Project 2: Augmenting functional information about human genes using probabilistic phylogenetic modeling

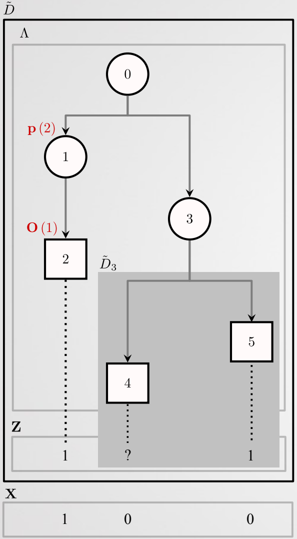
George G. Vega Yon
vegayon@usc.edu

Duncan Thomas   Paul D. Thomas   Paul Marjoram   Huaiyu Mi   John Morrison

Department of Preventive Medicine
University of Southern California

November 14th, 2018

Notation



Symbol	Description
$\Lambda \equiv (\mathcal{N}, \mathcal{E})$	Phylogenetic Tree.
$\mathbf{p}(n)$	Parent of node n.
$\mathbf{O}(n)$	Offspring of node n.
$\mathbf{X} \equiv \{x_n\}_{n\in\mathcal{N}}$	True annotations.
$\mathbf{Z} \equiv \{z_n\}_{n\in\mathcal{N}}$	Experimental annotations.
$D \equiv (\Lambda, \mathbf{X})$	Annotated Phylogenetic Tree.
$\tilde{D} \equiv (\Lambda, \mathbf{Z})$	Experimentally Annotated Phylogenetic Tree.
$\tilde{D}_n$	Induced Experimentally Annotated Subtree of node n.
$\tilde{D}_n^c$	Complement of $\tilde{D}_n$.

Table 1: Mathematical Notation

Recap: Model

1. A probabilistic model of gene function evolution,
2. The probability that the root node has the function is π,
3. Conditional on its parent state, the probabilities that any given node has to either gain or lose a function are (μ_{01}, μ_{10}),
4. Finally, at the leaf node, the probability that a node with no function is mislabeled as having the function is ψ_{01}. Conversely, the probability that a node with a function is mislabeled as not having the function is ψ_{10}.

Parameter	Probability
π	The root node has the function
μ_{01}	Gaining a function
μ_{10}	losing a function
ψ_{01}	Mislabeling a 0
ψ_{10}	Mislabeling a 1

Table 2: Model parameters

Recap: Model

1. A probabilistic model of gene function evolution,
2. The probability that the root node has the function is π,
3. Conditional on its parent state, the probabilities that any given node has to either gain or lose a function are (μ_{01}, μ_{10}),
4. ~~Finally~~, at the leaf node, the probability that a node with no function is mislabeled as having the function is ψ_{01}. Conversely, the probability that a node with a function is mislabeled as not having the function is ψ_{10}.
5. Finally, curators will report their discovery of function *present*/*absent* with probability η_0/η_1.

Parameter	Probability
π	The root node has the function
μ_{01}	Gaining a function
μ_{10}	losing a function
ψ_{01}	Mislabeling a 0
ψ_{10}	Mislabeling a 1
η_0	Propensity to report a 0
η_1	Propensity to report a 1

Table 2: Model parameters

Changes from last year

From the formal (statistical) stand point

- Prediction function: Right mathematical definition of the model prediction.
- New set of parameters: Propensity to report a finding.
- Flexible model specification: Definition of the likelihood function for different sets of parameters

By-products generated during the implementation

- The `sluRm` R package: A light-weight interface to slurm.
- Improvements on the `amcmc` R package, notably: automatic stop.

Recap: The aphylo R package

Features:

- Provides a representation of *annotated* partially ordered trees.

Recap: The aphylo R package

Features:

- Provides a representation of *annotated* partially ordered trees.
- Integrates the ape package (most used Phylogenetics R package with ~25K downloads/month)

Recap: The aphylo R package

Features:

- Provides a representation of *annotated* partially ordered trees.
- Integrates the ape package (most used Phylogenetics R package with $\sim$25K downloads/month)
- Implements the loglikelihood calculation of our model (with C++ under-the-hood).

Recap: The aphylo R package

Features:

- Provides a representation of *annotated* partially ordered trees.
- Integrates the ape package (most used Phylogenetics R package with $\sim$25K downloads/month)
- Implements the loglikelihood calculation of our model (with C++ under-the-hood).

Some new features

- Model specification via formula.
- Added the propensity to report discovery parameters.
- Two implementations of the prediction function (using a post-order algorithm as suggested by Prof. Suchard), and a brute force method... we use this for unit tests.
- (in the `amcmc` R package) Convergence monitoring and automatic stop of the MCMC algorithm

Nice visualizations

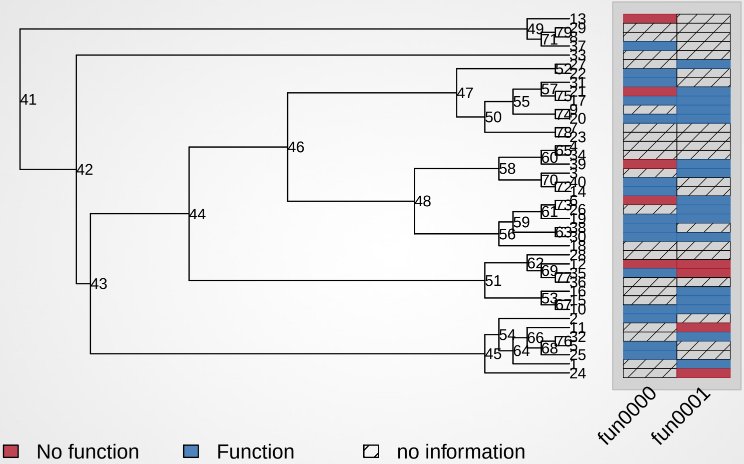


Figure 1: Annotated Phylogenetic Tree

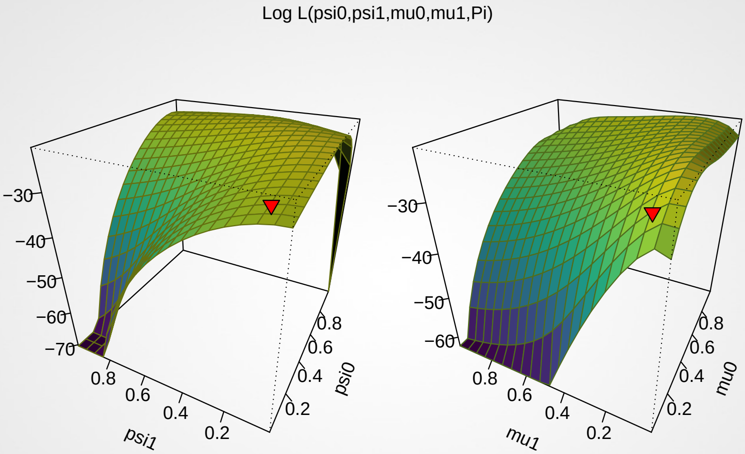


Figure 2: Surface of the likelihood of a given annotated tree.

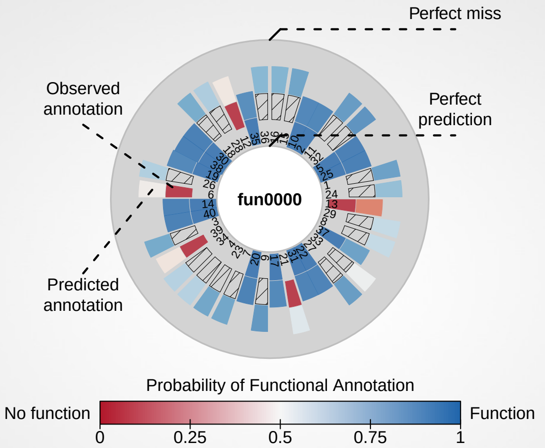


Figure 3: Prediction Accuracy: Observed versus predicted values

Flexible model specification

Automatic specification of the likelihood function, e.g.

- `x ~ mu` baseline model
- `x ~ mu + psi + Pi` model including mislabeling and root node probabilities
- `x ~ mu + Pi` same as before, but excluding mislabeling
- `x ~ mu + psi(1) + Pi` mislabeling of 1 is fixed
- `x ~ mu + psi(0, 1) + Pi` mislabeling of 0s and 1s is fixed

Flexible model specification

```
##
## ESTIMATION OF ANNOTATED PHYLOGENETIC TREE
##
##  Call: aphylo_mcmc(model = x ~ mu + psi + Pi, priors = bprior())
##  ll: -15.1028 ,
##  Method used: mcmc (748 iterations)
##  Leafs:
##  # of Functions 2
##          Estimate  Std. Err.
##  psi0    0.0998    0.0782
##  psi1    0.0955    0.0679
##  mu0     0.2379    0.0902
##  mu1     0.0499    0.0379
##  Pi      0.0888    0.0781
```

Simulation study

Using the panther dataset, we applied our model's data generating process to annotate trees

Simulation study

Using the panther dataset, we applied our model's data generating process to annotate trees

Four different scenarios:

Simulation study

Using the panther dataset, we applied our model's data generating process to annotate trees

Four different scenarios:

1. Gold standard: Estimation of the model on fully annotated trees

Simulation study

Using the panther dataset, we applied our model's data generating process to annotate trees

Four different scenarios:

1. Gold standard: Estimation of the model on fully annotated trees
2. Missing data: Estimation of the model with missing annotations [from 10% to 90% missigness]

Simulation study

Using the panther dataset, we applied our model's data generating process to annotate trees

Four different scenarios:

1. Gold standard: Estimation of the model on fully annotated trees
2. Missing data: Estimation of the model with missing annotations [from 10% to 90% missigness]
3. Propensity to report (a): Same data as scenario 2, but we drop more observations with probabilities η_0, η_1. Estimation does not include η.

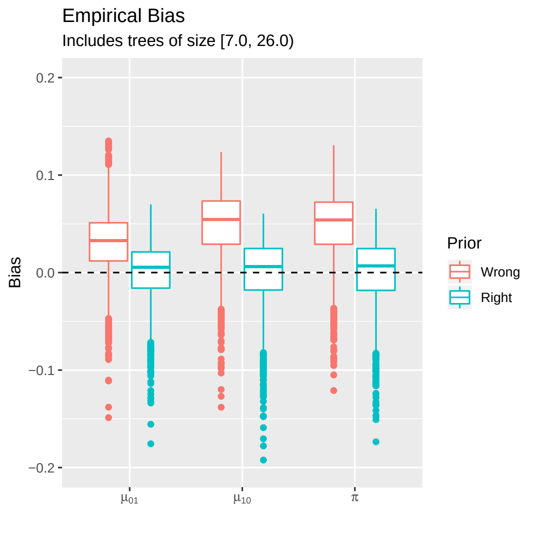
Simulation study

Using the panther dataset, we applied our model's data generating process to annotate trees

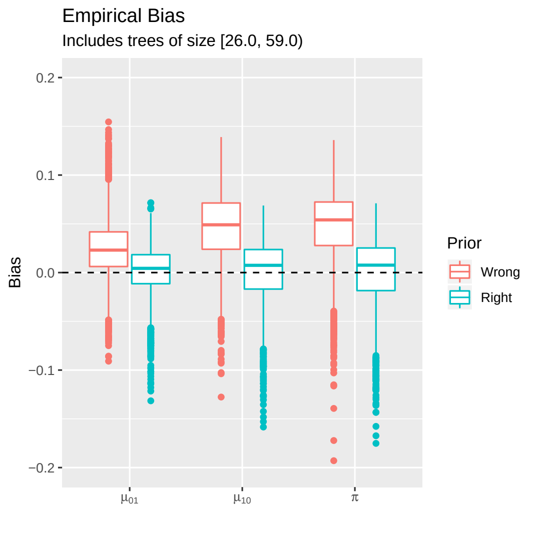
Four different scenarios:

1. Gold standard: Estimation of the model on fully annotated trees
2. Missing data: Estimation of the model with missing annotations [from 10% to 90% missigness]
3. Propensity to report (a): Same data as scenario 2, but we drop more observations with probabilities η_0, η_1. Estimation does not include η.
4. Propensity to report (b): Sames as scenario 3, but we include η.

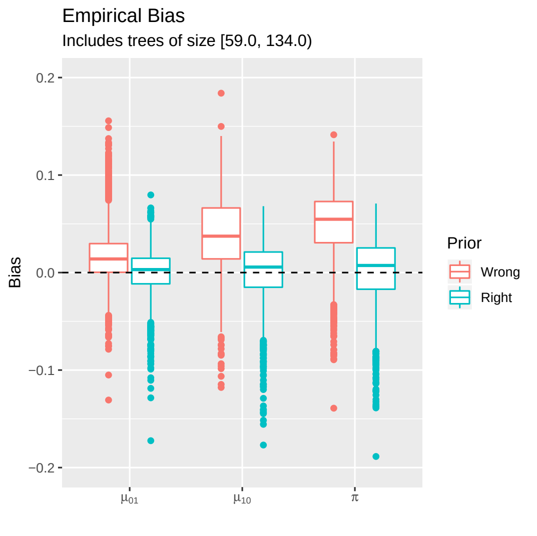
Gold standard: Bias (small trees)



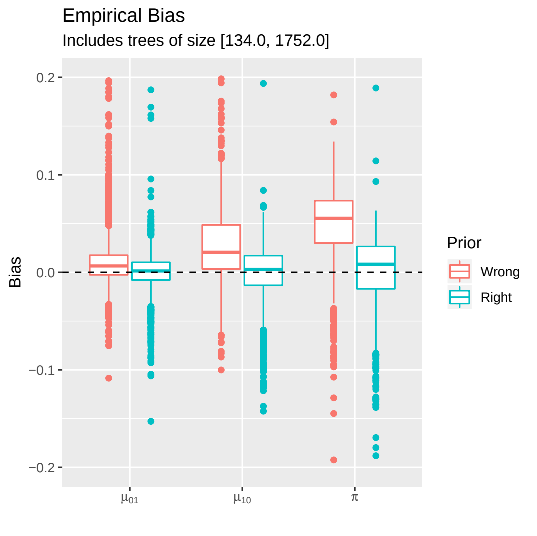
Gold standard: Bias (mid-small trees)



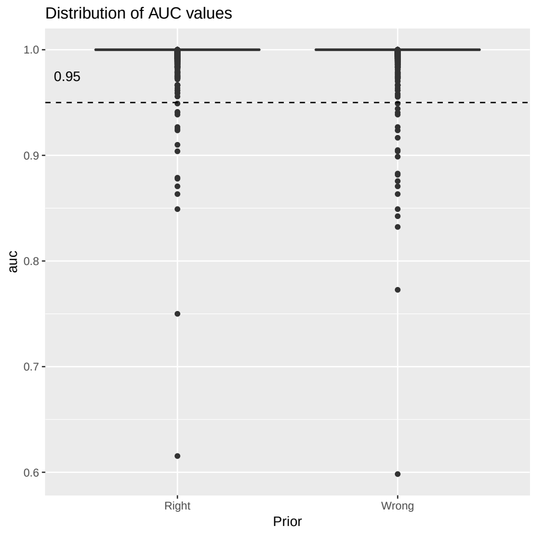
Gold standard: Bias (mid-large trees)



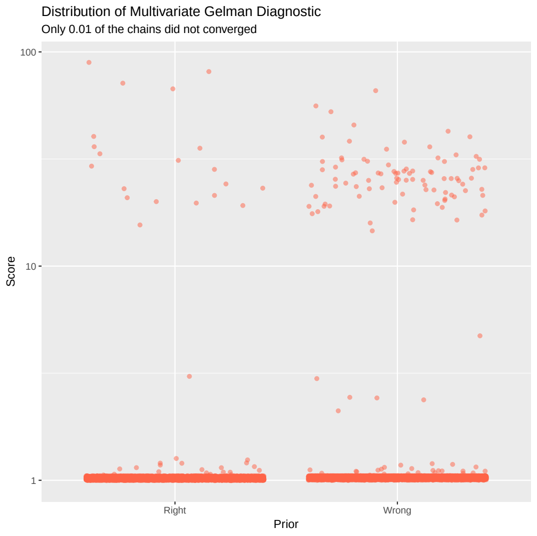
Gold standard: Bias (large trees)



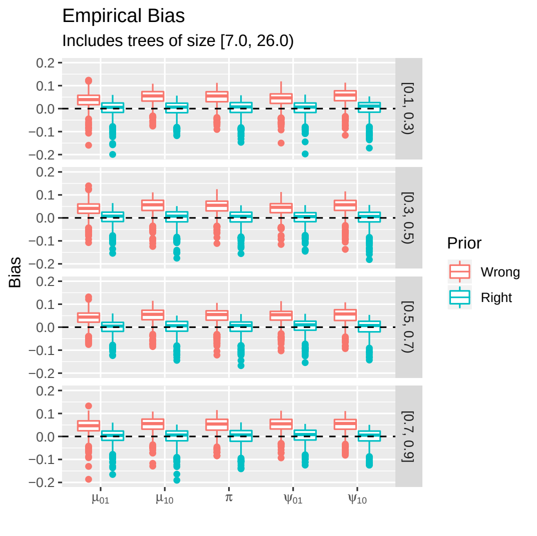
Gold standard: Prediction



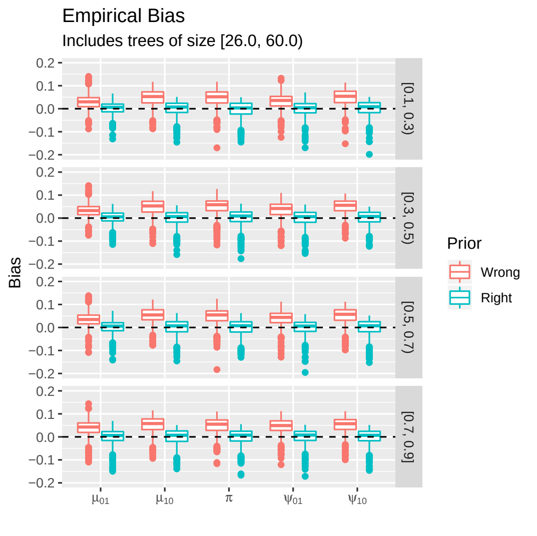
Gold standard: Convergence



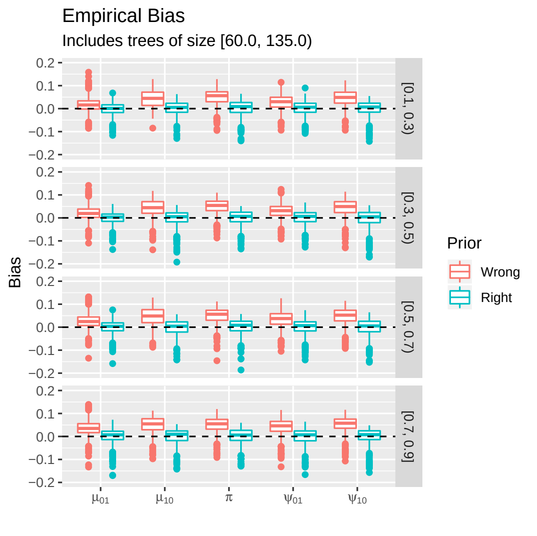
Missing data: Bias (small trees)



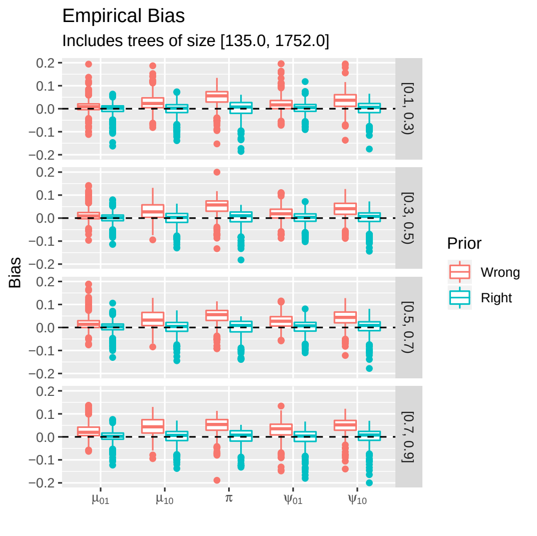
Missing data: Bias (mid-small trees)



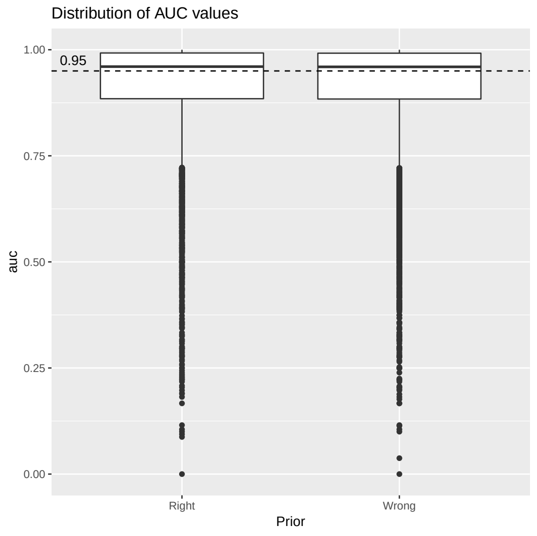
Missing data: Bias (mid-large trees)



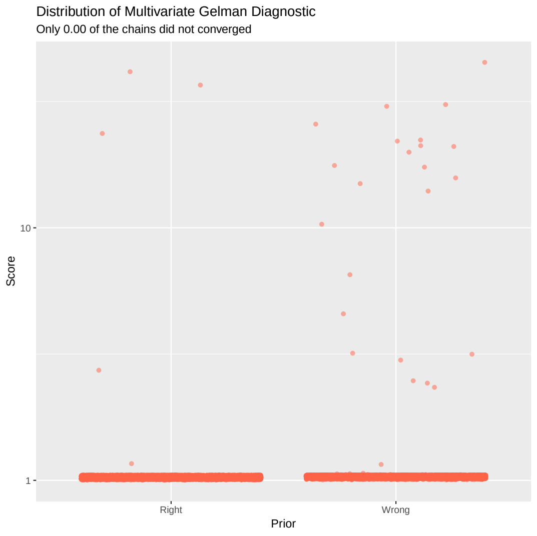
Missing data: Bias (large trees)



Missing data: Prediction



Missing data: Convergence



Does η improves the model? Prediction

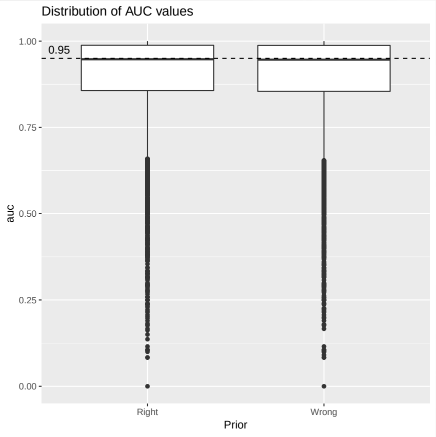


Figure 4: Misspecified model (does not include η)

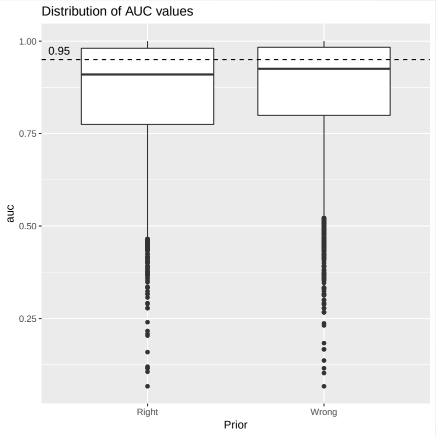
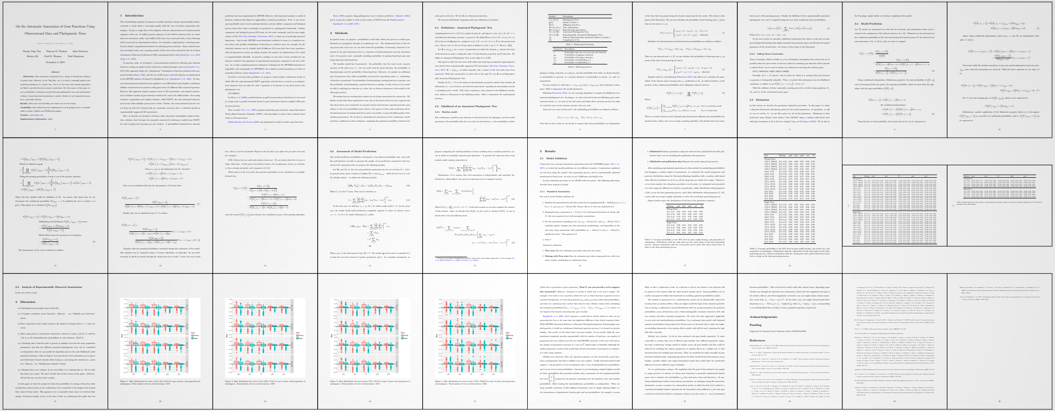


Figure 5: Correct specification (includes η)

Status of the paper



On the Automatic Annotation of Gene Functions Using Observational Data and Phylogenetic Trees

George Vega Yon Duncan D. Thomas John Morrison Huaiyu Mi Paul D. Thomas Paul Marjoram

November 8, 2018

Concluding remarks

A parsimonious model of gene functions: easy to apply on a large scale (we already ran some simulations using all 13,000 trees from PantherDB. . . and it took us less than 1 ~~week~~ hour with ~~10~~ 240 processors ~~only~~).

Concluding remarks

A parsimonious model of gene functions: easy to apply on a large scale (we already ran some simulations using all 13,000 trees from PantherDB… and it took us less than 1 ~~week~~ hour with ~~10~~ 240 processors ~~only~~).

- Already implemented, we are currently in the stage of ~~writing the paper and setting up the simulation study~~ finishing and submiting the paper.

Concluding remarks

A parsimonious model of gene functions: easy to apply on a large scale (we already ran some simulations using all 13,000 trees from PantherDB... and it took us less than 1 ~~week~~ hour with ~~10~~ 240 processors ~~only~~).

- Already implemented, we are currently in the stage of ~~writing the paper and setting up the simulation study~~ finishing and submiting the paper.
- For the next steps, we are evaluating whether to include or how to include:

Concluding remarks

A parsimonious model of gene functions: easy to apply on a large scale (we already ran some simulations using all 13,000 trees from PantherDB... and it took us less than 1 ~~week~~ hour with ~~10~~ 240 processors ~~only~~).

- Already implemented, we are currently in the stage of ~~writing the paper and setting up the simulation study~~ finishing and submiting the paper.
- For the next steps, we are evaluating whether to include or how to include:
  - Type of node: speciation, duplication, horizontal transfer.
  - Branch lengths
  - Correlation structure between functions
  - ~~Using Taxon Constraints to improve predictions~~
  - Hierarchical model: Use fully annotated trees by curators as prior information.
- We are still unsure about how to procede with the software: R journal? Journal of Open Source Software? Journal of Statistical Software? Bioinformatics? etc.

Thank you!

Project 2: Augmenting functional information about human genes using probabilistic phylogenetic modeling

George G. Vega Yon
vegayon@usc.edu

Duncan Thomas   Paul D. Thomas   Paul Marjoram   Huaiyu Mi   John Morrison

Department of Preventive Medicine
University of Southern California

November 14th, 2018