Essays on Bioinformatics and Social Network Analysis

Statistical and Computational Methods for Complex Systems

George G Vega Yon

University of Southern California, Department of Preventive Medicine

January 30, 2020

Keck School of Medicine of USC

On the prediction of gene functions using phylogenetic trees

Joint with: Paul D Thomas, Paul Marjoram, Huaiyu Mi, Duncan Thomas, and John Morrison

Genes

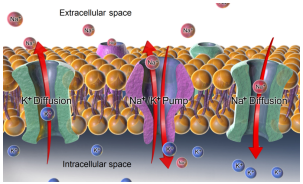
Encode the synthesis of genetic products that ultimately are related to a particular aspect of life, for example

Genes

Encode the synthesis of genetic products that ultimately are related to a particular aspect of life, for example

Molecular function

Active transport GO:0005215

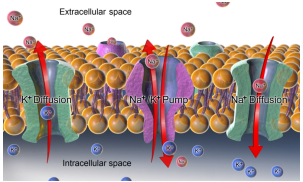


Genes

Encode the synthesis of genetic products that ultimately are related to a particular aspect of life, for example

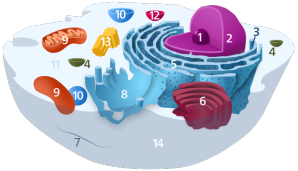
Molecular function

Active transport GO:0005215



Cellular component

Mitochondria GO:0004016

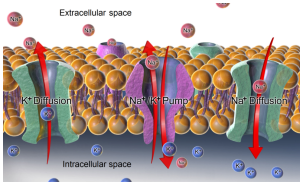


Genes

Encode the synthesis of genetic products that ultimately are related to a particular aspect of life, for example

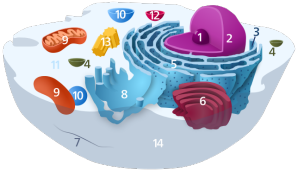
Molecular function

Active transport GO:0005215



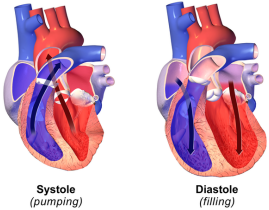
Cellular component

Mitochondria GO:0004016



Biological process

Heart contraction GO:0060047



Keck School of Medicine of USC

GENEONTOLOGY
Unifying Biology

- The GO project has $\sim$ 44,700 validated terms more, $\sim$ 7.3M annotations on $\sim$ 4,500 species.

source: Statistics from `pantherdb.org` and `geneontology.org`

Keck School of Medicine of USC

GENEONTOLOGY
Unifying Biology

- The GO project has $\sim$ 44,700 validated terms ▸ more, $\sim$ 7.3M annotations on $\sim$ 4,500 species.
- About $\sim$ 500,000 are on human genes.

source: Statistics from `pantherdb.org` and `geneontology.org`

Keck School of Medicine of USC

GENEONTOLOGY
Unifying Biology

- The GO project has $\sim$ 44,700 validated terms ▸ more, $\sim$ 7.3M annotations on $\sim$ 4,500 species.
- About $\sim$ 500,000 are on human genes.
- Roughly half of human genes ($\sim$ 10,000 / 20,000) have some form of annotation.

source: Statistics from `pantherdb.org` and `geneontology.org`

The Gene Ontology Project

Keck School of Medicine of USC

GENEONTOLOGY
Unifying Biology

- The GO project has $\sim$ 44,700 validated terms ▸ more, $\sim$ 7.3M annotations on $\sim$ 4,500 species.
- About $\sim$ 500,000 are on human genes.
- Roughly half of human genes ($\sim$ 10,000 / 20,000) have some form of annotation.
- We know something of less than 10% of known genes (near 1.7M across species).

source: Statistics from `pantherdb.org` and `geneontology.org`

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

You know what is interesting about this function?

These four species have a gene with that function...

Felis catus pthr10037

Oryzias latipes **pthr11521**

Anolis carolinensis **pthr11521**

Equus caballus pthr24356

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

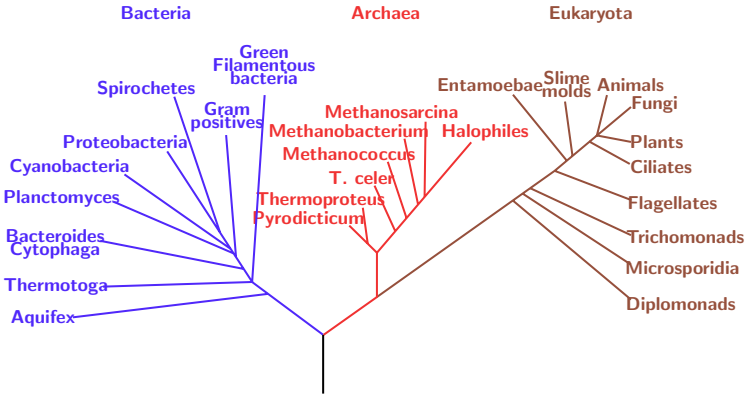


Figure 1 A phylogenetic tree of living things, based on RNA data and proposed by Carl Woese, showing the separation of bacteria, archaea, and eukaryotes (wiki)

Phylogenetic Trees: The PANTHER classification system

Phylogenetic Trees: The PANTHER classification system

Keck School of Medicine of USC

- The PANTHER project (part of GO) provides information about evolutionary structure of 1.7 million genes

Phylogenetic Trees: The PANTHER classification system

Keck School of Medicine of USC

- The PANTHER project (part of GO) provides information about evolutionary structure of 1.7 million genes
- These genes are grouped in 15,524 phylogenetic trees (families)

Phylogenetic Trees: The PANTHER classification system

Keck School of Medicine of USC

- The PANTHER project (part of GO) provides information about evolutionary structure of 1.7 million genes
- These genes are grouped in 15,524 phylogenetic trees (families)
- A single family can host multiple species

Phylogenetic Trees: The PANTHER classification system

- The PANTHER project (part of GO) provides information about evolutionary structure of 1.7 million genes
- These genes are grouped in 15,524 phylogenetic trees (families)
- A single family can host multiple species

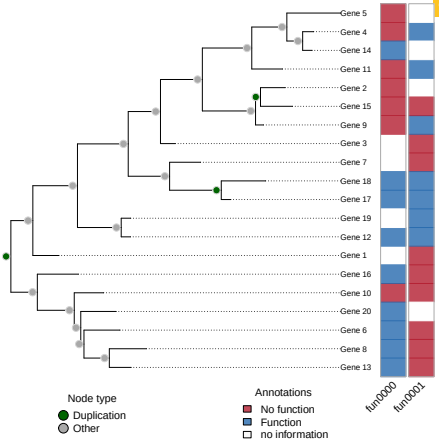


Figure 2 Simulated phylogenetic tree and gene annotations.

Speciation

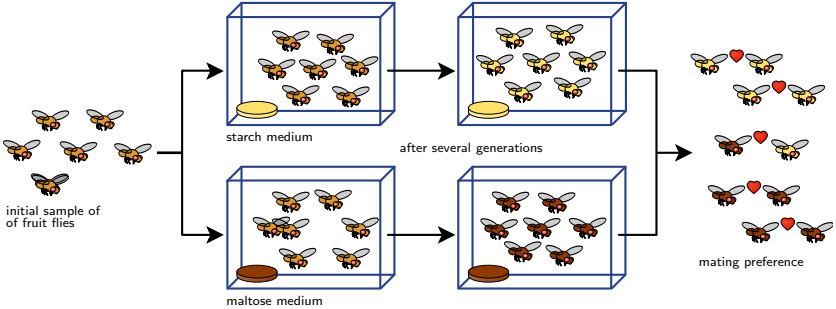


Figure 3 Dodd 1989: After one year of isolation, flies showed a significant level or assortativity in mating (wikimedia)

go back

Duplication

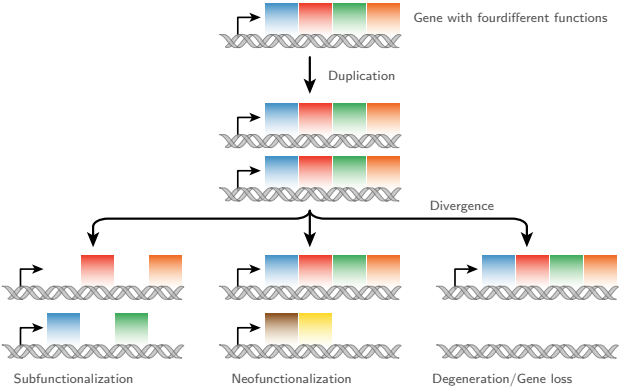


Figure 4 A key part of molecular innovation, gene duplication provides opportunity for new functions to emerge (wikimedia)

go back

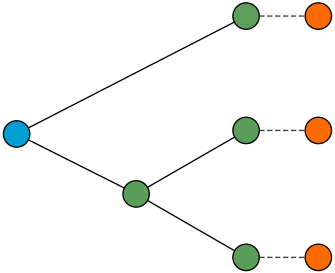
We can use

evolutionary trees

to inform a model for predicting

genetic annotations!

An evolutionary model of gene functions

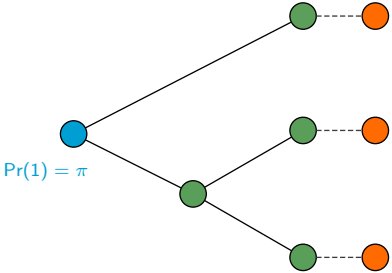


other models

other view

An evolutionary model of gene functions

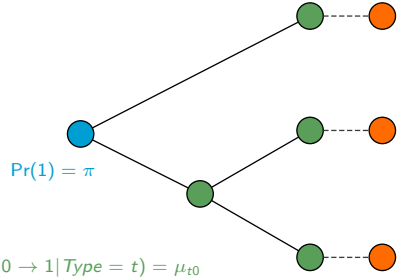
- Initial (spontaneous) gain of function.



other models | other view

An evolutionary model of gene functions

- Initial (spontaneous) gain of function.
- Loss/gain of offspring depends on: (a) the state of their parents (**Markov process**), and (b) the type of node



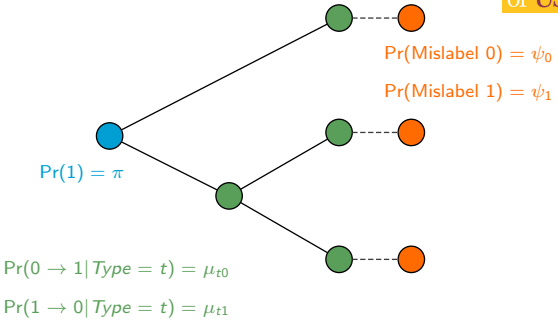
$\Pr(0 \to 1 | \mathit{Type} = t) = \mu_{t0}$

$\Pr(1 \to 0 | \mathit{Type} = t) = \mu_{t1}$

other models
other view

An evolutionary model of gene functions

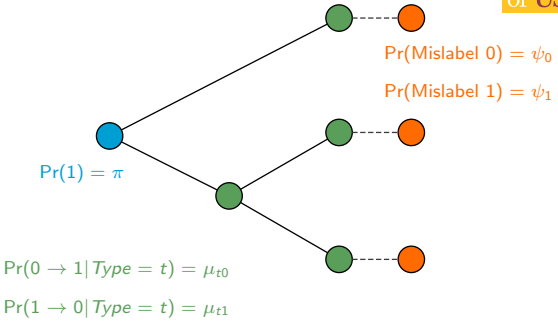
- Initial (spontaneous) gain of function.
- Loss/gain of offspring depends on: (a) the state of their parents (**Markov process**), and (b) the type of node
- We control for human error.



other models
other view

An evolutionary model of gene functions

- Initial (spontaneous) gain of function.
- Loss/gain of offspring depends on: (a) the state of their parents (**Markov process**), and (b) the type of node
- We control for human error.



We implemented the model using Felsenstein's' pruning algorithm (linear complexity) in the R package `aphylo`.

other models
other view

The `aphylo` R package

- Simulation and visualization of annotated phylogenetic trees.

The `aphylo` R package

Keck School of Medicine of USC

- Simulation and visualization of annotated phylogenetic trees.
- Pruning algorithm implemented in C++ using the `pruner` template library (by-product).

The `aphylo` R package

Keck School of Medicine of USC

- Simulation and visualization of annotated phylogenetic trees.
- Pruning algorithm implemented in C++ using the `pruner` template library (by-product).
- Uses meta-programming: users can specify different formulas, including pooled-data models.

The `aphylo` R package

Keck School of Medicine of USC

- Simulation and visualization of annotated phylogenetic trees.
- Pruning algorithm implemented in C++ using the `pruner` template library (by-product).
- Uses meta-programming: users can specify different formulas, including pooled-data models.
- The estimation is done using either Maximum Likelihood, Maximum A Posteriory, or MCMC.

The `aphylo` R package

Keck School of Medicine of USC

- Simulation and visualization of annotated phylogenetic trees.
- Pruning algorithm implemented in C++ using the `pruner` template library (by-product).
- Uses meta-programming: users can specify different formulas, including pooled-data models.
- The estimation is done using either Maximum Likelihood, Maximum A Posteriory, or MCMC.
- The MCMC estimation is done via the `fmcmc` R package using adaptive MCMC (also implemented as part of this project):

The `aphylo` R package

- Simulation and visualization of annotated phylogenetic trees.
- Pruning algorithm implemented in C++ using the `pruner` template library (by-product).
- Uses meta-programming: users can specify different formulas, including pooled-data models.
- The estimation is done using either Maximum Likelihood, Maximum A Posteriory, or MCMC.
- The MCMC estimation is done via the `fmcmc` R package using adaptive MCMC (also implemented as part of this project):
  - Automatic stop via convergence check.

The `aphylo` R package

Keck School of Medicine of USC

- Simulation and visualization of annotated phylogenetic trees.
- Pruning algorithm implemented in C++ using the `pruner` template library (by-product).
- Uses meta-programming: users can specify different formulas, including pooled-data models.
- The estimation is done using either Maximum Likelihood, Maximum A Posteriory, or MCMC.
- The MCMC estimation is done via the `fmcmc` R package using adaptive MCMC (also implemented as part of this project):
  - Automatic stop via convergence check.
  - Out-of-the-box parallel chains using parallel computing.

The `aphylo` R package

Keck School of Medicine of USC

- Simulation and visualization of annotated phylogenetic trees.
- Pruning algorithm implemented in C++ using the `pruner` template library (by-product).
- Uses meta-programming: users can specify different formulas, including pooled-data models.
- The estimation is done using either Maximum Likelihood, Maximum A Posteriory, or MCMC.
- The MCMC estimation is done via the `fmcmc` R package using adaptive MCMC (also implemented as part of this project):
  - Automatic stop via convergence check.
  - Out-of-the-box parallel chains using parallel computing.
  - User-defined transition kernel (in our case, Adaptive Kernel).

Some preliminary results

Joint with: Paul D Thomas, Paul Marjoram, Huaiyu Mi, Duncan Thomas, and John Morrison

Prediction with real data

	Prior	
	Uniform	Beta
Mislab. prob.		
ψ_0	0.23	0.25
ψ_1	0.01	0.01
Gain/Loss at dupl.		
μ_{d0}	0.97	0.96
μ_{d1}	0.52	0.58
Gain/Loss at spec.		
μ_{s0}	0.05	0.06
μ_{s1}	0.01	0.02
Root node		
π	0.81	0.45
Leave-one-out AUC		
Mean	0.69	0.67
Median	0.81	0.75

Table 2 Parameter estimates using different priors.

- 141 pooled functions (trees) with 7,388 genes with 0/1 annotations.

Prediction with real data

	Prior	
	Uniform	Beta
Mislab. prob.		
ψ_0	0.23	0.25
ψ_1	0.01	0.01
Gain/Loss at dupl.		
μ_{d0}	0.97	0.96
μ_{d1}	0.52	0.58
Gain/Loss at spec.		
μ_{s0}	0.05	0.06
μ_{s1}	0.01	0.02
Root node		
π	0.81	0.45
Leave-one-out AUC		
Mean	0.69	0.67
Median	0.81	0.75

Table 2 Parameter estimates using different priors.

- 141 pooled functions (trees) with 7,388 genes with 0/1 annotations.
- Parameter estimates are actually probabilities.

Prediction with real data

	Prior	
	Uniform	Beta
Mislab. prob.		
ψ_0	0.23	0.25
ψ_1	0.01	0.01
Gain/Loss at dupl.		
μ_{d0}	0.97	0.96
μ_{d1}	0.52	0.58
Gain/Loss at spec.		
μ_{s0}	0.05	0.06
μ_{s1}	0.01	0.02
Root node		
π	0.81	0.45
Leave-one-out AUC		
Mean	0.69	0.67
Median	0.81	0.75

Table 2 Parameter estimates using different priors.

- 141 pooled functions (trees) with 7,388 genes with 0/1 annotations.
- Parameter estimates are actually probabilities.
- Data driven results (uninformative prior).

Prediction with real data

	Prior	
	Uniform	Beta
Mislab. prob.		
ψ_0	0.23	0.25
ψ_1	0.01	0.01
Gain/Loss at dupl.		
μ_{d0}	0.97	0.96
μ_{d1}	0.52	0.58
Gain/Loss at spec.		
μ_{s0}	0.05	0.06
μ_{s1}	0.01	0.02
Root node		
π	0.81	0.45
Leave-one-out AUC		
Mean	0.69	0.67
Median	0.81	0.75

Table 2 Parameter estimates using different priors.

- 141 pooled functions (trees) with 7,388 genes with 0/1 annotations.
- Parameter estimates are actually probabilities.
- Data driven results (uninformative prior).
- Biologically meaningful results.

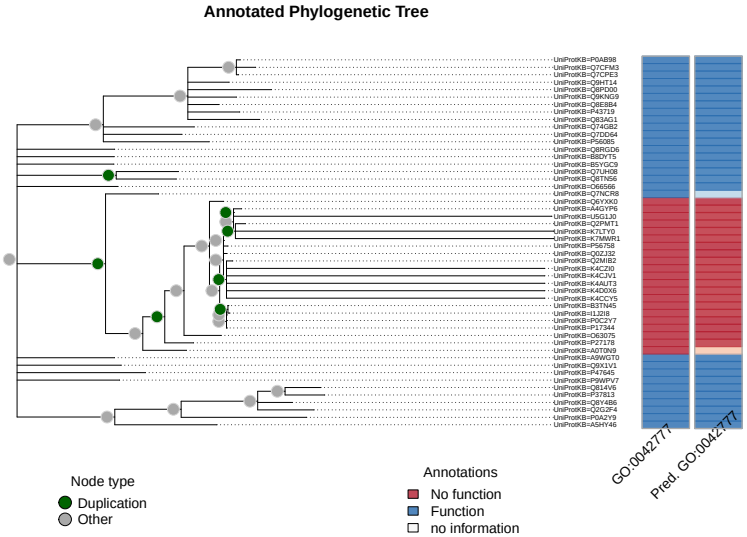
Prediction with real data

	Prior	
	Uniform	Beta
Mislab. prob.		
ψ_0	0.23	0.25
ψ_1	0.01	0.01
Gain/Loss at dupl.		
μ_{d0}	0.97	0.96
μ_{d1}	0.52	0.58
Gain/Loss at spec.		
μ_{s0}	0.05	0.06
μ_{s1}	0.01	0.02
Root node		
π	0.81	0.45
Leave-one-out AUC		
Mean	0.69	0.67
Median	0.81	0.75

Table 2 Parameter estimates using different priors.

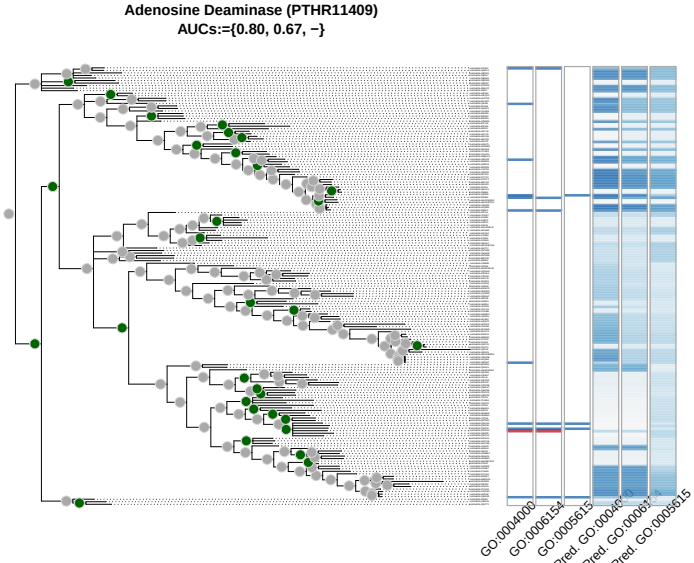
- 141 pooled functions (trees) with 7,388 genes with 0/1 annotations.
- Parameter estimates are actually probabilities.
- Data driven results (uninformative prior).
- Biologically meaningful results.
- Took about 5 minutes each.

Prediction with real data: Leave-one-out



Prediction with real data: Out-of-sample prediction

Keck School of Medicine of USC



Keck School of Medicine of USC

Key takeaways

- A parsimonious model for predicting gene functions using phylogenetics.
- Computationally scalable. SIFTER (our benchmark) would take about 66 years (yes, years) to estimate a model for 100 families of size 300, we take about 5 minutes.
- Meaningful biological results.
- Preliminary accuracy results comparable to state-of-the-art phylo-based models.

Keck School of Medicine of USC

Key takeaways

- A parsimonious model for predicting gene functions using phylogenetics.
- Computationally scalable. SIFTER (our benchmark) would take about 66 years (yes, years) to estimate a model for 100 families of size 300, we take about 5 minutes.
- Meaningful biological results.
- Preliminary accuracy results comparable to state-of-the-art phylo-based models.

Future Research: phylogenetic models

- Make the model hierarchical when pooling trees

Future Research: phylogenetic models

Keck School of Medicine of USC

- Make the model hierarchical when pooling trees
  - Different mutation rates per class of tree/function
  - Can be complicated to fit/justify (how many classes?)

Future Research: phylogenetic models

- Make the model hierarchical when pooling trees
  - Different mutation rates per class of tree/function
  - Can be complicated to fit/justify (how many classes?)
- Use a framework similar to Exponential Random Graph Models:

Future Research: phylogenetic models

Keck School of Medicine of USC

- Make the model hierarchical when pooling trees
  - Different mutation rates per class of tree/function
  - Can be complicated to fit/justify (how many classes?)
- Use a framework similar to Exponential Random Graph Models:

$$\mathbb{P}\left(\mathbf{X} = \{x_{n1}, x_{n2}, \ldots\} \;\middle|\; x_{\mathbf{p}(n1,\ldots)}\right) = \frac{\exp\left\{\mu^T s(\mathbf{x}|x_{\mathbf{p}(\cdot)})\right\}}{\sum_{\mathbf{x}'} \exp\left\{\mu^T s(\mathbf{x}'|x_{\mathbf{p}(\cdot)})\right\}}$$

  - A generalization of the model.
  - Extends to account for joint dist of functions+siblings.
  - Can incorporate additional information such as branch lengths.
  - Yet computationally more compact compared to SIFTER (finite number of parameters).

Future Research: phylogenetic models

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

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

Future Research: phylogenetic models

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

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

Sufficient statistics

Future Research: phylogenetic models

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

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

Sufficient statistics

# Gains	1	2

Future Research: phylogenetic models

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

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

Sufficient statistics

# Gains	1	2
# only one offspring changes	1	0

Future Research: phylogenetic models

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

			Transitions to	
			Case 1	Case 2
Parent	A	$\begin{bmatrix} 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}$
	B	$\begin{bmatrix} 1 \end{bmatrix}$		
	C	$\begin{bmatrix} 0 \end{bmatrix}$		
Sufficient statistics				
# Gains			1	2
# only one offspring changes			1	0
# Swaps (0→1, 1→0)			2	4

Future Research: phylogenetic models

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

			Transitions to	
			Case 1	Case 2
Parent	A	$\begin{bmatrix} 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}$
	B	$\begin{bmatrix} 1 \end{bmatrix}$		
	C	$\begin{bmatrix} 0 \end{bmatrix}$		
Sufficient statistics				
# Gains			1	2
# only one offspring changes			1	0
# Swaps (0→1, 1→0)			2	4

In SIFTER, for modelling 3 functions, we need $2^{2\times 3} = 64$ parameters.

Essays on Bioinformatics and Social Network Analysis

Statistical and Computational Methods for Complex Systems

George G Vega Yon

University of Southern California, Department of Preventive Medicine

January 30, 2020

Keck School of Medicine of USC

Thanks!

References I

Dodd, Diane M. B. (1989). "Reproductive Isolation as a Consequence of Adaptive Divergence in Drosophila pseudoobscura". In: Evolution 43.6, pp. 1308–1311. ISSN: 00143820, 15585646. URL: http://www.jstor.org/stable/2409365.

Engelhardt, Barbara E. et al. (2011). "Genome-scale phylogenetic function annotation of large and diverse protein families". In: Genome Research 21.11, pp. 1969–1980. ISSN: 10889051. DOI: 10.1101/gr.104687.109.

Engelhardt, Barbara E et al. (2005). "Protein Molecular Function Prediction by Bayesian Phylogenomics". In: PLOS Computational Biology 1.5. DOI: 10.1371/journal.pcbi.0010045. URL: https://doi.org/10.1371/journal.pcbi.0010045.

Jiang, Yuxiang et al. (Dec. 2016). "An expanded evaluation of protein function prediction methods shows an improvement in accuracy". In: Genome Biology 17.1, p. 184. ISSN: 1474-760X. DOI: 10.1186/s13059-016-1037-6. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1037-6.

References II

Oliver, Stephen (Feb. 2000). "Guilt-by-association goes global". In: Nature 403.6770, pp. 601–602. ISSN: 0028-0836. DOI: 10.1038/35001165. URL: http://www.nature.com/articles/35001165.

Pesaranghader, Ahmad et al. (May 2016). "simDEF: definition-based semantic similarity measure of gene ontology terms for functional similarity analysis of genes". In: Bioinformatics 32.9, pp. 1380–1387. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv755. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv755.

Piovesan, Damiano et al. (July 2015). "INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity". In: Nucleic Acids Research 43.W1, W134–W140. ISSN: 0305-1048. DOI: 10.1093/nar/gkv523. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv523.

References III

Yu, Chun et al. (Jan. 2018). "Assessing the Performances of Protein Function Prediction Algorithms from the Perspectives of Identification Accuracy and False Discovery Rate". In: International Journal of Molecular Sciences 19.1, p. 183. ISSN: 1422-0067. DOI: 10.3390/ijms19010183. URL: http://www.mdpi.com/1422-0067/19/1/183.

Predicting gene functions

There various approaches for this, some to highlight

- Text analysis like in Pesaranghader et al. 2016
- Protein-protein interaction networks like in Oliver 2000; Piovesan et al. 2015.
- Phylogenetic based like SIFTER Barbara E. Engelhardt et al. 2011, 2005.
  - Parameters to estimate: 2^{2P}, where P is the number of functions.

(a nice literature review in Jiang et al. 2016; Yu et al. 2018) go back

An evolutionary model of gene functions (algorithmic view)

Keck School of Medicine of USC

Data: A phylogenetic tree, $\{\pi, \mu, \psi\}$(Model probabilities)

Result: An annotated tree

```
for n ∈ PostOrder(N) do
    Nodes gain/loss function depending on their parent;
    switch class of n do
        case root node do
            Gain function with probability π;
        case interior node do
            if Parent has the function then Keep it with prob. (1 − μ1);
            else Gain it with prob. μ0;
    end
    Finally, we allow for mislabeling;
    if n is leaf then
        if has the function then Mislabel with prob. ψ1;
        else Mislabel with prob. ψ0;
end
```

go back

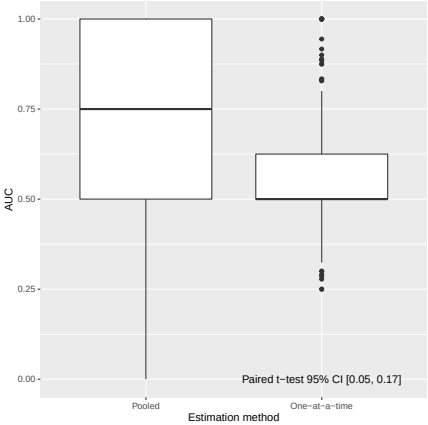


Figure 5 Comparing LOOCV AUC when performing predictions using either the estimates from the pooled model or each trees' own set of estimates obtained when fitting the model individually go back.