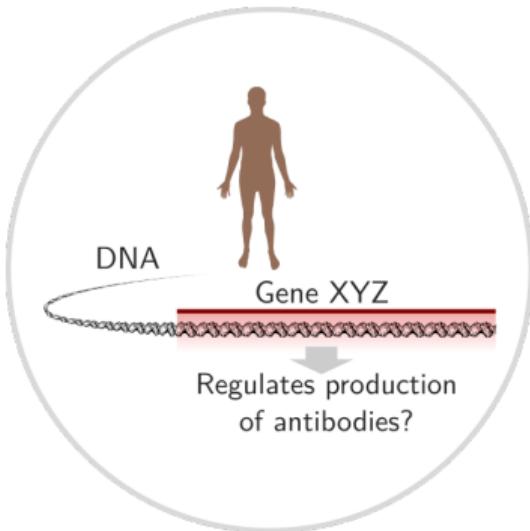


But we may know from other species

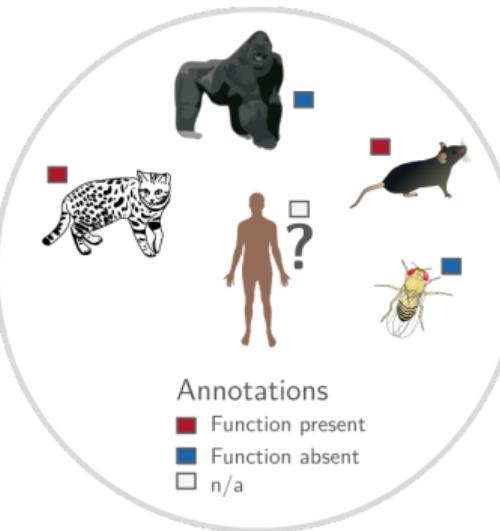


And we further know how these *genetically connected*

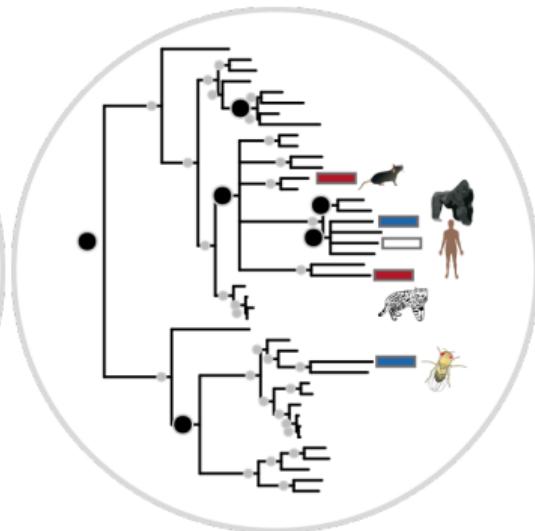
Is gene *XYZ* involved in process *ABC*?



Complex to directly assess



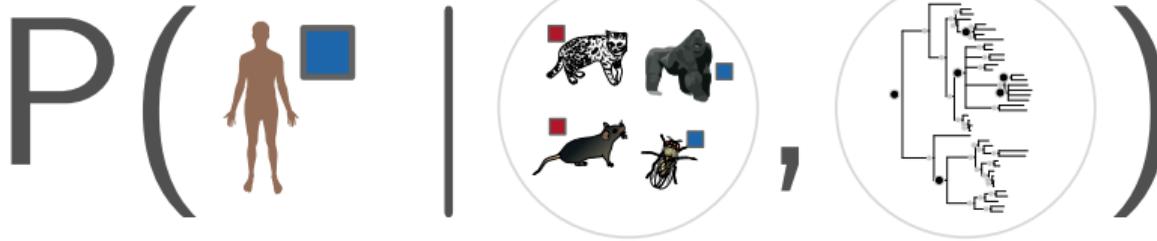
But we may know from other species



And we further know how these *genetically connected*

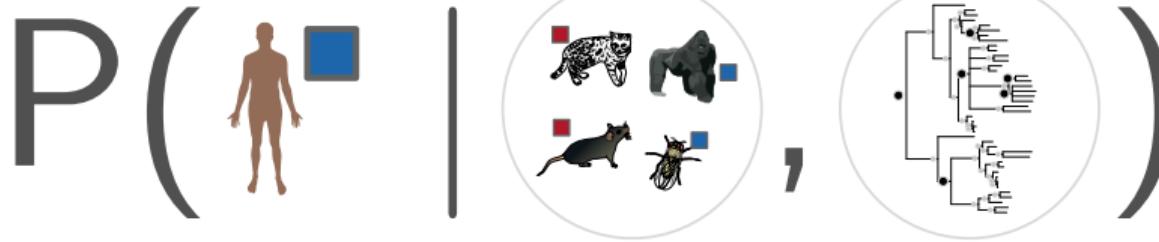
... let's rephrase the question.

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

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



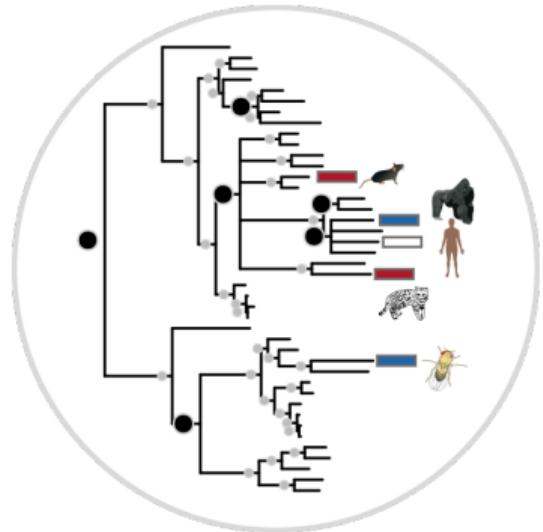
Annotations

- Function present
- Function absent
- n/a

... Where is all this data?

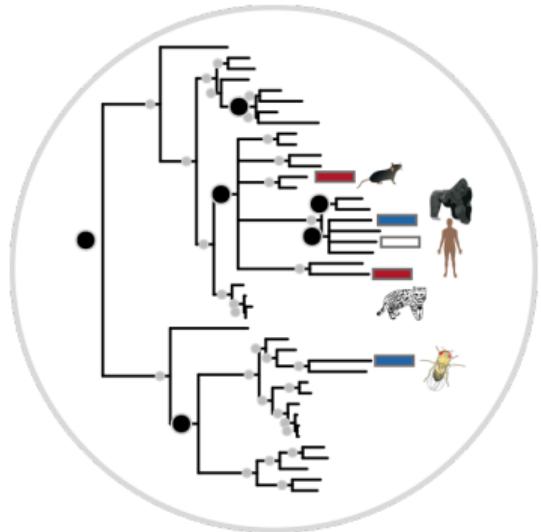
▶ more

The Gene Ontology Project



▶ more

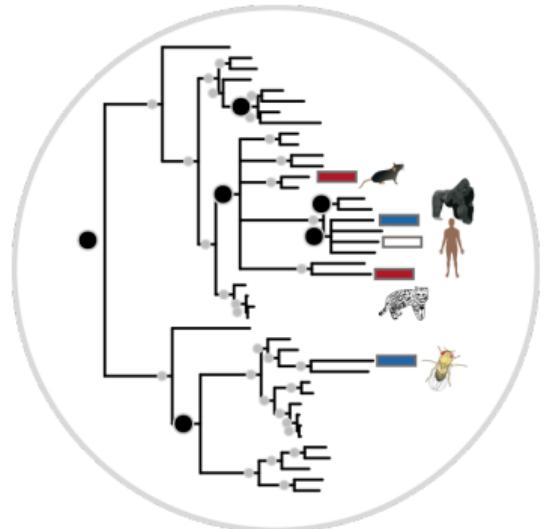
The Gene Ontology Project



► ~ 15,000 phylogenetic trees

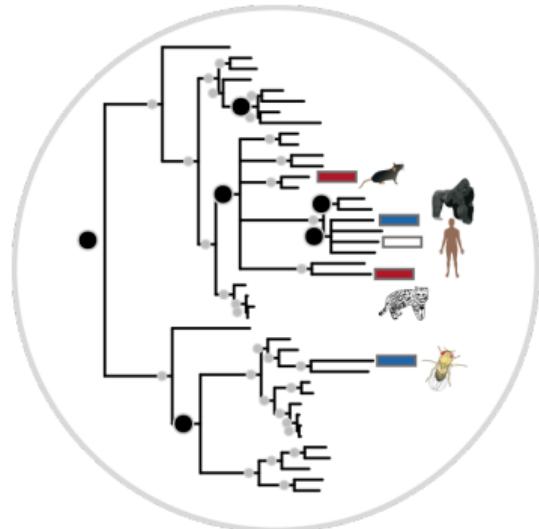
► more

The Gene Ontology Project



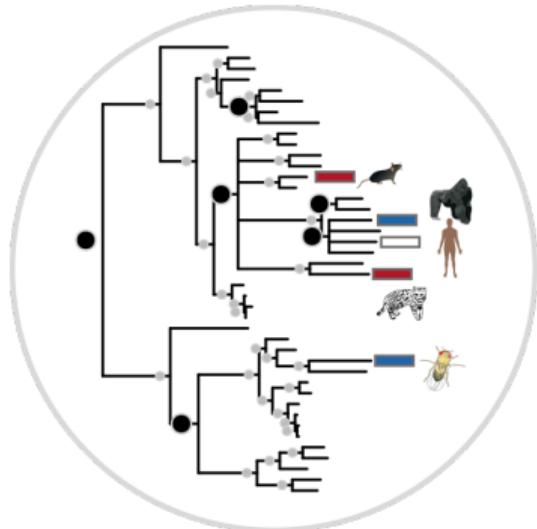
- ▶ ~ 15,000 phylogenetic trees
- ▶ ~ 8 million annotations

► more



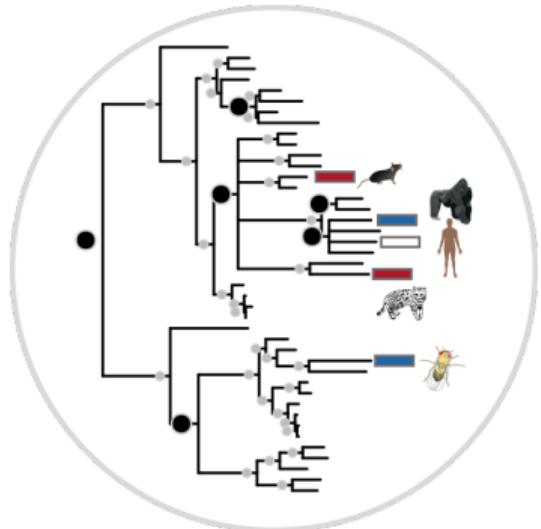
- ▶ ~ 15,000 phylogenetic trees
- ▶ ~ 8 million annotations
- ▶ ~ 600 thousand on human genes

▶ more



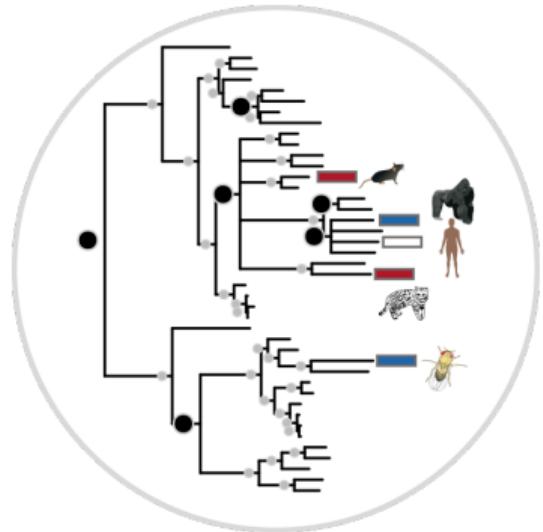
- ▶ ~ 15,000 phylogenetic trees
- ▶ ~ 8 million annotations
- ▶ ~ 600 thousand on human genes
- ▶ ~ < 10% are based on experimental evidence...

▶ more



- ▶ ~ 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

▶ more



- ▶ ~ 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 2021, 2,500+ Cancer papers using the GO (Google Scholar)

▶ more

An evolutionary model of gene functions

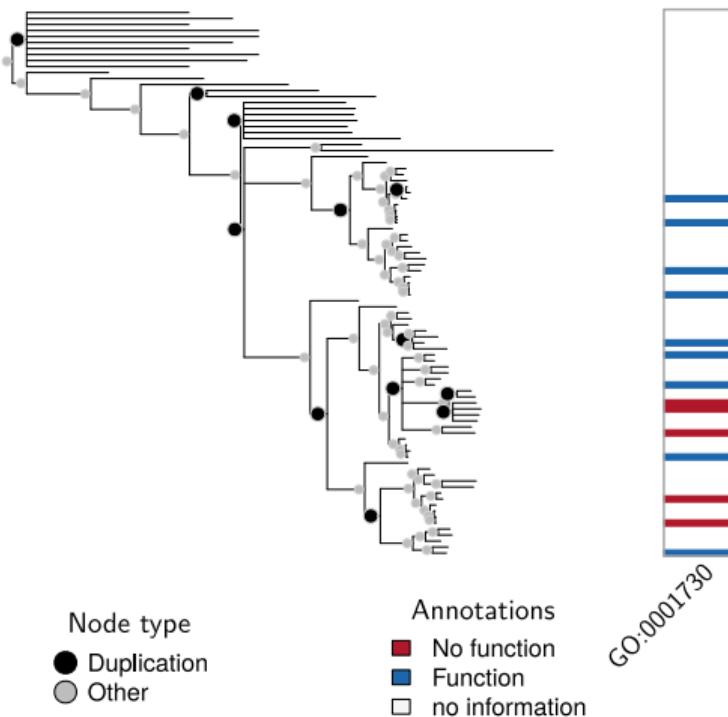
Family: PTHR11258

Type: Molecular Function

Name: 2'-5'-oligoadenylate synthetase activity

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

[see details](#)



An evolutionary model of gene functions

Family: PTHR11258

Type: Molecular Function

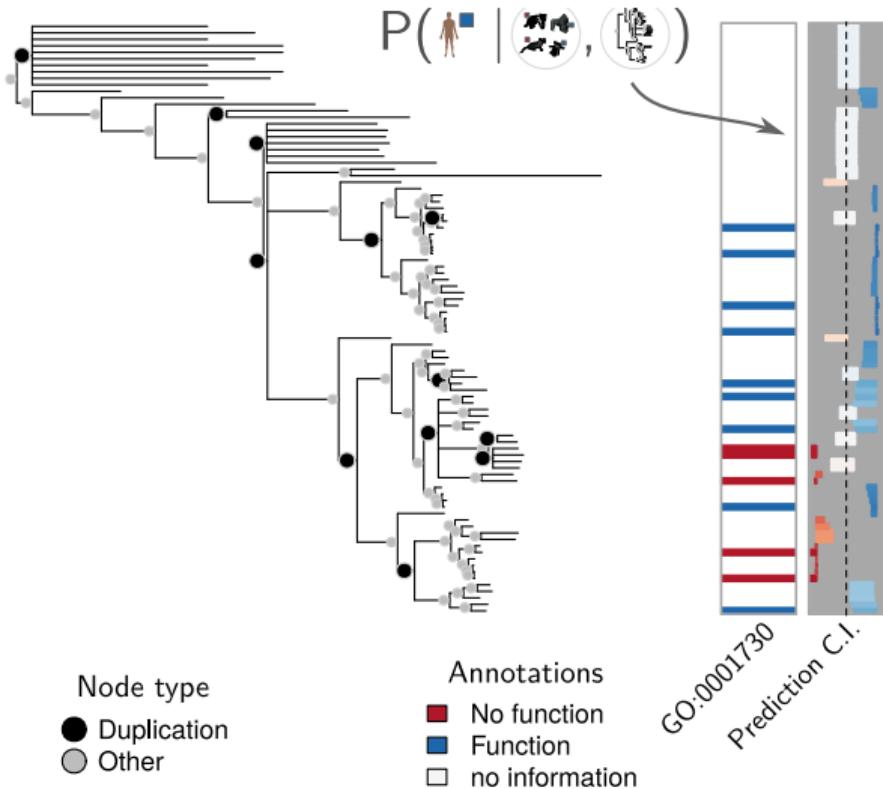
Name: 2'-5'-oligoadenylate synthetase activity

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

MAE: 0.34

AUC: 0.91

[see details](#)



An evolutionary model of gene functions

Family: PTHR11258

Type: Molecular Function

Name: 2'-5'-oligoadenylate synthetase activity

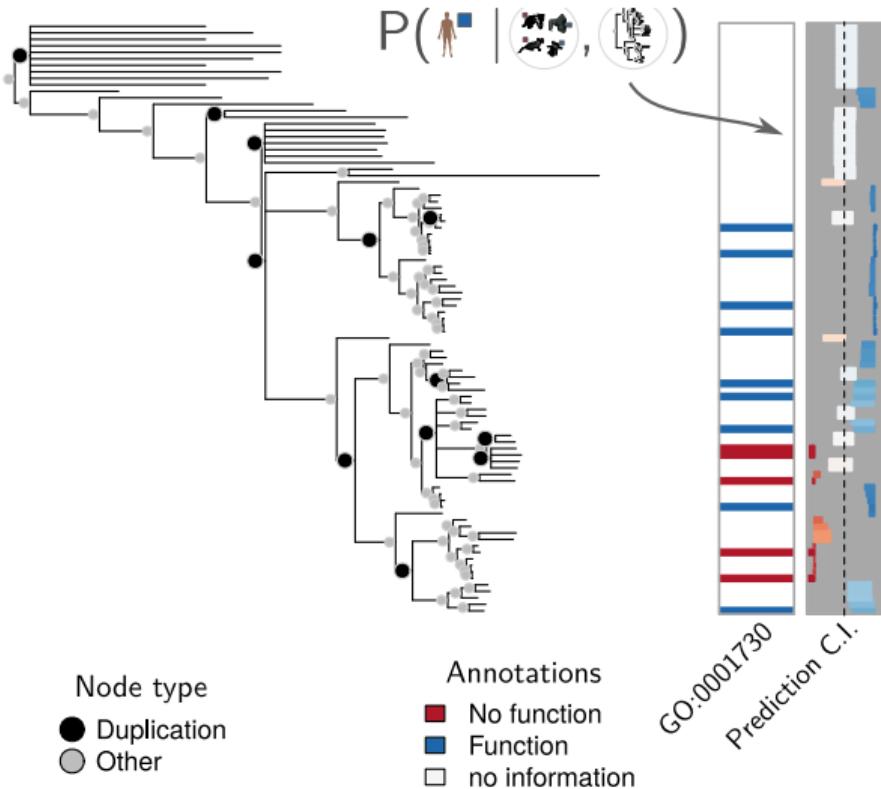
Desc: GO:0001730 involved in the process of cellular antiviral activity (wiki on [interferon](#)).

MAE: 0.34

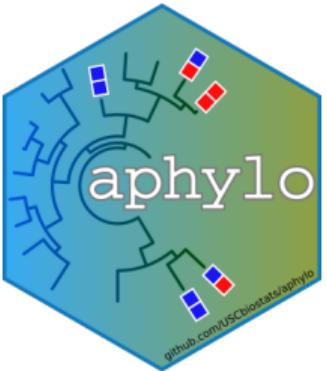
AUC: 0.91

I implemented this model in the **aphylo R package**

[see details](#)



Results: What does aphylo brings to the table?

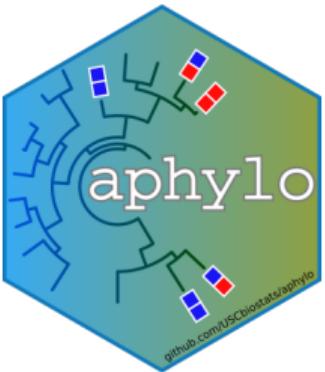


Large scale

Estimate pooled-data
models involving hundreds
of families
(1,300 genes at a time)

▶ comp. feats ▶ details

Results: What does aphylo brings to the table?



Large scale

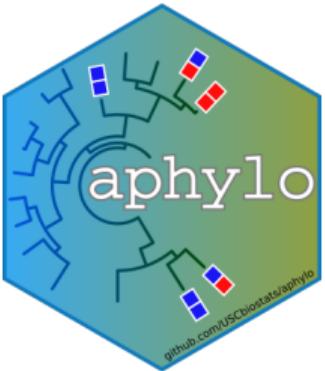
Interpretable

Estimate pooled-data
models involving hundreds
of families
(1,300 genes at a time)

Pooled-data model
provides inference aligned
with theoretical results
(gene duplication is key)

▶ comp. feats ▶ details

Results: What does aphylo brings to the table?



Large scale

Estimate **pooled-data**
models involving hundreds
of families
(1,300 genes at a time)

Interpretable

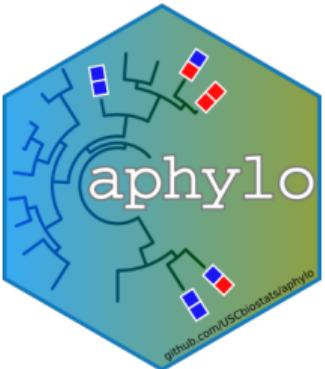
Pooled-data model
provides inference aligned
with theoretical results
(gene duplication is key)

Fast

Computational efficiency
allows making **inference**
and prediction fast
(1 second vs 2 hours)

▶ comp. feats ▶ details

Results: What does aphylo brings to the table?



Large scale

Estimate **pooled-data** models involving **hundreds of families** (1,300 genes at a time)

Interpretable

Pooled-data model provides inference aligned with theoretical results (gene duplication is key)

Fast

Computational efficiency allows making **inference and prediction fast** (1 second vs 2 hours)

Accuracy

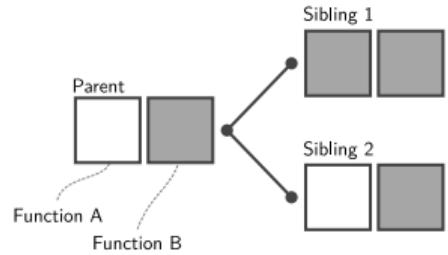
Outperforms state-of-the-art phylo-models (0.72 vs 0.60 AUC)

▶ comp. feats

▶ details

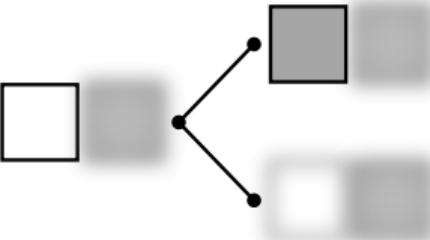
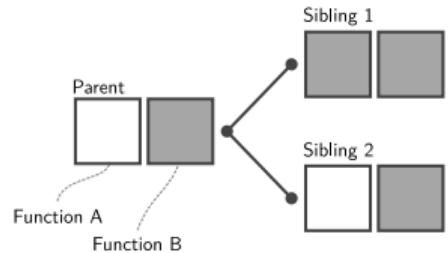
A general framework for modeling functional evolution

Phylogenetics Modeling Strategies



- White box: Has the function
- Gray box: Doesn't have the function

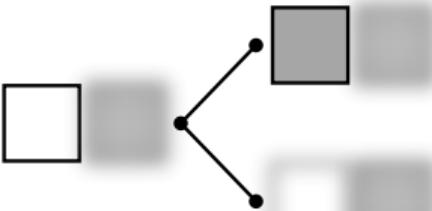
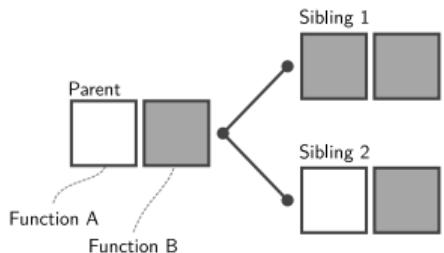
Phylogenetics Modeling Strategies



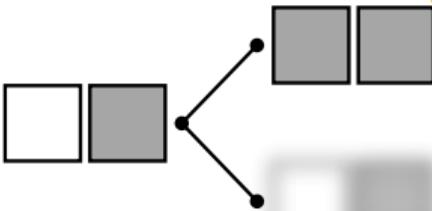
(a) Sibling and Function
Conditional Independence

- [White square] Has the function
- [Gray square] Doesn't have the function

Phylogenetics Modeling Strategies



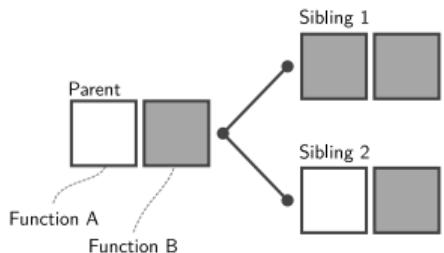
(a) Sibling and Function Conditional Independence



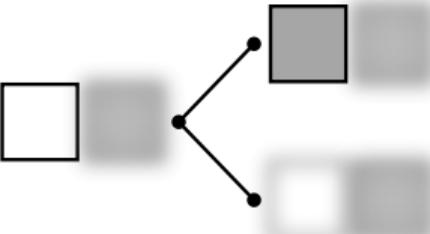
(b) Sibling Conditional Independence

- [White Square] Has the function
- [Gray Square] Doesn't have the function

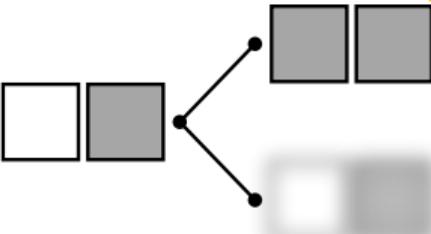
Phylogenetics Modeling Strategies



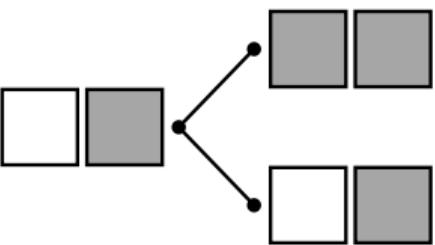
Has the function
Doesn't have the function



(a) Sibling and Function Conditional Independence



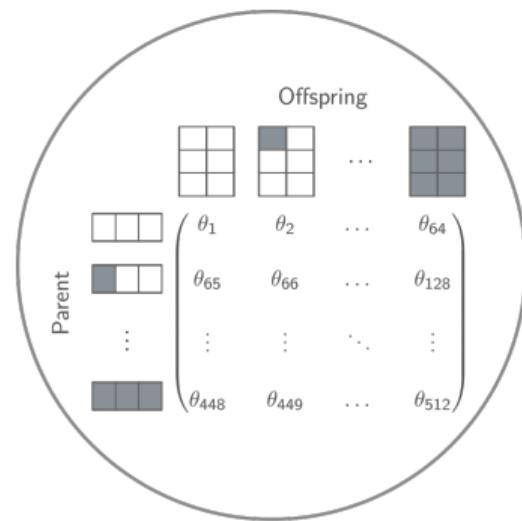
(b) Sibling Conditional Independence



(c) No conditional independence

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

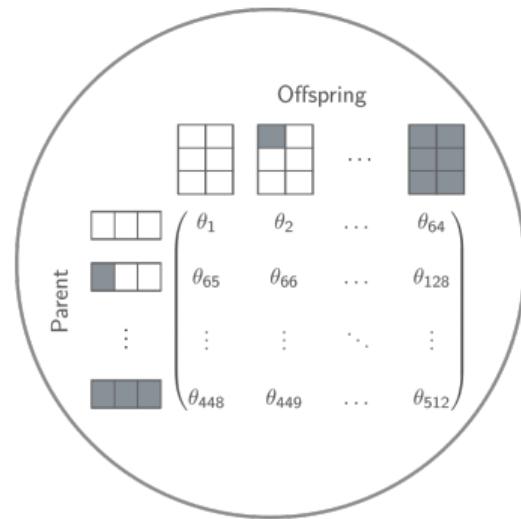
Full Markov Transition Matrix



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

Full Markov Transition Matrix

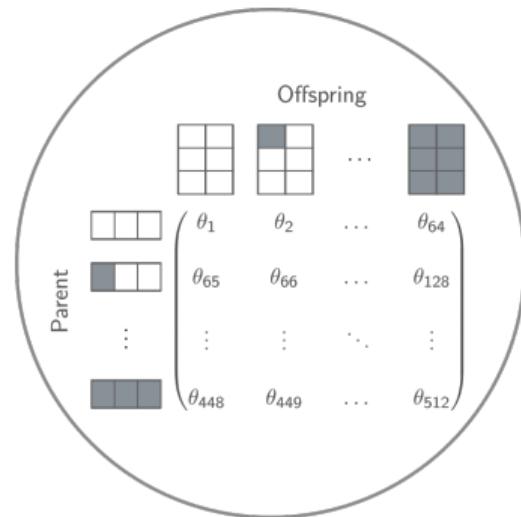
► 512 parameters



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

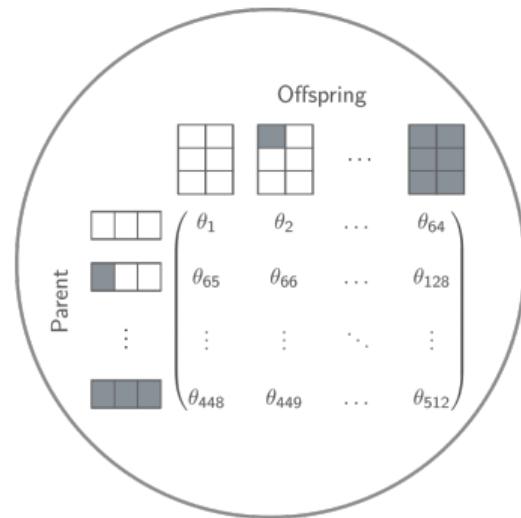
Full Markov Transition Matrix

- ▶ 512 parameters
- ▶ Finding this many parameters not easy.



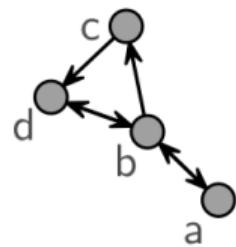
If we wanted to build a model with 3 functions, we would need to estimate...

Full Markov Transition Matrix



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

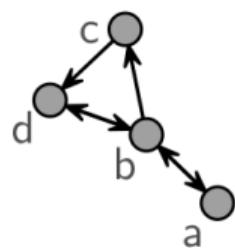
Social Network



	a	b	c	d
a				
b				
c				
d				

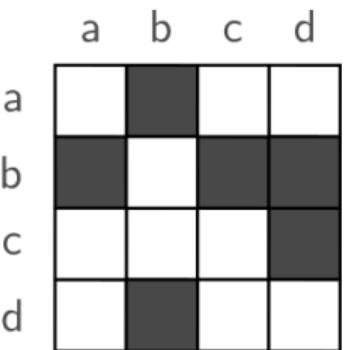
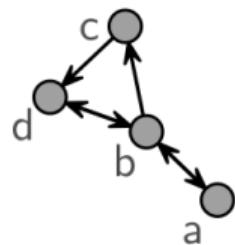
Social Network

► Not about individual ties.

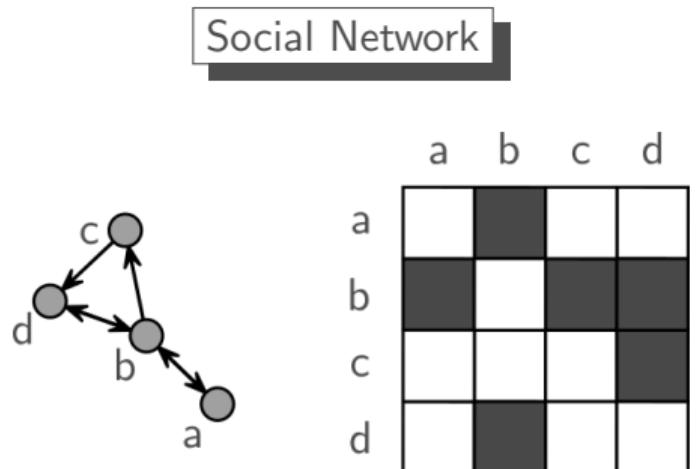


	a	b	c	d
a				
b				
c				
d				

Social Network



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

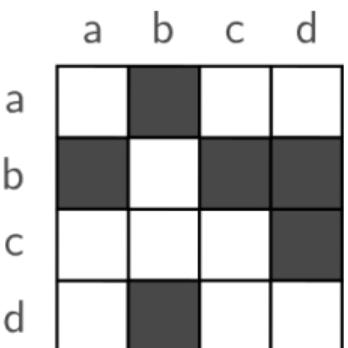
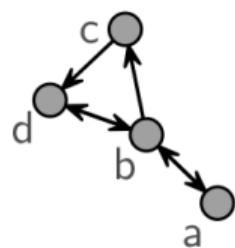


- ▶ Not about individual ties.
- ▶ Statistical inference on *motifs* (triangles, dyads, homophily, etc.)

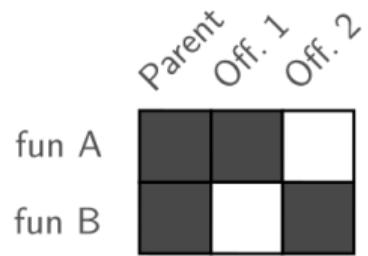
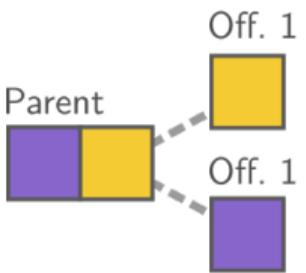
Ultimately...

ERGM \equiv **Modeling binary arrays**

Social Network



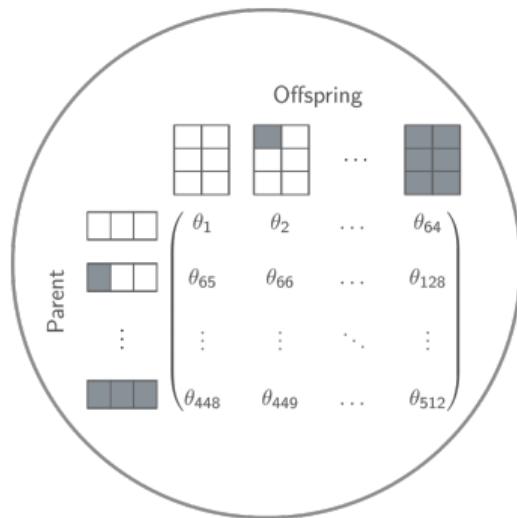
Evolutionary Event



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

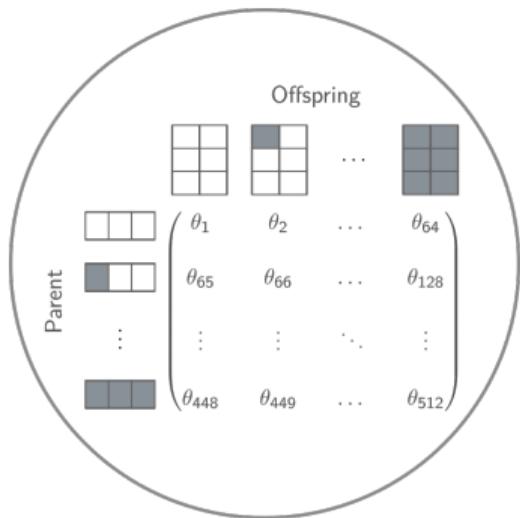
If we wanted to build a model with 3 functions, we would need to estimate...

Full Markov Transition Matrix



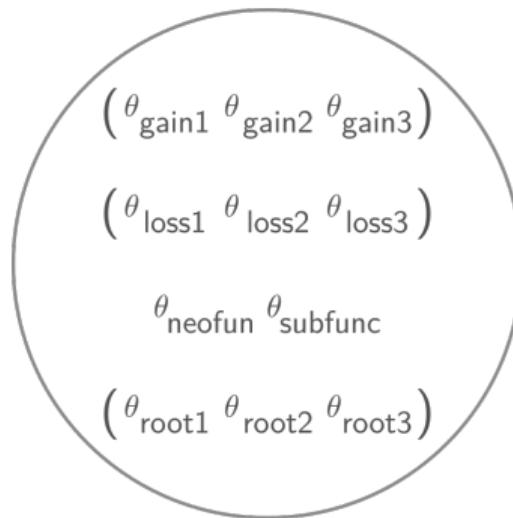
If we wanted to build a model with 3 functions, we would need to estimate...

Full Markov Transition Matrix



512 parameters

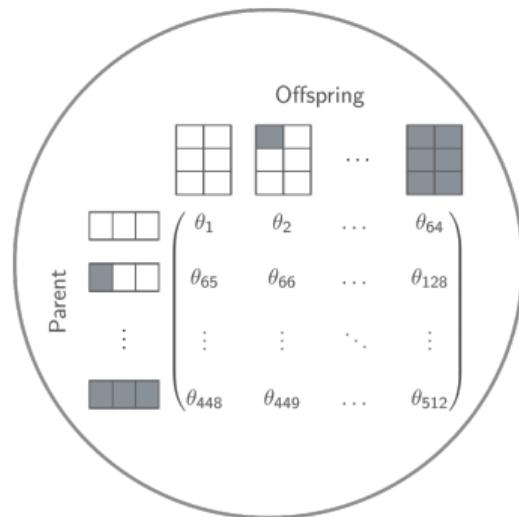
Sufficient statistics



11 parameters (for example)

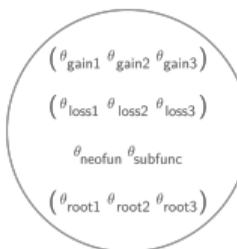
If we wanted to build a model with 3 functions, we would need to estimate...

Full Markov Transition Matrix

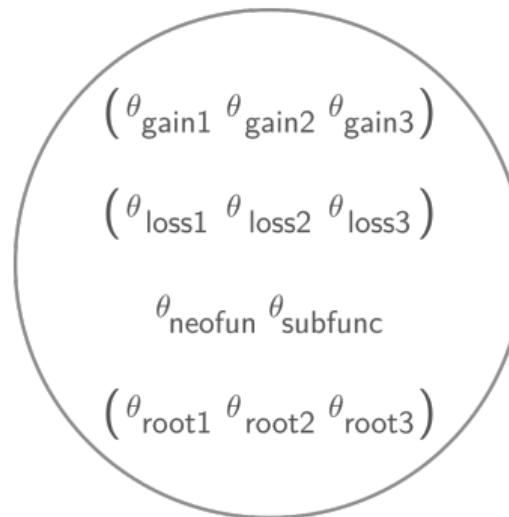


512 parameters

Sufficient statistics



Easier to fit
Easier to interpret



11 parameters (for example)

◀ numeric example

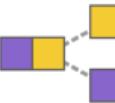
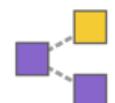
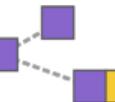
Representation	Description	Definition
	Gain of function	$(1 - x_p) \sum_{n:n \in Off} x_n$
	Loss of function	$x_p \sum_{n:n \in Off} (1 - x_n)$
	Subfunctionalization	$x_p^k x_p^j \sum_{n \neq m} x_n^k (1 - x_n^j) (1 - x_m^k) x_m^j$
	Neofunctionalization	$x_p^k (1 - x_p^j) \sum_{n \neq m} x_n^k (1 - x_n^j)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

Table 1 Example of sufficient statistics for evolutionary transitions.

Tree likelihoods: Felsenstein's Pruning algorithm

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

All possible transitions
from \mathbf{x}_n

Transition Probability
(ERGM)

Tree likelihoods: Felsenstein's Pruning algorithm

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

All possible transitions
from \mathbf{x}_n

Transition Probability
(ERGM)

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

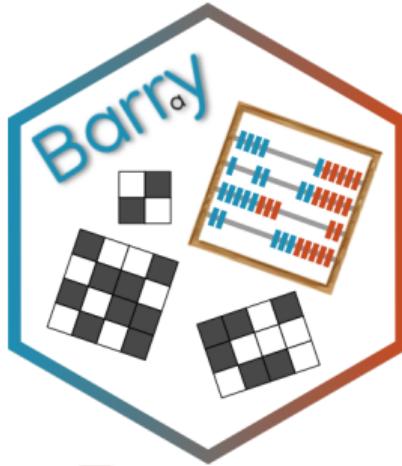
Model
Parameters

Vector of
Sufficient Statistics

Normalizing
Constant

the *lingua franca* of SNA

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



Barry

C++ header-only library for counting structures in binary arrays

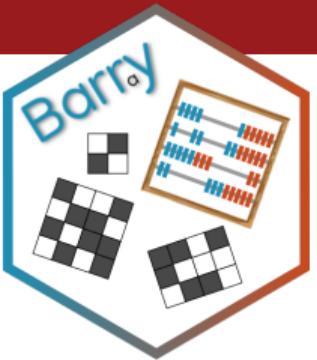
“The Sniffing Accountant” (Seinfeld, Season 5, Episode 4)

Computational features of **barry**



Core

Computational features of **barry**



Core

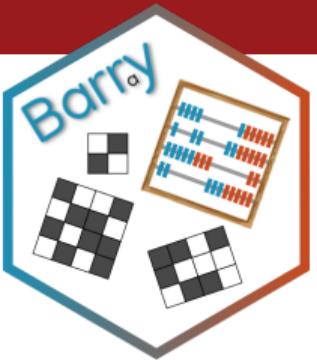
- ▶ C++ header-only template library.

Computational features of **barry**



Core

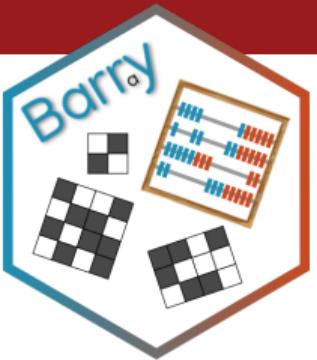
- ▶ C++ header-only template library.
- ▶ Arrays \equiv sparse matrices, i.e., small and large.



Core

- ▶ C++ header-only template library.
- ▶ Arrays \equiv sparse matrices, i.e., small and large.
- ▶ Full enumeration and support of binary arrays.

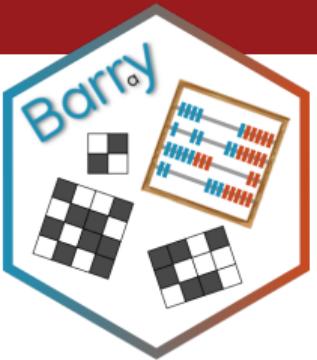
Computational features of **barry**



Core

- ▶ C++ header-only template library.
- ▶ Arrays \equiv sparse matrices, i.e., small and large.
- ▶ Full enumeration and support of binary arrays.
- ▶ Arbitrary constraints for enumeration.

Computational features of **barry**

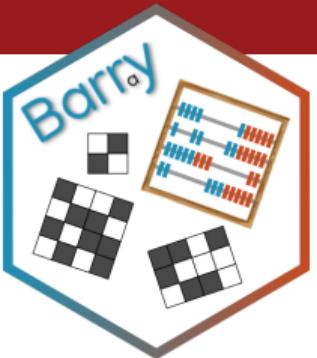


Core

- ▶ C++ header-only template library.
- ▶ Arrays \equiv sparse matrices, i.e., small and large.
- ▶ Full enumeration and support of binary arrays.
- ▶ Arbitrary constraints for enumeration.

Modeling features

Computational features of **barry**



Core

- ▶ C++ header-only template library.
- ▶ Arrays \equiv sparse matrices, i.e., small and large.
- ▶ Full enumeration and support of binary arrays.
- ▶ Arbitrary constraints for enumeration.

Modeling features

- ▶ Arbitrary Model terms (suff. stats).

Computational features of **barry**

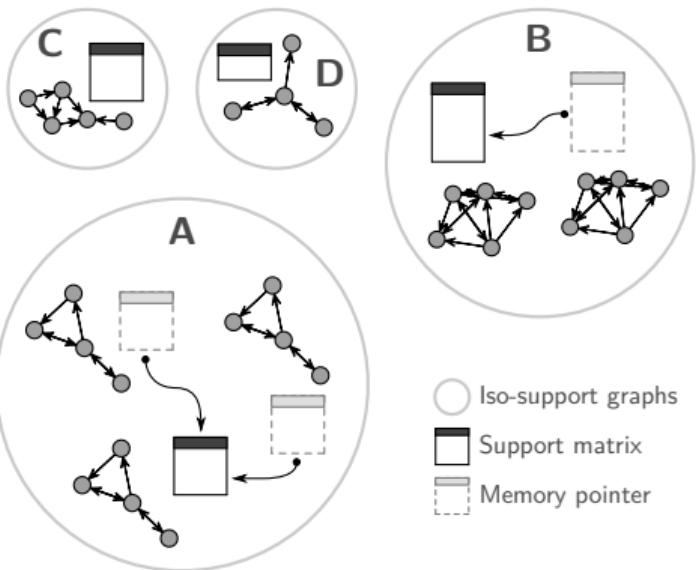


Core

- ▶ C++ header-only template library.
- ▶ Arrays \equiv sparse matrices, i.e., small and large.
- ▶ Full enumeration and support of binary arrays.
- ▶ Arbitrary constraints for enumeration.

Modeling features

- ▶ Arbitrary Model terms (suff. stats).
- ▶ Hashmap recycles support, i.e., pooled data models.



Example: Simple model with two functions

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

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,

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 gain with some prob. at a dupl. event,
4. Function 1 is gain as neofunctionalization (from 0) at a dupl. event,

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 gain with some prob. at a dupl. event,
4. Function 1 is gain as neofunctionalization (from 0) at a dupl. event,
5. There is a higher chance of changes at duplication.
6. There is low chance root node starts off with either 0 or 1.

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 gain with some prob. at a dupl. event,
4. Function 1 is gain as neofunctionalization (from 0) at a dupl. event,
5. There is a higher chance of changes at duplication.
6. There is low chance root node starts off with either 0 or 1.

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

Example: Simple model with two functions

posterior distributions

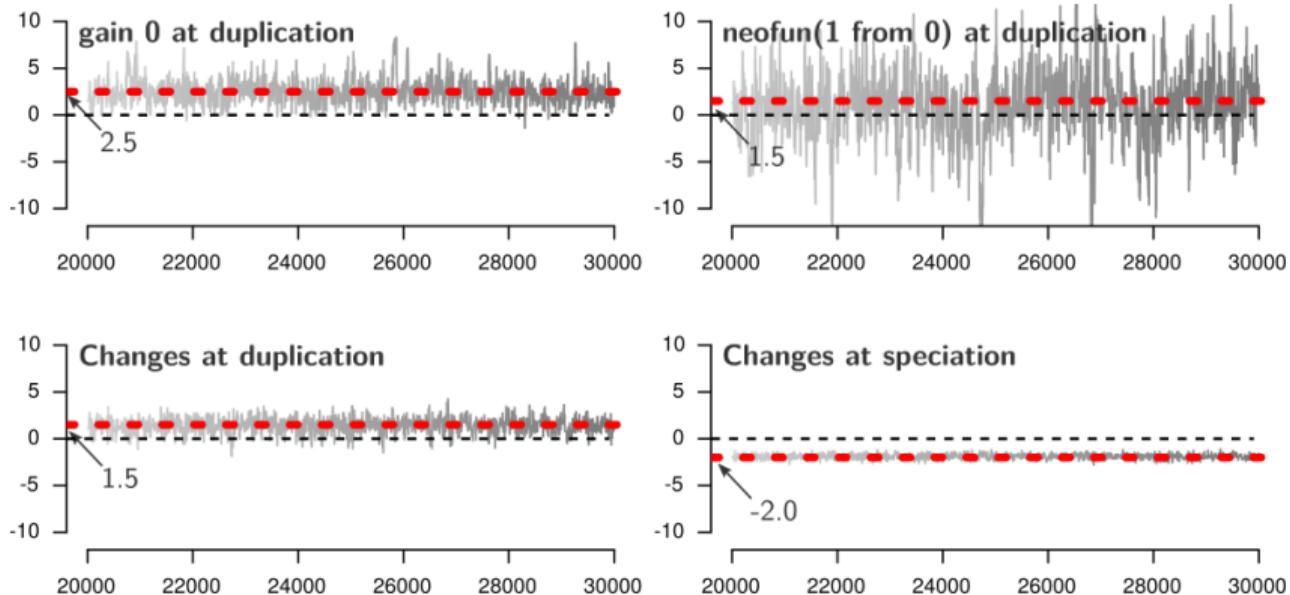


Figure 1 MCMC Trace of the functional gain of 0, neofunctionalization (1 from 0), and change rate (by event type).

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

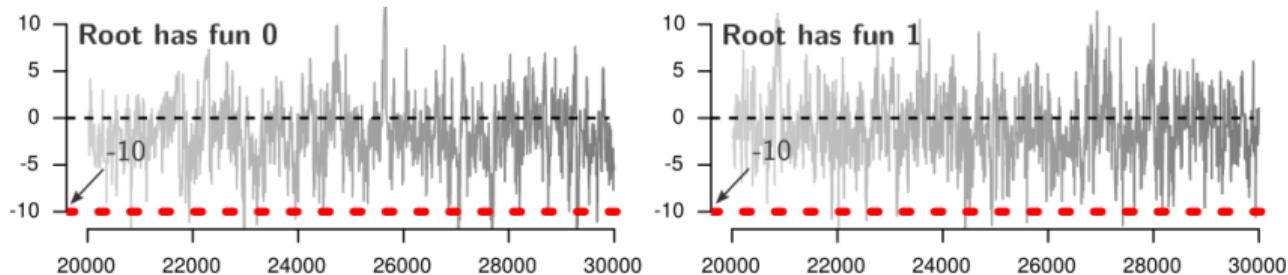


Figure 2 MCMC Trace of root parameters. The true population parameters are $(\theta_{root0}, \theta_{root1}) = (-10.0, -10.0)$.
Root node probabilities are always hard to get.

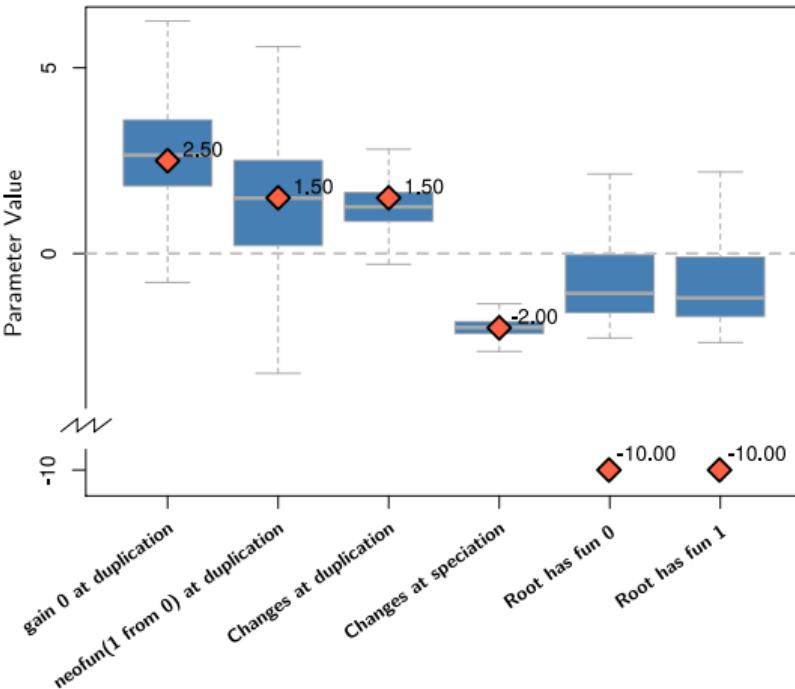
Repeated this experiment 5,000 times:

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

Figure 3 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.



What questions?

With this modeling framework, we could tackle, e.g.,

What questions?

With this modeling framework, we could tackle, e.g.,

- ▶ Potentially improve prediction accuracy.

What questions?

With this modeling framework, we could tackle, e.g.,

- ▶ Potentially improve prediction accuracy.
- ▶ Make inferences of the sort of:

What questions?

With this modeling framework, we could tackle, e.g.,

- ▶ Potentially improve prediction accuracy.
- ▶ Make inferences of the sort of:
"Function A or function B, which came first?"

What questions?

With this modeling framework, we could tackle, e.g.,

- ▶ Potentially improve prediction accuracy.
- ▶ Make inferences of the sort of:
 - "Function A or function B, which came first?"
 - "When was subfunctionalization more likely to happen?"

What questions?

With this modeling framework, we could tackle, e.g.,

- ▶ Potentially improve prediction accuracy.
- ▶ Make inferences of the sort of:
 - "Function A or function B, which came first?"
 - "When was subfunctionalization more likely to happen?"
 - "Where functions A and B gained at the same time?"

What questions?

With this modeling framework, we could tackle, e.g.,

- ▶ Potentially improve prediction accuracy.
- ▶ Make inferences of the sort of:
 - "Function A or function B, which came first?"
 - "When was subfunctionalization more likely to happen?"
 - "Where functions A and B gained at the same time?"
- ▶ and much more...

Take home

Goal

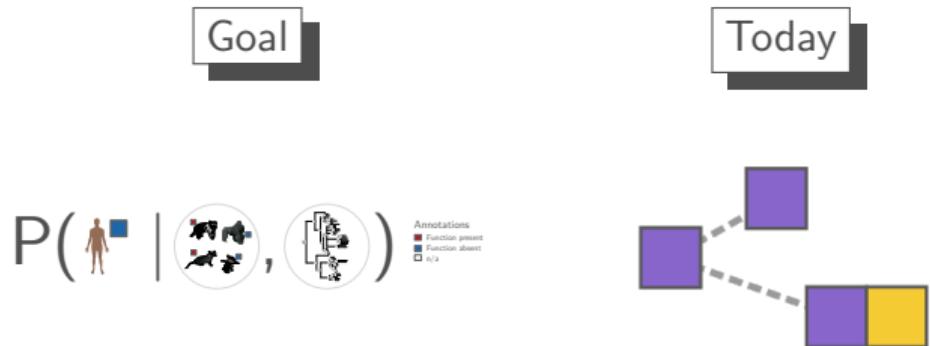
$$P(\text{ } | \text{ } , \text{ })$$

Annotations

- Function present
- Function absent
- n/a



- ▶ We are in a race for uncovering **what genes do.**
- ▶ **Automatic algorithms** provide a way.



- ▶ We are in a race for uncovering **what genes do.**
- ▶ **Automatic algorithms** provide a way.
- ▶ Many alternatives... many unrealistic **assumptions**.
- ▶ **barry** (Sufficient Stats.): a general framework for gene function evolution.

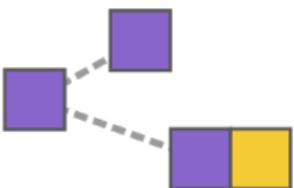
Goal

$$P(\text{ } | \text{ } , \text{ })$$

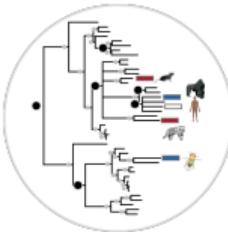
Annotations

- Function present
- Function absent
- n/a

Today



Next steps



- ▶ We are in a race for uncovering **what genes do**.
- ▶ **Automatic algorithms** provide a way.
- ▶ Many alternatives... many unrealistic **assumptions**.
- ▶ **barry** (Sufficient Stats.): a general framework for gene function evolution.
- ▶ Further study its properties (bias, power, accuracy).

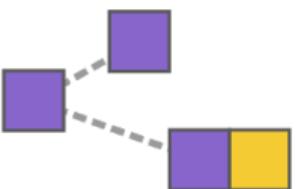
Goal

$$P(\text{ } | \text{ } , \text{ })$$

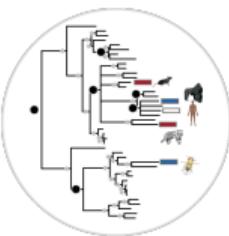
Annotations

- Function present
- Function absent
- n/a

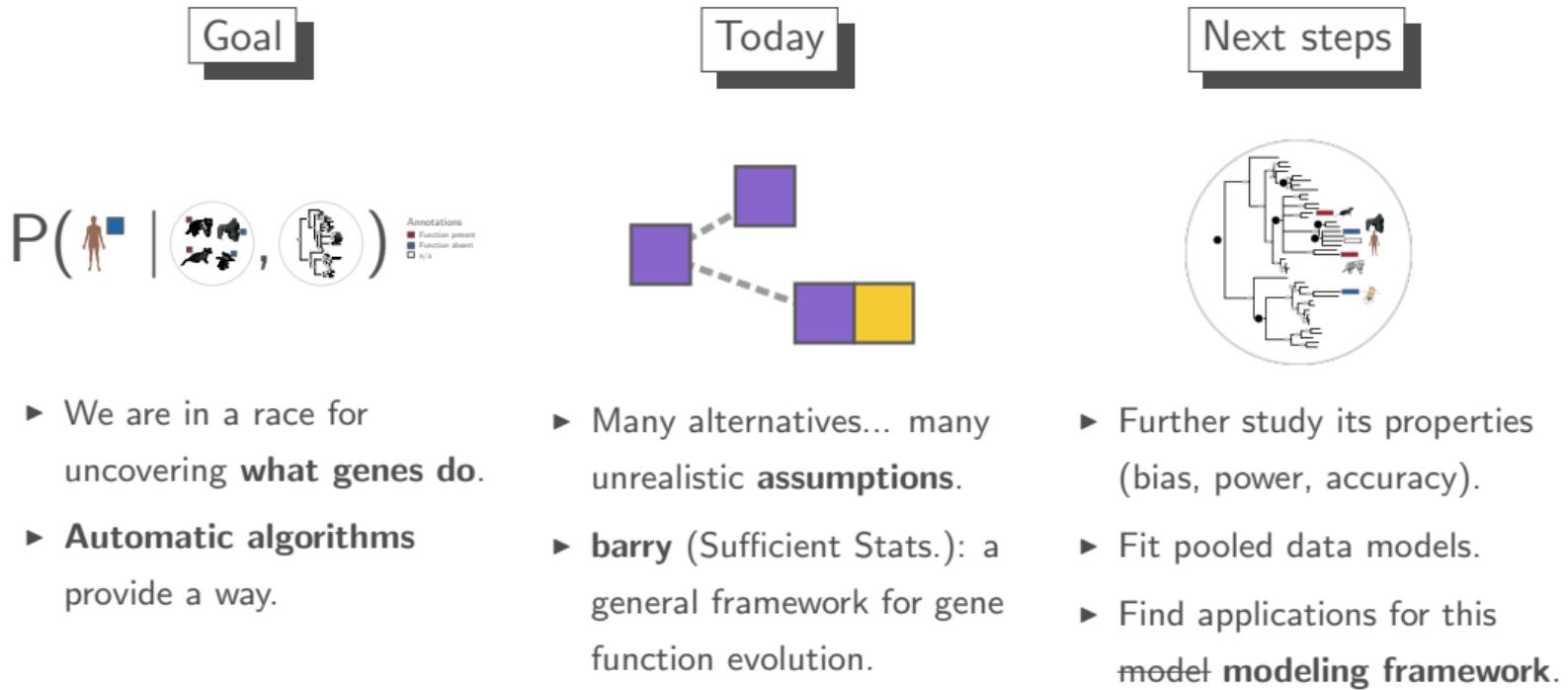
Today



Next steps



- ▶ We are in a race for uncovering **what genes do**.
- ▶ **Automatic algorithms** provide a way.
- ▶ Many alternatives... many unrealistic **assumptions**.
- ▶ **barry** (Sufficient Stats.): a general framework for gene function evolution.
- ▶ Further study its properties (bias, power, accuracy).
- ▶ Fit pooled data models.



Triads, Dyads, and Gene Functions

When Social Network Analysis Meets Phylogenetics

George G Vega Yon

<https://ggyv.cl>

vegayon@usc.edu

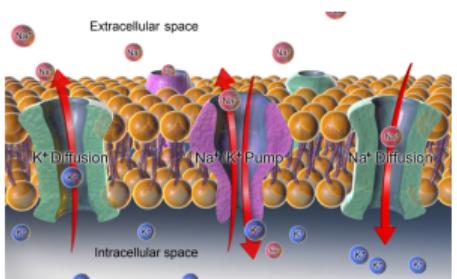


Thank you!

Gene functions can be classified in three types:

Molecular function

Active transport GO:0005215



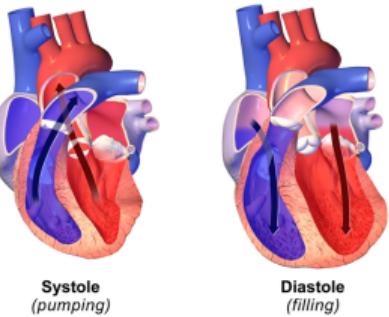
Cellular component

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 2 Heart Contraction Function. source: amigo.geneontology.org

You know what is interesting about this function?

◀ go back

These four species have a gene with that function... and two of these are part of the same evolutionary tree!



Felis catus pthr10037



Oryzias latipes pthr11521



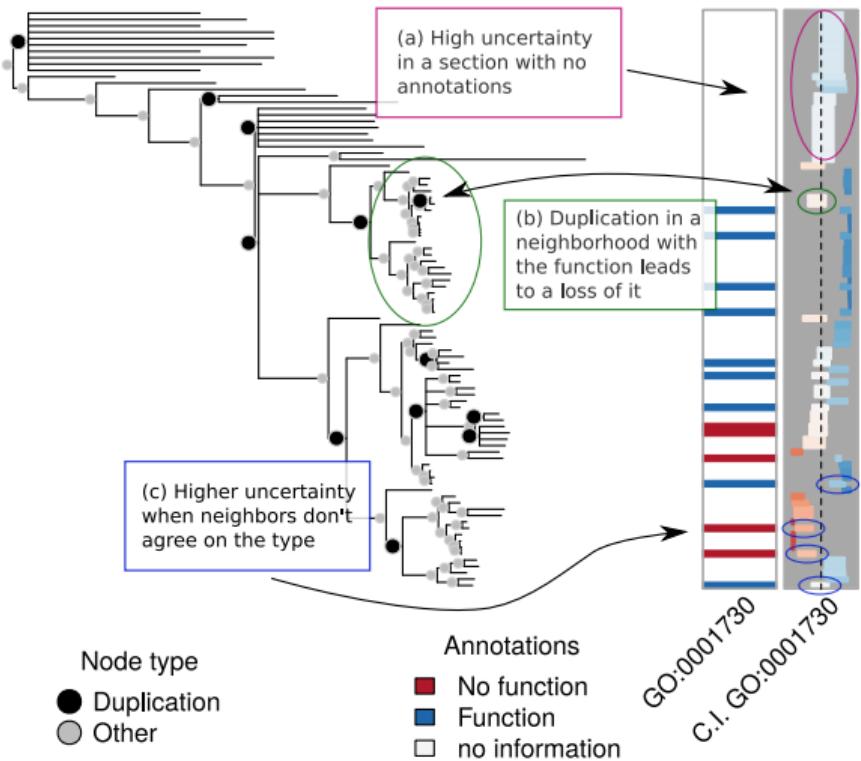
Anolis carolinensis pthr11521



Equus caballus pthr24356

[◀ go back](#)

Example of Data + Predictions

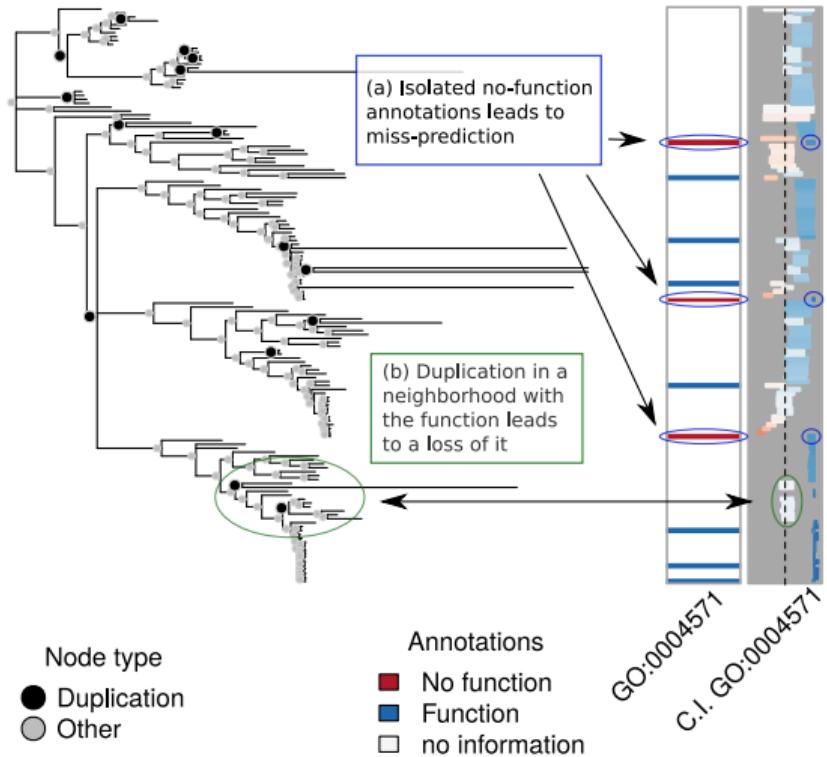
Family: PTHR11258**Type:** Molecular Function**Name:** 2'-5'-oligoadenylate synthetase activity**Desc:** GO:0001730 involved in the process of cellular antiviral activity (wiki on [interferon](#)).**MAE:** 0.34**AUC:** 0.91[see a bad one](#)[◀ go back](#)

Example 2: Bad quality prediction

MAE: 0.52

AUC: 0.33

Type: Molecular Function

Name: mannosyl-oligosaccharide
1,2-alpha-mannosidase activityDesc: GO:0004571 involved in
synthesis of glycoproteins ([wiki](#)
and [examples](#)).[◀ go back](#)

		Pooled-data	One-at-a-time	
		Beta prior	Unif. prior	Beta Prior
Pooled-data				
Unif. prior	Beta prior	[-0.02,-0.01]	[-0.14,-0.10]	[-0.06,-0.03]
	Beta prior	-	[-0.12,-0.09]	[-0.04,-0.01]
One-at-a-time				
Unif. prior	Beta prior	-	-	[0.06, 0.09]

Table 3 Differences in Mean Absolute Error [MAE]. Each cell shows the 95% confidence interval for the difference in MAE resulting from two methods (row method minus column method). Cells are color coded blue when the method on that row has a significantly smaller MAE than the method on that column; Conversely, cells are colored red when the method in that column outperforms the method in that row. Overall, predictions calculated using the parameter estimates from *pooled-data* predictions outperform *one-at-a-time*.

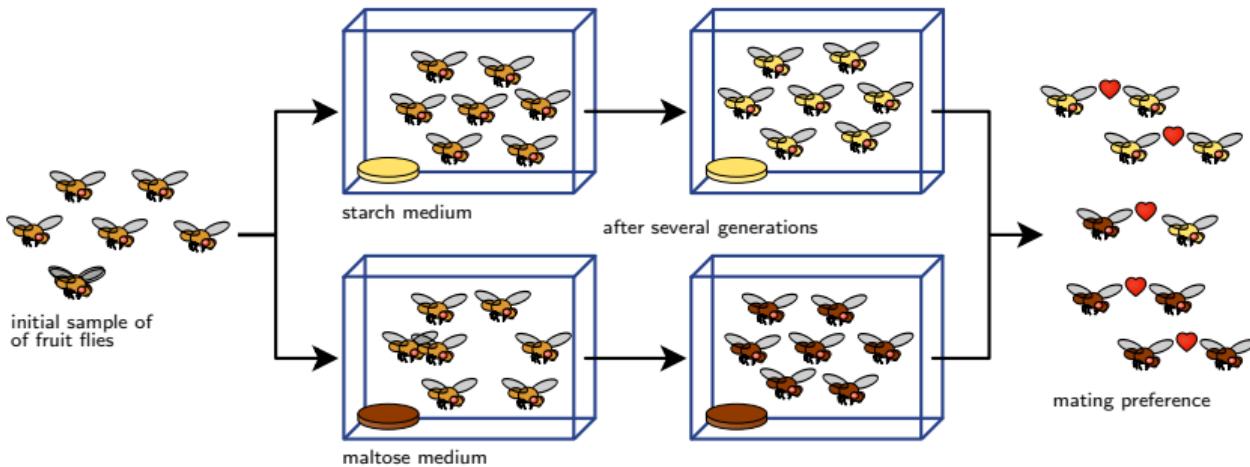


Figure 4 Dodd (1989): After one year of isolation, flies showed a significant level of assortativity in mating (wikimedia)

◀ go back

Duplication

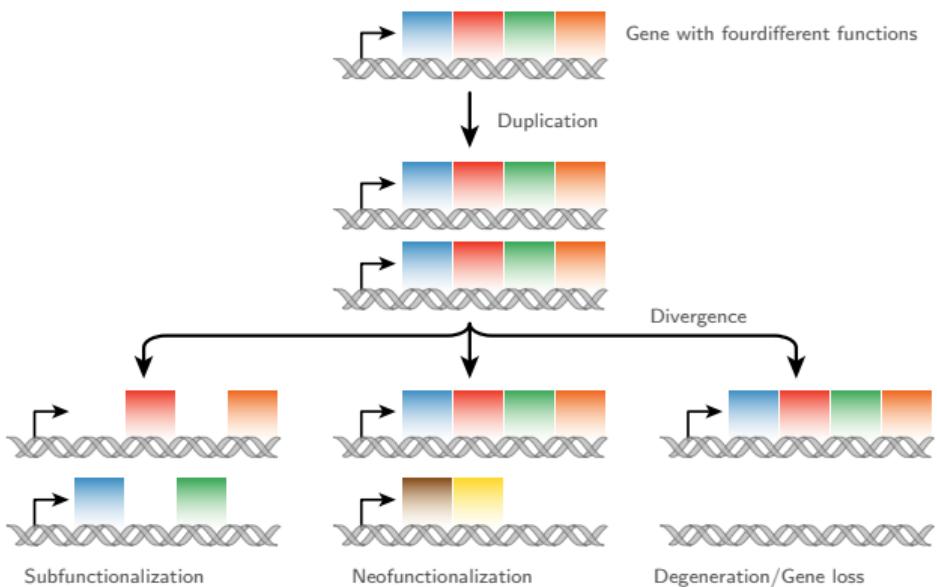


Figure 5 A key part of molecular innovation, gene duplication provides opportunity for new functions to emerge
(wikimedia)

◀ go back

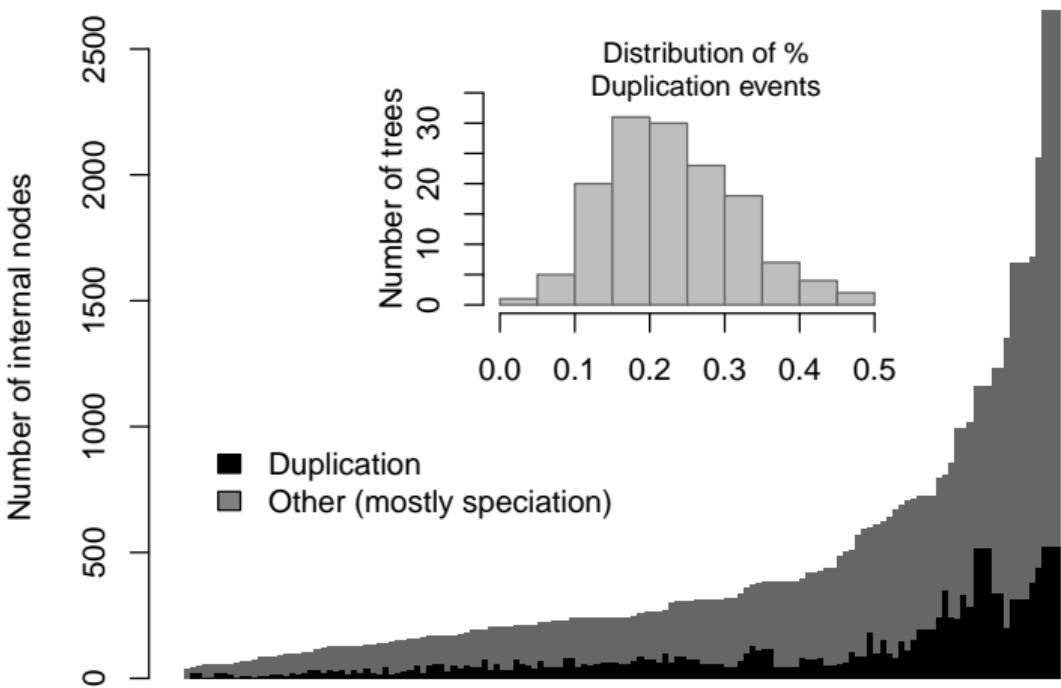
Data: Phylogenetic trees

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](#)

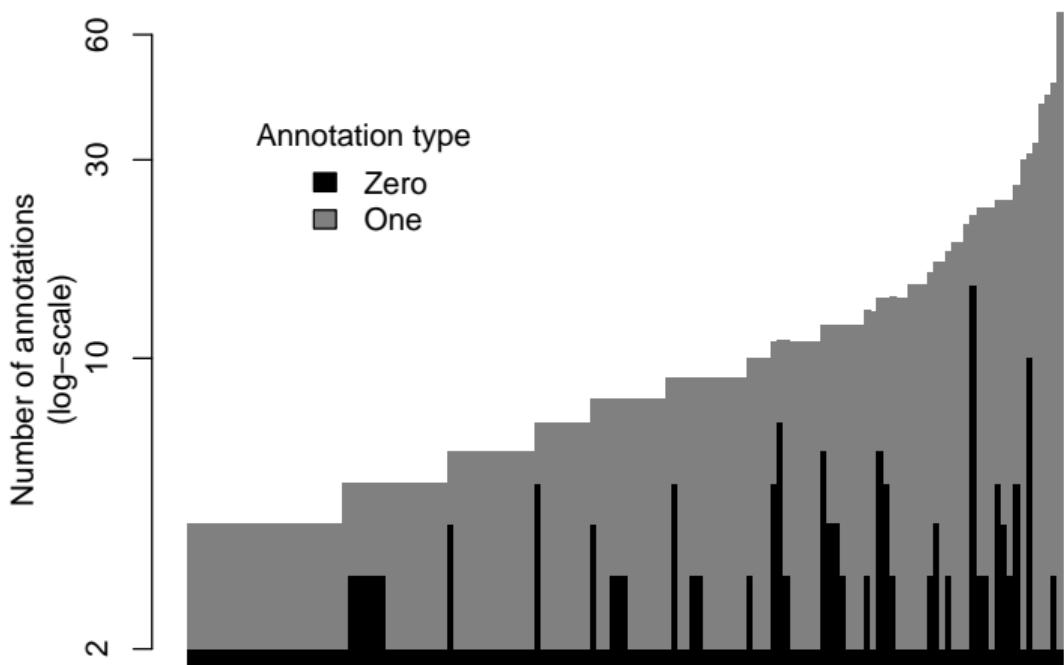
Data: Annotations (example)

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	

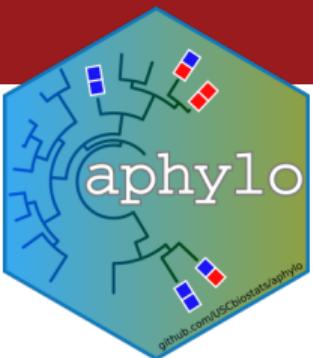
◀ go back

Data: Experimental Annotations



◀ go back

Computational features of aphylo

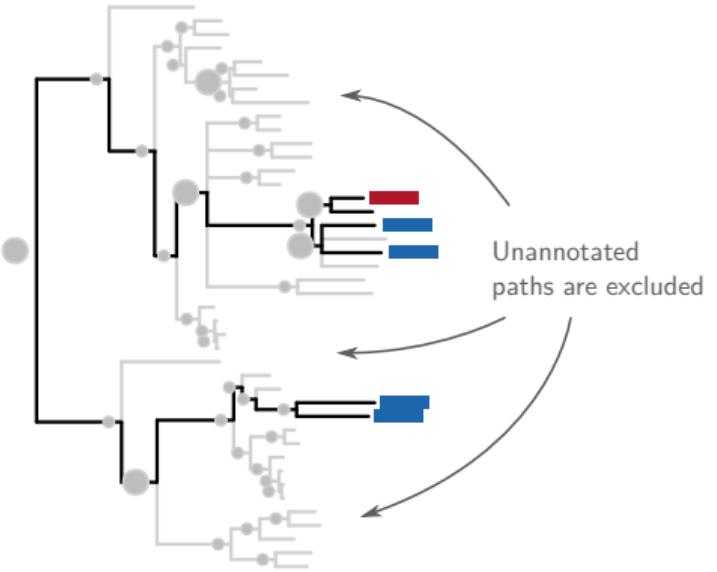


Baseline features

- ▶ Parsimony: Conditional independence across functions/siblings.
- ▶ Post-order Tree traversal: Linear complexity $O(|\text{tree}|)$.

Additional features

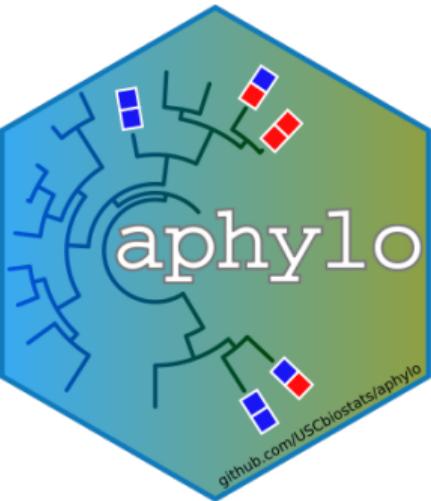
- ▶ Reduced pruning sequence: Induced sub-tree of nodes connected to annotated leafs
 \implies Complexity $O(|\text{Induced sub-tree}|) \leq O(|\text{tree}|)$
- ▶ Implemented in C++ (**pruner** library)



◀ return

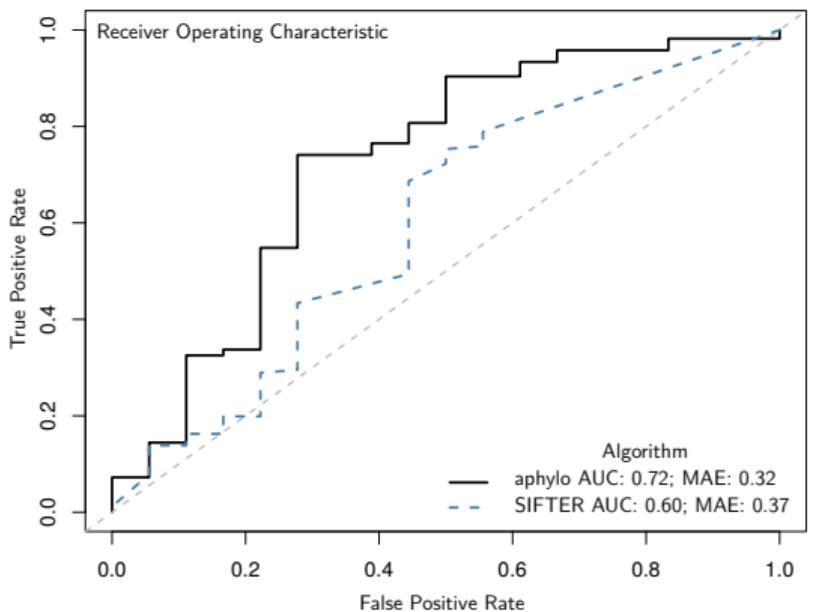
Results: Implementation and Large scale study

- ▶ Simulation, estimation, and prediction: **aphylo** R package.
- ▶ Large simulation study (all known trees, about 15,000) on USC's HPC cluster.
- ▶ Prediction quality assessment on $\sim 1,300$ genes involving ~ 130 families... estimation of parameters using a pooled-data model (< 5 min). [◀ modeling](#) [◀ estimates](#)
- ▶ In a subset of ~ 200 predictions we found 46 novel annotations

[▶ more](#)[◀ go back](#)

Results: Performance and Scalability

aphylo vs SIFTER (state-of-the-art phylo-based model) on 147 genes.



Fast 110 minutes (SIFTER) to calculate the posterior probabilities, aphylo took 1 second.

Accurate aphylo reported higher accuracy levels in LOO cross-validation (0.72 vs 0.60 AUC).

Overview of Prediction Results

	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

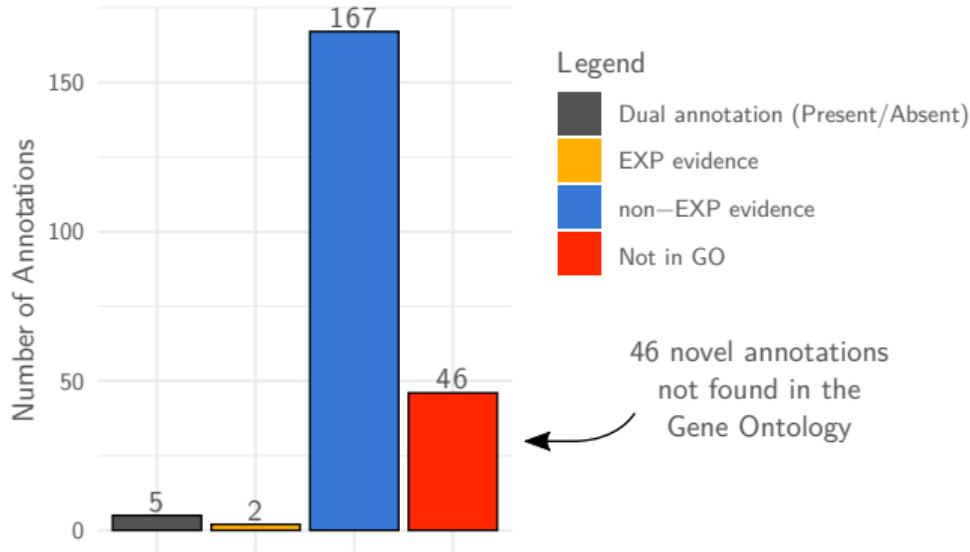


Figure 6 Distribution of predictions

◀ go back

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

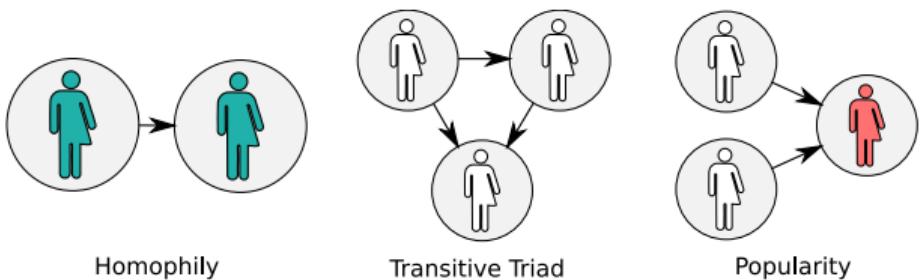
# Gains	1	1
Only one offspring changes (yes/no)	1	0
# Changes (gain+loss)	2	3
Subfunctionalizations (yes/no)	0	1

▶ return

What are Exponential Random Graph Models

Exponential Family Random Graph Models, aka **ERGMs** are:

- ▶ Statistical models of (social) networks.
- ▶ Social Network Analysis: What drives social connections?
- ▶ Not about individual ties, but about local structures (sufficient statistics).



- ▶ Social Networks \equiv Adjacency Matrix \equiv Binary arrays

What Drives Evolution: a game changer

In the model with 3 functions and 2 offspring per node:

- ▶ Full Markov transition matrix: $2^3 \times 2^6 = 512$
- ▶ Using sufficient statistics:

Pairwise co-evolution: 3 terms,

Pairwise Neofunctionalization: 3 terms,

Pairwise Subfunctionalization: 3 terms,

Function specific gain: 3 terms,

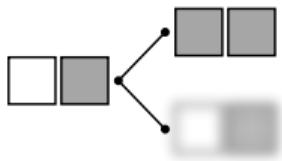
Function specific loss: 3 terms,

Total: 15 parameters.

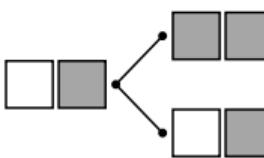
- ▶ Easier to fit and interpret.



(a) Sibling and Function Conditional Independence



(b) Sibling Conditional Independence



(c) No conditional independence