Seminario IMC UC

Predicción de funciones genéticas utilizando evidencia experimental y árboles filogenéticos: Un modelo evolutivo

o Ciencia de datos en la práctica

George G Vega Yon

Candidato a Doctor

University of Southern California, Department of Preventive Medicine

Abril 14, 2020

Keck School of Medicine of USC

Acerca de

- Ingeniero comercial UAI

Acerca de

- Ingeniero comercial UAI $\rightarrow$ Economista Caltech

Acerca de

- Ingeniero comercial UAI $\rightarrow$ Economista Caltech $\rightarrow$ Estadística Computacional USC.

Acerca de

- Ingeniero comercial UAI $\rightarrow$ Economista Caltech $\rightarrow$ Estadística Computacional USC.
- Funcionario público (5 años repartidos entre Mineduc, MDS, S Pensiones).

Acerca de

- Ingeniero comercial UAI $\rightarrow$ Economista Caltech $\rightarrow$ Estadística Computacional USC.
- Funcionario público (5 años repartidos entre Mineduc, MDS, S Pensiones).
- Nerd de R (fundé el grupo de usuarios de R en Chile el 2013).

Acerca de

- Ingeniero comercial UAI $\rightarrow$ Economista Caltech $\rightarrow$ Estadística Computacional USC.
- Funcionario público (5 años repartidos entre Mineduc, MDS, S Pensiones).
- Nerd de R (fundé el grupo de usuarios de R en Chile el 2013).
- Casado, 2 hijos, 1 perro.

Acerca de

- Ingeniero comercial UAI $\rightarrow$ Economista Caltech $\rightarrow$ Estadística Computacional USC.
- Funcionario público (5 años repartidos entre Mineduc, MDS, S Pensiones).
- Nerd de R (fundé el grupo de usuarios de R en Chile el 2013).
- Casado, 2 hijos, 1 perro.
- Varios años en proyectos de "Data Science".

Acerca de

- Ingeniero comercial UAI → Economista Caltech → Estadística Computacional USC.
- Funcionario público (5 años repartidos entre Mineduc, MDS, S Pensiones).
- Nerd de R (fundé el grupo de usuarios de R en Chile el 2013).
- Casado, 2 hijos, 1 perro.
- Varios años en proyectos de "Data Science".
- Mi investigación se centra en estadística computacional con énfasis en: Computación en paralelo, análisis de redes sociales, desarrollo de software, y métodos gral.

Acerca de

- Ingeniero comercial UAI $\rightarrow$ Economista Caltech $\rightarrow$ Estadística Computacional USC.
- Funcionario público (5 años repartidos entre Mineduc, MDS, S Pensiones).
- Nerd de R (fundé el grupo de usuarios de R en Chile el 2013).
- Casado, 2 hijos, 1 perro.
- Varios años en proyectos de "Data Science".
- Mi investigación se centra en estadística computacional con énfasis en: Computación en paralelo, análisis de redes sociales, desarrollo de software, y métodos gral.

Más en **http://ggvy.cl**.

On the prediction of gene functions using phylogenetic trees

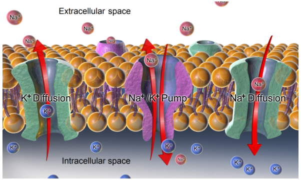
Joint with: Paul D Thomas, Paul Marjoram, Huaiyu Mi, Duncan Thomas, and John Morrison

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

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

Molecular function

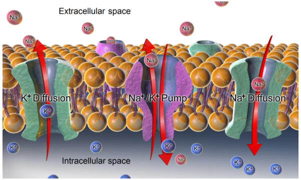
Active transport GO:0005215



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

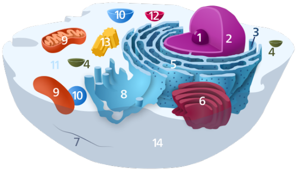
Molecular function

Active transport GO:0005215



Cellular component

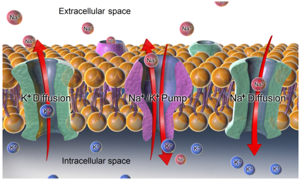
Mitochondria GO:0004016



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

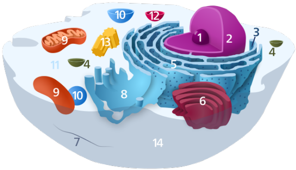
Molecular function

Active transport GO:0005215



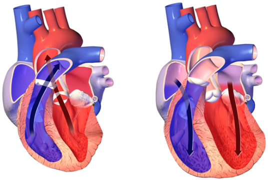
Cellular component

Mitochondria GO:0004016



Biological process

Heart contraction GO:0060047



Keck School of Medicine of USC

- 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`

- 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`

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

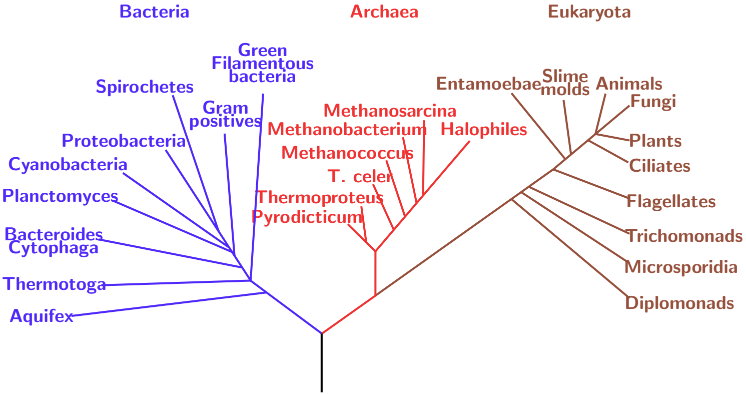


Figure 1 A phylogenetic tree of living things, based on RNA data and proposed by Carl Woese, showing the separation of bacteria, archaea, and eukaryotes (wiki)

Phylogenetic Trees: The PANTHER classification system

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

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)

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

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

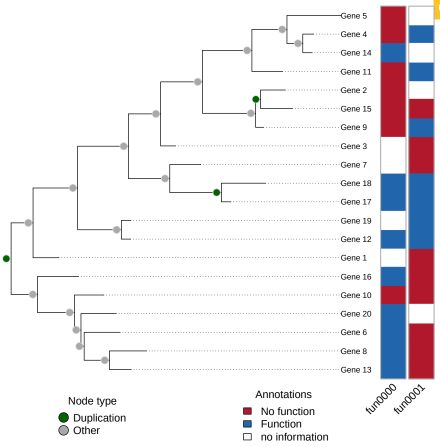


Figure 2 Simulated phylogenetic tree and gene annotations.

Keck School of Medicine of USC

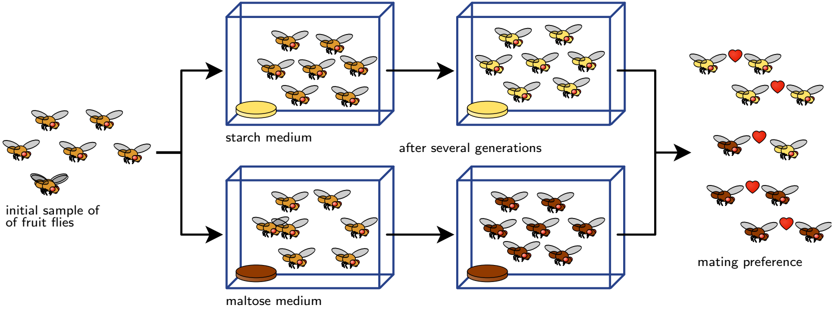


Figure 3 Dodd 1989: After one year of isolation, flies showed a significant level or assortativity in mating (wikimedia)

go back

Keck School of Medicine of USC

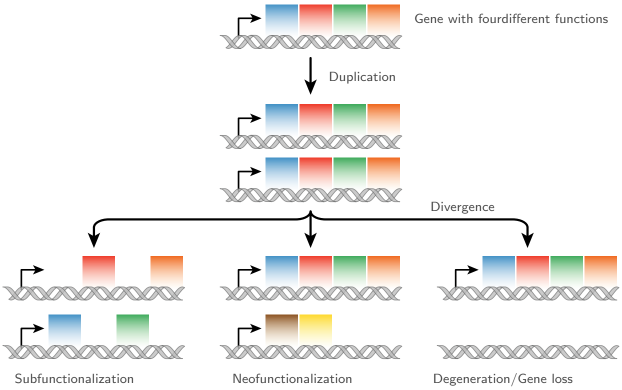
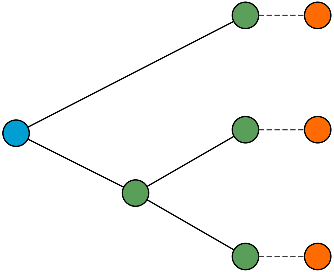


Figure 4 A key part of molecular innovation, gene duplication provides opportunity for new functions to emerge (wikimedia)

go back

An evolutionary model of gene functions

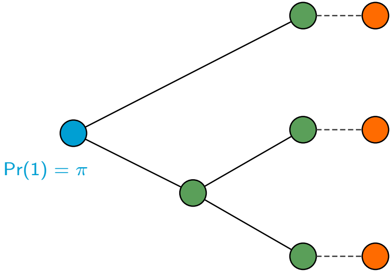


other models other view

An evolutionary model of gene functions

Keck School of Medicine of USC

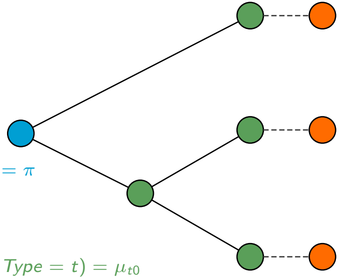
- Initial (spontaneous) gain of function.



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 (**(discrete) Markov process**), and (b) the type of node



$\Pr(1) = \pi$

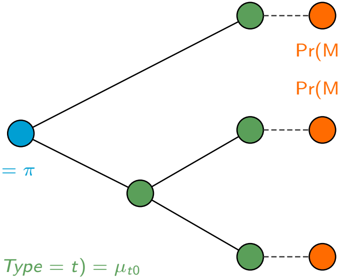
$\Pr(0 \to 1 | \mathit{Type} = t) = \mu_{t0}$

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

▸ 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 (**(discrete) Markov process**), and (b) the type of node
- We control for human error.



$\Pr(\text{Mislabel } 0) = \psi_0$

$\Pr(\text{Mislabel } 1) = \psi_1$

$\Pr(1) = \pi$

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

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

▸ other models ▸ other view

Formally

We need to calculate the probability of observing $\tilde{D} = (\Lambda, \mathbf{Z})$ (a partially annotated phylogeny) as a function of the model parameters ψ (mislabel), μ (gain/loss), π (root node):

- Probability of the induced sub-tree:

$$\mathbb{P}\left(\tilde{D}_n \,\middle|\, x_n, \psi, \mu\right) = \prod_{m \in \mathbf{O}(n)} \mathbb{P}\left(\tilde{D}_m \,\middle|\, x_n\right), \tag{1}$$

  where

$$\mathbb{P}\left(\tilde{D}_m \,\middle|\, x_n\right) = \begin{cases} \sum_{x_m \in \{0,1\}} \mathbb{P}\left(\tilde{D}_m \,\middle|\, x_m, \psi, \mu\right) \mathbb{P}\left(x_m \mid x_n, \mu\right) & \text{if } m \text{ is an interior node,} \\ \sum_{x_m \in \{0,1\}} \mathbb{P}\left(x_m \mid z_m, \psi\right) \mathbb{P}\left(x_m \mid x_n, \mu\right) & \text{if } m \text{ is a leaf node.} \end{cases}$$

- The exact likelihood:

$$\mathrm{L}\left(\psi, \mu, \pi \,\middle|\, \tilde{D}\right) = \sum_{x_0 \in \{0,1\}} \mathbb{P}\left(x_0 \mid \pi\right) \mathbb{P}\left(\tilde{D}_0 \,\middle|\, x_0, \psi, \mu\right) \tag{2}$$

This likelihood can be computed in $O(n)$, n number of nodes. This is known as Post-order tree-traversal, or Felsenstein's Pruning algorithm.

Implementation

Software, algorithms, and analysis

Software and algorithms

- The likelihood function is computed using the C++ template library `pruner` (by-product).

Software and algorithms

- The likelihood function is computed using the C++ template library `pruner` (by-product).
- To fit the model we use:

Keck School of Medicine of USC

Software and algorithms

- The likelihood function is computed using the C++ template library `pruner` (by-product).
- To fit the model we use:
  - *Maximum Likelihood Estimation (MLE) and Maximum A Posteriory (MAP)*: Maximized using L-BFGS-B.

Software and algorithms

- The likelihood function is computed using the C++ template library `pruner` (by-product).
- To fit the model we use:
  - *Maximum Likelihood Estimation (MLE) and Maximum A Posteriory (MAP)*: Maximized using L-BFGS-B.
  - *Markov Chain Monte Carlo (MCMC)*: Using the `fmcmc` R package (by-product), and in particular, Haario's Adaptive Metropolis.

Keck School of Medicine of USC

Software and algorithms

- The likelihood function is computed using the C++ template library `pruner` (by-product).
- To fit the model we use:
  - *Maximum Likelihood Estimation (MLE) and Maximum A Posteriory (MAP)*: Maximized using L-BFGS-B.
  - *Markov Chain Monte Carlo (MCMC)*: Using the `fmcmc` R package (by-product), and in particular, Haario's Adaptive Metropolis.

Keck School of Medicine of USC

Software and algorithms

- The likelihood function is computed using the C++ template library `pruner` (by-product).
- To fit the model we use:
  - *Maximum Likelihood Estimation (MLE) and Maximum A Posteriory (MAP)*: Maximized using L-BFGS-B.
  - *Markov Chain Monte Carlo (MCMC)*: Using the `fmcmc` R package (by-product), and in particular, Haario's Adaptive Metropolis.

Analysis

- Conducted a large simulation study fitting 15,000 models with MCMC

Implementation

Keck School of Medicine of USC

Software and algorithms

- The likelihood function is computed using the C++ template library `pruner` (by-product).
- To fit the model we use:
  - *Maximum Likelihood Estimation (MLE) and Maximum A Posteriory (MAP)*: Maximized using L-BFGS-B.
  - *Markov Chain Monte Carlo (MCMC)*: Using the `fmcmc` R package (by-product), and in particular, Haario's Adaptive Metropolis.

Analysis

- Conducted a large simulation study fitting 15,000 models with MCMC
- The analysis was performed in USC's HPCC and took about 4 hours (using 400 cores, i.e. 2 month equiv to core hours)

Keck School of Medicine of USC

Software and algorithms

- The likelihood function is computed using the C++ template library `pruner` (by-product).
- To fit the model we use:
  - *Maximum Likelihood Estimation (MLE) and Maximum A Posteriory (MAP)*: Maximized using L-BFGS-B.
  - *Markov Chain Monte Carlo (MCMC)*: Using the `fmcmc` R package (by-product), and in particular, Haario's Adaptive Metropolis.

Analysis

- Conducted a large simulation study fitting 15,000 models with MCMC
- The analysis was performed in USC's HPCC and took about 4 hours (using 400 cores, i.e. 2 month equiv to core hours)
- We used the `slurmR` package (also by-product) to implement the pipe-line.

Data

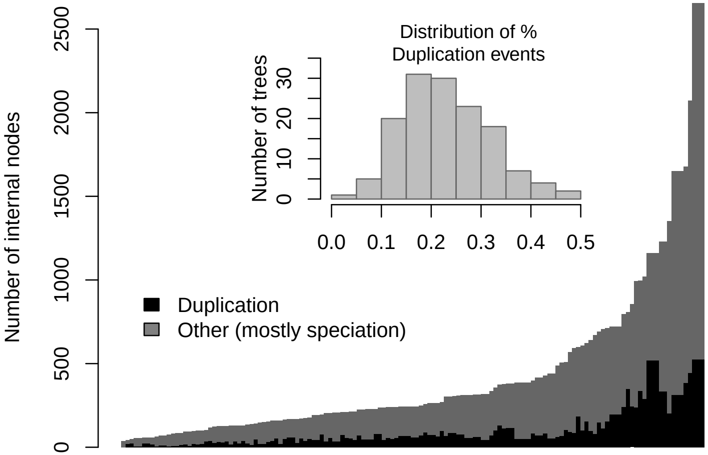
Phylogenetic trees and Experimental Annotations

Data: Phylogenetic trees

Sample of annotations (first 10 in a single tree)

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

Data: Node type (events)



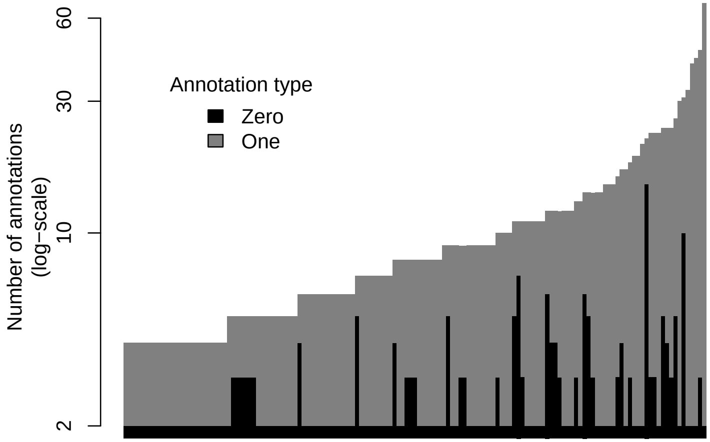
Data: Annotations (example)

This is the first 10 of $\sim$ 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	

Data: Experimental Annotations

Keck School of Medicine of USC



Some preliminary results

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

	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.

	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).

Keck School of Medicine of USC

	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.

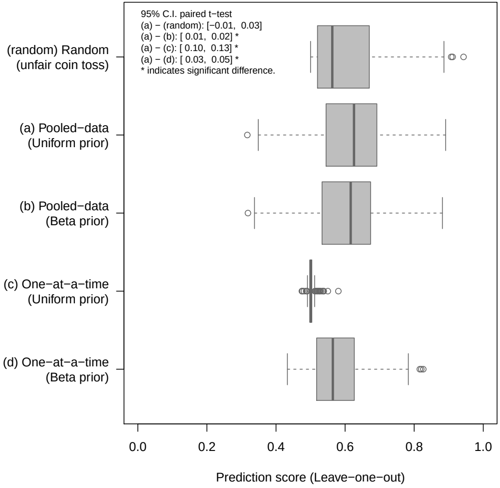
Keck School of Medicine of USC

	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.

Pooled estimation (worth it?)



Keck School of Medicine of USC

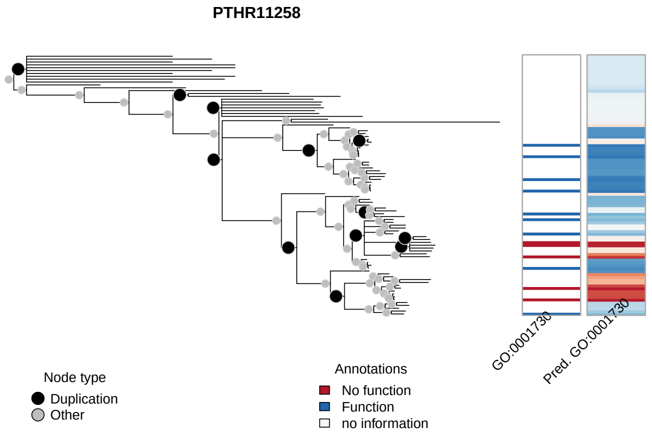
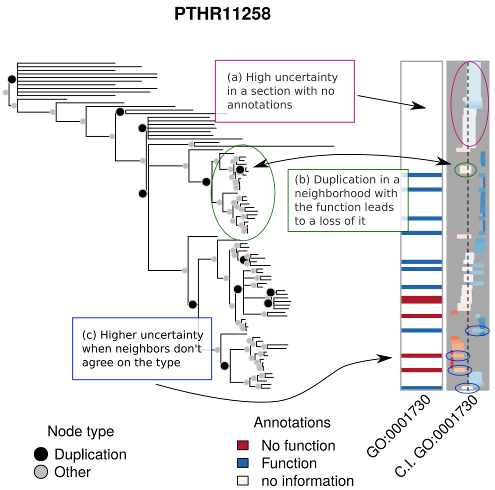


Figure 5 This family contains the human gene OAS1 (chromosome 12) "a member of the 2-5A synthetase family, essential proteins involved in the innate immune response to viral infection" (wiki)

Prediction with real data (leave-one-out CI)



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.

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

- 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:

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

  - 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)

$$
\begin{array}{cccccc}
 & & & & \multicolumn{2}{c}{\textbf{Transitions to}} \\
 & & & & \text{Case 1} & \text{Case 2} \\
\hline
\textbf{Parent} & \begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \end{array} & \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} & & \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} & \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}
\end{array}
$$

Future Research: phylogenetic models

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

$$
\begin{array}{rc|c c}
 & & \multicolumn{2}{c}{\textbf{Transitions to}} \\
 & & \text{Case 1} & \text{Case 2} \\
\hline
\textbf{Parent} \begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \end{array} & \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} & \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}
\end{array}
$$

Sufficient statistics

Future Research: phylogenetic models

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

$$
\begin{array}{llccc}
 & & & \multicolumn{2}{c}{\textbf{Transitions to}} \\
 & & & \text{Case 1} & \text{Case 2} \\
\textbf{Parent} & \begin{array}{c} A \\ B \\ C \end{array} & \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} & \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix} \\
\textbf{Sufficient statistics} & & & & \\
\quad \# \text{ Gains} & & & 1 & 2
\end{array}
$$

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	$\begin{array}{c} A \\ B \\ C \end{array} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}$
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	$\begin{array}{c} A \\ B \\ C \end{array} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}$
Sufficient statistics			
# Gains		1	2
# only one offspring changes		1	0
# changes (total)		2	4

Future Research: phylogenetic models

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

$$
\begin{array}{ccc}
 & & \textbf{Transitions to} \\
 & & \text{Case 1} \qquad\qquad \text{Case 2} \\
\textbf{Parent} & \begin{array}{c} A \\ B \\ C \end{array} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}
\end{array}
$$

Sufficient statistics	Case 1	Case 2
# Gains	1	2
# only one offspring changes	1	0
# changes (total)	2	4

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

Seminario IMC UC

Predicción de funciones genéticas utilizando evidencia experimental y árboles filogenéticos: Un modelo evolutivo

o Ciencia de datos en la práctica

George G Vega Yon

Candidato a Doctor

University of Southern California, Department of Preventive Medicine

Abril 14, 2020

Keck School of Medicine of USC

Thanks!

Keck School of Medicine of USC

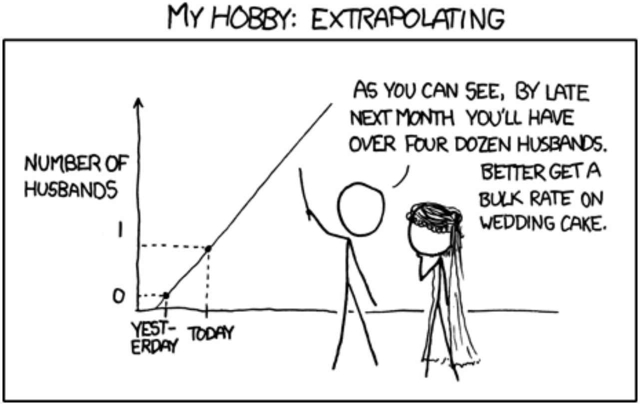


Figure 6 Fuente: https://xkcd.com/605/

Figure 7 Fuente: https://xkcd.com/208/

Keck School of Medicine of USC

Pero si insisten...

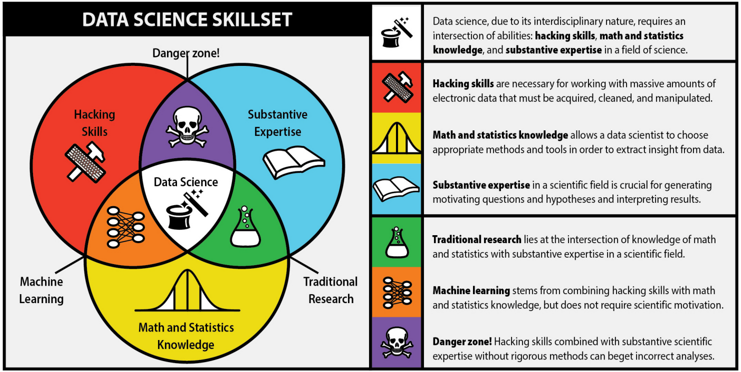


Figure 8 Fuente: https://berkeleysciencereview.com/2013/07/how-to-become-a-data-scientist-before-you-graduate/ Original de Drew Conway.

Finalmente

Reality Behind Data Science

Keck School of Medicine of USC

- 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.

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.

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 $n \in PostOrder(N)$ **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. $(1 - \mu_1)$;

            else Gain it with prob. μ_0;

    end

    Finally, we allow for mislabeling;

    if n is leaf **then**

        if has the function **then** Mislabel with prob. ψ_1;

        else Mislabel with prob. ψ_0;

end

go back