Predicting of Gene Functions by Leveraging Biological Insights with Mechanistic Machine Learning

George G. Vega Yon, Ph.D.
george.vegayon@utah.edu
Division of Epidemiology @ University of Utah

Jan 11th, 2023 @ Data Science Seminar

Collaborators: Paul Thomas, Paul Marjoram, Huaiyu Mi, Christopher Williams (USC), Alun Thomas (UofU)

Preliminaries
Evolution of Gene Function
Mechanistic Machine Learning
Proof of Concept

Table of Contents

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

You can download the slides from `https://ggv.cl/UofUDS2023`

Table of Contents

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

You can download the slides from `https://ggv.cl/UofUDS2023`

But first...

whoami

- Research Assistant Professor in Epidemiology since Nov. 2021.

whoami

- Research Assistant Professor in Epidemiology since Nov. 2021.
- A methodologist researching Statistical Computing applied to Complex Systems Modeling.

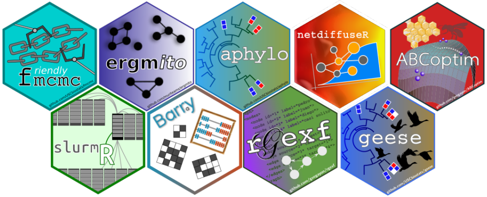
whoami

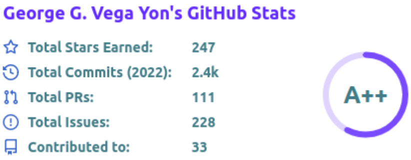
- Research Assistant Professor in Epidemiology since Nov. 2021.
- A methodologist researching Statistical Computing applied to Complex Systems Modeling.
- Got a Ph.D. in Biostats from USC and an M.Sc. in Economics from Caltech.

whoami

- Research Assistant Professor in Epidemiology since Nov. 2021.
- A methodologist researching Statistical Computing applied to Complex Systems Modeling.
- Got a Ph.D. in Biostats from USC and an M.Sc. in Economics from Caltech.
- Working with networks + programming since ~2010.

whoami





- Research Assistant Professor in Epidemiology since Nov. 2021.
- A methodologist researching Statistical Computing applied to Complex Systems Modeling.
- Got a Ph.D. in Biostats from USC and an M.Sc. in Economics from Caltech.
- Working with networks + programming since ∼2010.
- Scientific software developer (R/C++) with (¿ 500K downloads + 9 CRAN packages).

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

Table of Contents

Preliminaries

Evolution of Gene Function

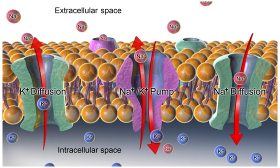
Mechanistic Machine Learning

Proof of Concept

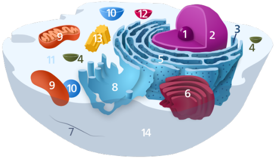
Gene Function

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

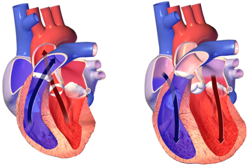
Molecular function
Active transport
GO:0005215



Cellular component
Mitochondria GO:0004016



Biological process
Heart contraction
GO:0060047



Gene Function: the Gene Ontology Project

- The GO project has $\sim$ 43,000 validated terms, $\sim$ 7.4M annotations on $\sim$ 5,200 species.

source: Statistics from `http://pantherdb.org/panther/summaryStats.jsp` and `http://geneontology.org/stats.html`

Gene Function: the Gene Ontology Project

GENEONTOLOGY
Unifying Biology

- The GO project has $\sim$ 43,000 validated terms, $\sim$ 7.4M annotations on $\sim$ 5,200 species.
- About $\sim$ 700,000 annotations are on human genes.

source: Statistics from `http://pantherdb.org/panther/summaryStats.jsp` and `http://geneontology.org/stats.html`

Gene Function: the Gene Ontology Project

- The GO project has $\sim$ 43,000 validated terms, $\sim$ 7.4M annotations on $\sim$ 5,200 species.
- About $\sim$ 700,000 annotations are on human genes.
- Only half of the human gene annotations are based on experimental evidence.

source: Statistics from `http://pantherdb.org/panther/summaryStats.jsp` and `http://geneontology.org/stats.html`

Gene Function: the Gene Ontology Project

GENEONTOLOGY
Unifying Biology

- The GO project has $\sim$ 43,000 validated terms, $\sim$ 7.4M annotations on $\sim$ 5,200 species.
- About $\sim$ 700,000 annotations are on human genes.
- Only half of the human gene annotations are based on experimental evidence.
- About $\sim$ 173,000 publications have used the GO.

source: Statistics from `http://pantherdb.org/panther/summaryStats.jsp` and `http://geneontology.org/stats.html`

Predicting Gene Function: State-of-the-art

Sequences, phylogenomics, and ML.

- **BLAST**:[2] Prediction by sequence homology ($\sim$ 105,000 citations).

Predicting Gene Function: State-of-the-art

Sequences, phylogenomics, and ML.

- **BLAST**:[2] Prediction by sequence homology ($\sim$ 105,000 citations).
- **SIFTER**:[6,5] An evolutionary model of gene function/loss using phylogenetics.

Predicting Gene Function: State-of-the-art

Sequences, phylogenomics, and ML.

- **BLAST**:[2] Prediction by sequence homology ($\sim$ 105,000 citations).
- **SIFTER**:[6,5] An evolutionary model of gene function/loss using phylogenetics.
- **aphylo**[15] (by yours truly): Another phylo-based method. Leverages negative annotations and pooled trees.

Predicting Gene Function: State-of-the-art

Sequences, phylogenomics, and ML.

- **BLAST**:[2] Prediction by sequence homology ($\sim$ 105,000 citations).
- **SIFTER**:[6,5] An evolutionary model of gene function/loss using phylogenetics.
- **aphylo**[15] (by yours truly): Another phylo-based method. Leverages negative annotations and pooled trees.
- **Phylo-PFP**:[10] A BLAST-based adding phylogenetic based distances.

Predicting Gene Function: State-of-the-art

Sequences, phylogenomics, and ML.

- **BLAST**:[2] Prediction by sequence homology ($\sim$ 105,000 citations).
- **SIFTER**:[6,5] An evolutionary model of gene function/loss using phylogenetics.
- **aphylo**[15] (by yours truly): Another phylo-based method. Leverages negative annotations and pooled trees.
- **Phylo-PFP**:[10] A BLAST-based adding phylogenetic based distances.
- **DeepGOPlus**:[12]One of the top-performing models in the literature, uses a 2D convolutional neural network on gene sequences.

Predicting Gene Function: State-of-the-art

Sequences, phylogenomics, and ML.

- **BLAST**:[2] Prediction by sequence homology ($\sim$ 105,000 citations).
- **SIFTER**:[6,5] An evolutionary model of gene function/loss using phylogenetics.
- **aphylo**[15] (by yours truly): Another phylo-based method. Leverages negative annotations and pooled trees.
- **Phylo-PFP**:[10] A BLAST-based adding phylogenetic based distances.
- **DeepGOPlus**:[12]One of the top-performing models in the literature, uses a 2D convolutional neural network on gene sequences.
- **GOLabeler**:[17] Top performing tool according to the *Critical Assessment of Function Annotation* [CAFA] challenge,[18] is an ensemble of various simple ML methods, including K-means and logistic regression.

Predicting Gene Function: State-of-the-art

Sequences, phylogenomics, and ML.

- **BLAST**:[2] Prediction by sequence homology ($\sim$ 105,000 citations).
- **SIFTER**:[6,5] An evolutionary model of gene function/loss using phylogenetics.
- **aphylo**[15] (by yours truly): Another phylo-based method. Leverages negative annotations and pooled trees.
- **Phylo-PFP**:[10] A BLAST-based adding phylogenetic based distances.
- **DeepGOPlus**:[12]One of the top-performing models in the literature, uses a 2D convolutional neural network on gene sequences.
- **GOLabeler**:[17] Top performing tool according to the *Critical Assessment of Function Annotation* [CAFA] challenge,[18] is an ensemble of various simple ML methods, including K-means and logistic regression.
- **DeepFRI**:[8] Uses Graph Convolutional Neural Networks (GCNs) to predict function based on protein structure and genetic sequence.

Predicting Gene Function: State-of-the-art

Sequences, phylogenomics, and ML.

- **BLAST**:[2] Prediction by sequence homology ($\sim$ 105,000 citations).
- **SIFTER**:[6,5] An evolutionary model of gene function/loss using phylogenetics.
- **aphylo**[15] (by yours truly): Another phylo-based method. Leverages negative annotations and pooled trees.
- **Phylo-PFP**:[10] A BLAST-based adding phylogenetic based distances.
- **DeepGOPlus**:[12]One of the top-performing models in the literature, uses a 2D convolutional neural network on gene sequences.
- **GOLabeler**:[17] Top performing tool according to the *Critical Assessment of Function Annotation* [CAFA] challenge,[18] is an ensemble of various simple ML methods, including K-means and logistic regression.
- **DeepFRI**:[8] Uses Graph Convolutional Neural Networks (GCNs) to predict function based on protein structure and genetic sequence.

In the latest CAFA, **none** of the top-performing methods scored an AUC above 0.60, and most were outperformed by BLAST,[19] which annotates using homology based on sequence similarity.

Table of Contents

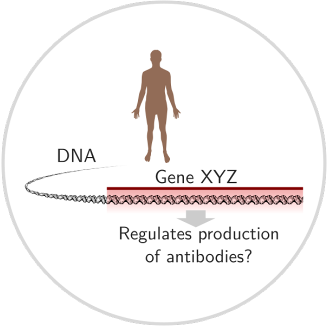
Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

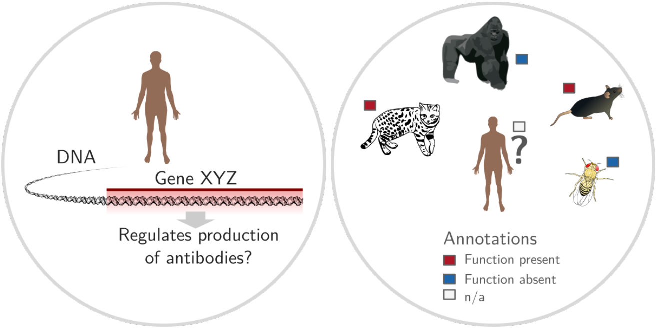
Proof of Concept

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



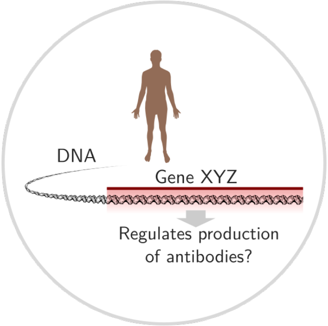
Complex to directly assess

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

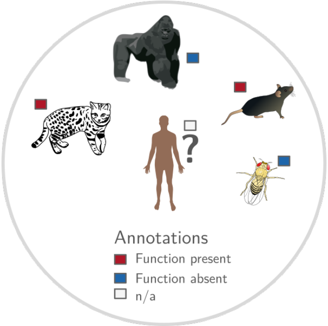


Complex to directly assess   But we may know from other species

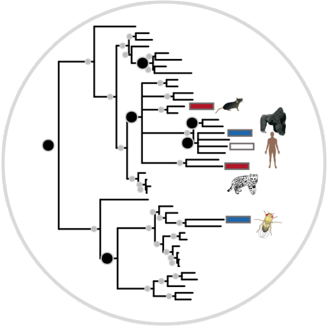
Is gene *XYZ* involved in process *ABC*?



Complex to directly assess

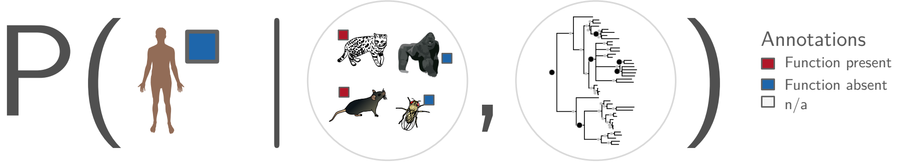


But we may know from other species



And we further know how these are *evolutionary* connected

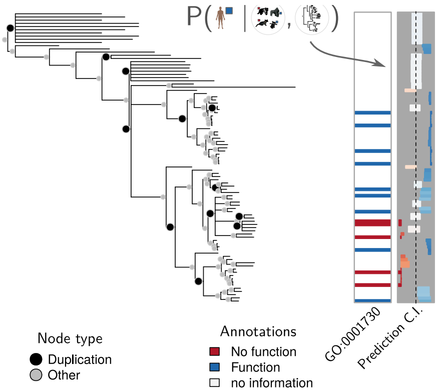
Is the human gene **XYZ** involved in process **ABC**, given what we know about that for other *related* species?



Evolution of Gene function (of one function)

Built a big model (lots of trees and annotations) called aphylo:

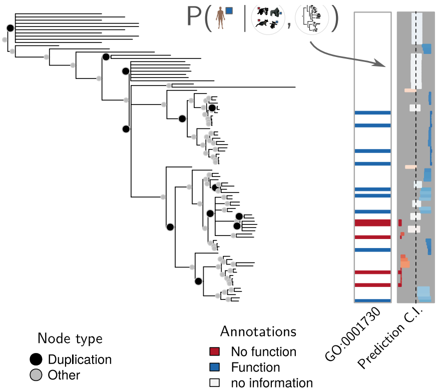
- Only two sources of data: Phylogenetic tree (`pantherdb.org`) and functional annotations (`geneontology.org`).



Evolution of Gene function (of one function)

Built a big model (lots of trees and annotations) called aphylo:

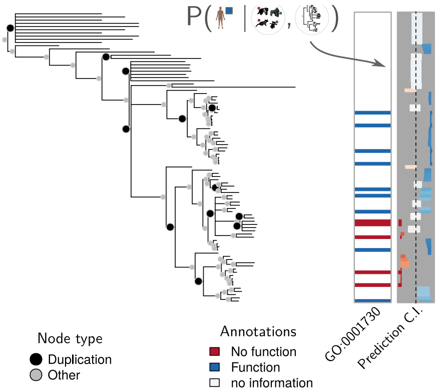
- Only two sources of data: Phylogenetic tree (`pantherdb.org`) and functional annotations (`geneontology.org`).
- Leverage negative annotation of GO terms (NOT).



Evolution of Gene function (of one function)

Built a big model (lots of trees and annotations) called aphylo:

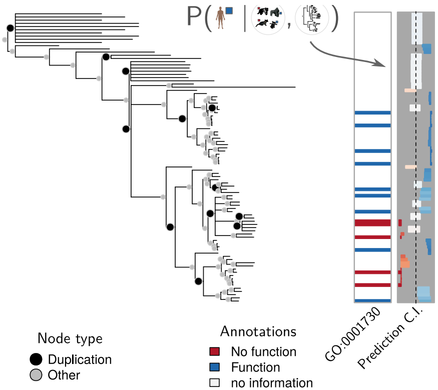
- Only two sources of data: Phylogenetic tree (`pantherdb.org`) and functional annotations (`geneontology.org`).
- Leverage negative annotation of GO terms (NOT).
- Use Felsenstein's tree pruning algorithm to compute tree likelihood.



Evolution of Gene function (of one function)

Built a big model (lots of trees and annotations) called aphylo:

- Only two sources of data: Phylogenetic tree (`pantherdb.org`) and functional annotations (`geneontology.org`).
- Leverage negative annotation of GO terms (NOT).
- Use Felsenstein's tree pruning algorithm to compute tree likelihood.
- Fit pooled models featuring thousands of annotations in hundreds of trees (with split-second prediction capability).



... But what if we wanted to deal with multiple functions?

Evolution of Gene function (multiple functions)

Tapping into Evol. Theory

- A fundamental part of Fun. Evol. is Duplication Events.

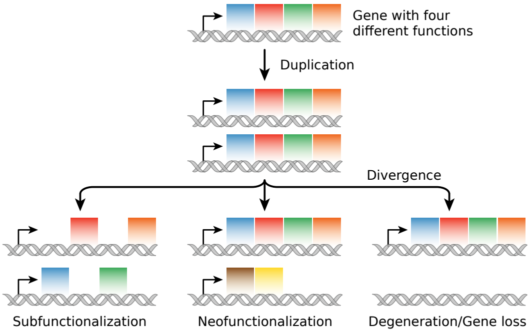


Figure: A key part of molecular innovation, gene duplication provides an opportunity for new functions to emerge (wikimedia)

Evolution of Gene function (multiple functions)

Tapping into Evol. Theory

- A fundamental part of Fun. Evol. is Duplication Events.
- Furthermore, knowing what happened to gene A (*e.g.*, neofunctionalization) is highly informative to learn about the functional state of B.

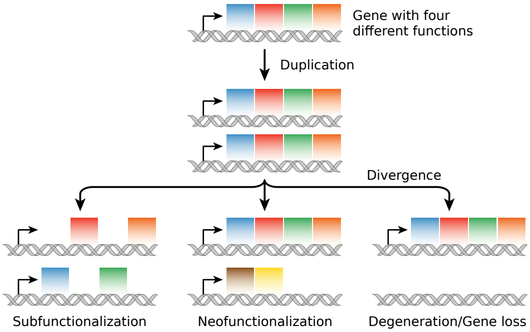


Figure: A key part of molecular innovation, gene duplication provides an opportunity for new functions to emerge (wikimedia)

Evolution of Gene function (multiple functions)

Tapping into Evol. Theory

- A fundamental part of Fun. Evol. is Duplication Events.
- Furthermore, knowing what happened to gene A (*e.g.*, neofunctionalization) is highly informative to learn about the functional state of B.
- One way to model this is using a Markov Transition Model (as in SIFTER).

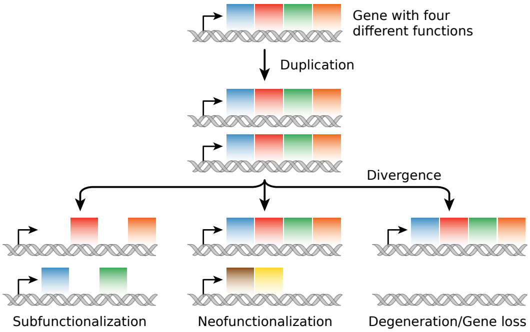
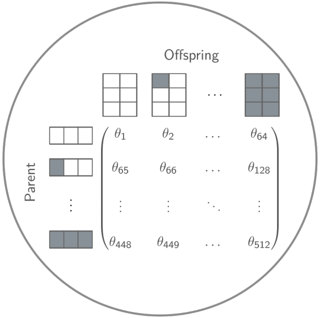


Figure: A key part of molecular innovation, gene duplication provides an opportunity for new functions to emerge (wikimedia)

Evolution of Gene function (multiple functions)

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

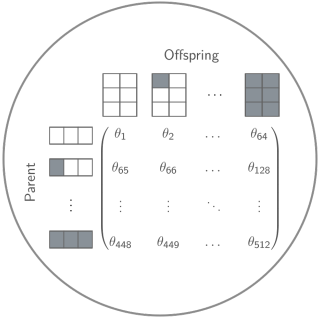
Full Markov Transition Matrix



Evolution of Gene function (multiple functions)

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

Full Markov Transition Matrix

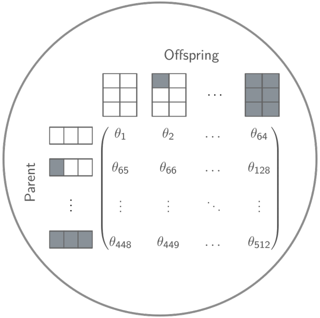


- 512 parameters

Evolution of Gene function (multiple functions)

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

Full Markov Transition Matrix

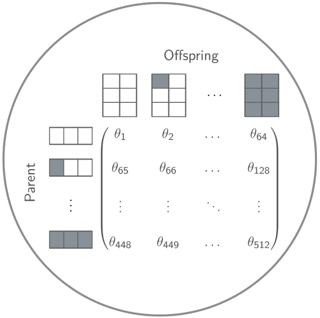


- 512 parameters
- Finding this many parameters is not easy.

Evolution of Gene function (multiple functions)

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

Full Markov Transition Matrix

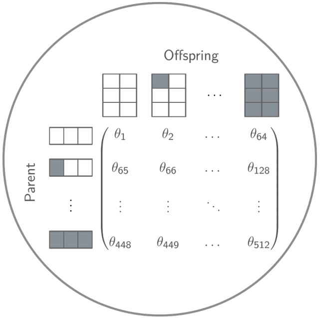


- 512 parameters
- Finding this many parameters is not easy.
- Even if you can, interpretation is awkward.

Evolution of Gene function (multiple functions)

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

Full Markov Transition Matrix

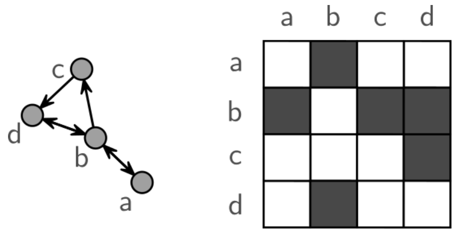


- 512 parameters
- Finding this many parameters is not easy.
- Even if you can, interpretation is awkward.

Social Network Analysis may help us...

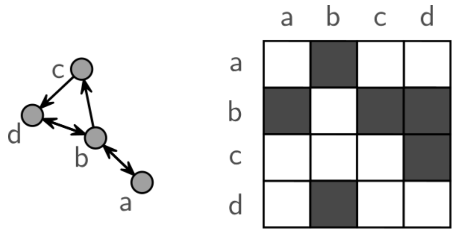
Exponential Random Graph Models (ERGMs)

Social Network



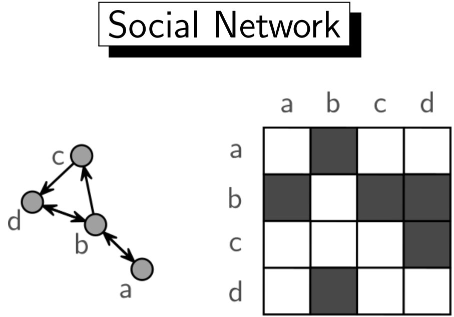
Exponential Random Graph Models (ERGMs)

Social Network



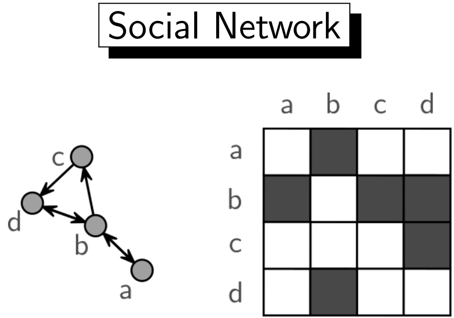
- Not about individual ties.

Exponential Random Graph Models (ERGMs)



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

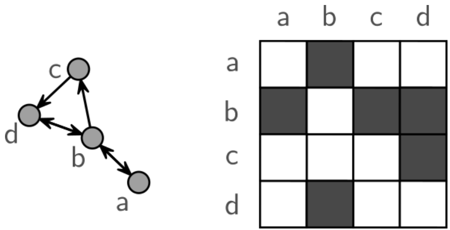
Exponential Random Graph Models (ERGMs)



- Not about individual ties.
- Statistical inference on *motifs* (triangles, dyads, homophily, etc.)
- Literature about ERGMs is vast, a.k.a. a low-hanging fruit.

Exponential Random Graph Models (ERGMs)

Social Network



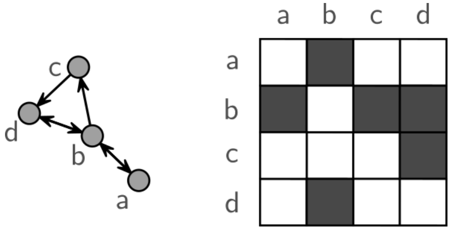
- Not about individual ties.
- Statistical inference on *motifs* (triangles, dyads, homophily, etc.)
- Literature about ERGMs is vast, a.k.a. a low-hanging fruit.

Ultimately...
ERGM $\equiv$ Modeling binary arrays

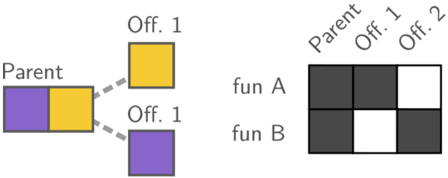
Preliminaries
Evolution of Gene Function
Mechanistic Machine Learning
Proof of Concept

Exponential Random Graph Models (ERGMs)

Social Network



Evolutionary Event



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

Preliminaries

Evolution of Gene Function

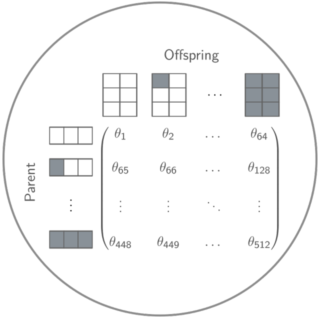
Mechanistic Machine Learning

Proof of Concept

Evolution of Gene function (multiple functions) (cont.)

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

Full Markov Transition Matrix

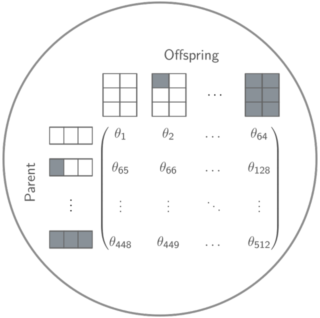


512 parameters

Evolution of Gene function (multiple functions) (cont.)

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

Full Markov Transition Matrix



512 parameters

Sufficient statistics

$$\left(\theta_{\text{gain1}}\ \theta_{\text{gain2}}\ \theta_{\text{gain3}}\right)$$

$$\left(\theta_{\text{loss1}}\ \theta_{\text{loss2}}\ \theta_{\text{loss3}}\right)$$

$$\theta_{\text{neofun}}\ \theta_{\text{subfunc}}$$

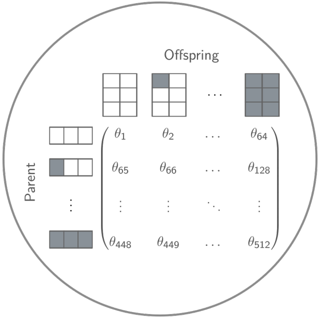
$$\left(\theta_{\text{root1}}\ \theta_{\text{root2}}\ \theta_{\text{root3}}\right)$$

11 parameters (for example)

Evolution of Gene function (multiple functions) (cont.)

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

Full Markov Transition Matrix



512 parameters

$(\theta_{\text{gain1}}\ \theta_{\text{gain2}}\ \theta_{\text{gain3}})$

$(\theta_{\text{loss1}}\ \theta_{\text{loss2}}\ \theta_{\text{loss3}})$

$\theta_{\text{neofun}}\ \theta_{\text{subfunc}}$

$(\theta_{\text{root1}}\ \theta_{\text{root2}}\ \theta_{\text{root3}})$

Easier to fit
Easier to interpret

Sufficient statistics

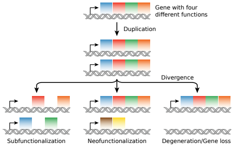
$(\theta_{\text{gain1}}\ \theta_{\text{gain2}}\ \theta_{\text{gain3}})$

$(\theta_{\text{loss1}}\ \theta_{\text{loss2}}\ \theta_{\text{loss3}})$

$\theta_{\text{neofun}}\ \theta_{\text{subfunc}}$

$(\theta_{\text{root1}}\ \theta_{\text{root2}}\ \theta_{\text{root3}})$

11 parameters (for example)



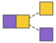
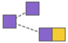
Rep.	Description	Definition
	Gain of function	$(1 - x_p) \sum_{n : n \in Off} x_n$
	Loss of function	$x_p \sum_{n : n \in Off} (1 - x_n)$
	Subfunctionalization	$x_p^k x_p^j \sum_{n \neq m} x_n^k (1 - x_n^j)(1 - x_m^k) x_m^j$
	Neofunctionalization	$x_p^k (1 - x_p^j) \sum_{n \neq m} x_n^k (1 - x_n^j)(1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

Table: Example of sufficient statistics for evolutionary transitions.

GEESE: **GE**ne functional **E**volution using **S**ufici**E**ncy

I implemented what I just described in a C++ library with a companion R package called geese. The question is: How much do we earn by using these motifs?

GEESE: **GE**ne functional **E**volution using **S**ufici**E**ncy

I implemented what I just described in a C++ library with a companion R package called geese. The question is: How much do we earn by using these motifs?

- Using 37 phylogenetic trees featuring 401 go annotations.

GEESE: **GE**ne functional **E**volution using **S**ufici**E**ncy

I implemented what I just described in a C++ library with a companion R package called geese. The question is: How much do we earn by using these motifs?

- Using 37 phylogenetic trees featuring 401 go annotations.
- **aphylo**: Fitted a *simple gain/loss* of function model.

GEESE: **GE**ne functional **E**volution using **S**ufici**E**ncy

I implemented what I just described in a C++ library with a companion R package called geese. The question is: How much do we earn by using these motifs?

- Using 37 phylogenetic trees featuring 401 go annotations.
- **aphylo**: Fitted a *simple gain/loss* of function model.
- **GEESE**: Fitted an evolutionary model controlling for *functional preservation* (i.e., like neofun or subfun.)

GEESE: **GE**ne functional **E**volution using **S**ufici**E**ncy

I implemented what I just described in a C++ library with a companion R package called geese. The question is: How much do we earn by using these motifs?

- Using 37 phylogenetic trees featuring 401 go annotations.
- **aphylo**: Fitted a *simple gain/loss* of function model.
- **GEESE**: Fitted an evolutionary model controlling for *functional preservation* (i.e., like neofun or subfun.)
- Fitted both of them using MCMC.

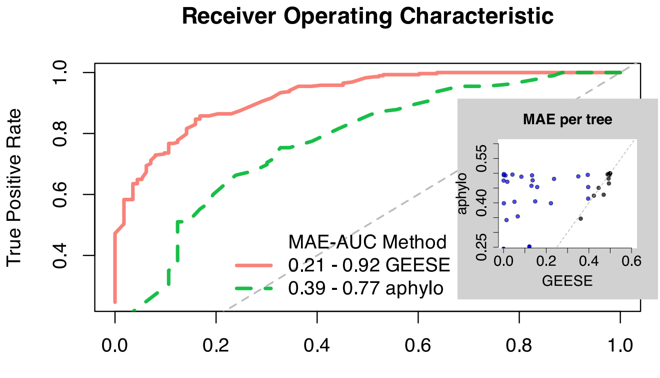
GEESE: **GE**ne functional **E**volution using **S**ufici**E**ncy

I implemented what I just described in a C++ library with a companion R package called geese. The question is: How much do we earn by using these motifs?

- Using 37 phylogenetic trees featuring 401 go annotations.
- **aphylo**: Fitted a *simple gain/loss* of function model.
- **GEESE**: Fitted an evolutionary model controlling for *functional preservation* (i.e., like neofun or subfun.)
- Fitted both of them using MCMC.
- Used LOO cross-validation to compute aggregated AUCs and MAE.

GEESE for predicting gene function (cont.)

How much can we gain from a joint dist. model?



Phylogenetic models predicting 401 GO annotations.

Just controlling for preservation (having only one duplicate changing) significantly improves our predictions.

Table of Contents

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

Mechanistic Machine Learning: State-of-the-art

- After all the data pouring, attention to causal inference and mechanistic models is coming back[3,14]
- Applications in Physics, Chemistry, Biomedical Imaging, and Biology[16,11,7,1] show the benefits of combining the two approaches.

Mechanistic Models	Machine Learning Models
Inference driven (causality)	Data-driven (prediction)
Great for small datasets	Lots of points to "learn"
Not the most accurate	Great for big data

Important: Mechanistic Machine Learning **is not** domain-knowledge aided feature engineering. You need a whole other model to complement the ML algorithm.

Three strategies

Figure: "A van Gogh-style painting of an android holding a large biology book in one hand and a computer in another, examining an evolutionary tree that, instead of leaves, have genes."–DALL-E's interpretation of my description (link)

Three strategies

a. **ML Correction**: Use machine learning to learn the errors of a mechanistic model.

Figure: "A van Gogh-style painting of an android holding a large biology book in one hand and a computer in another, examining an evolutionary tree that, instead of leaves, have genes."–DALL-E's interpretation of my description (link)

Three strategies

a. **ML Correction**: Use machine learning to learn the errors of a mechanistic model.
b. **Mechanistic Feature**: Add mechanistic predictions as a feature of a machine learning model.

Figure: "A van Gogh-style painting of an android holding a large biology book in one hand and a computer in another, examining an evolutionary tree that, instead of leaves, have genes."–DALL-E's interpretation of my description (link)

Three strategies

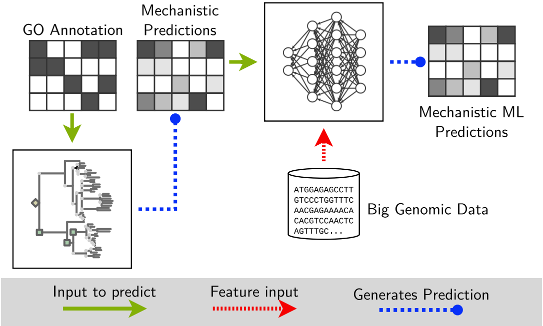
a. **ML Correction**: Use machine learning to learn the errors of a mechanistic model.
b. **Mechanistic Feature**: Add mechanistic predictions as a feature of a machine learning model.
c. **Mechanistic Penalty**: Add constraints to the ML algorithm based on a mechanistic model.

Figure: "A van Gogh-style painting of an android holding a large biology book in one hand and a computer in another, examining an evolutionary tree that, instead of leaves, have genes."–DALL-E's interpretation of my description (link)

Three strategies

a. ML Correction

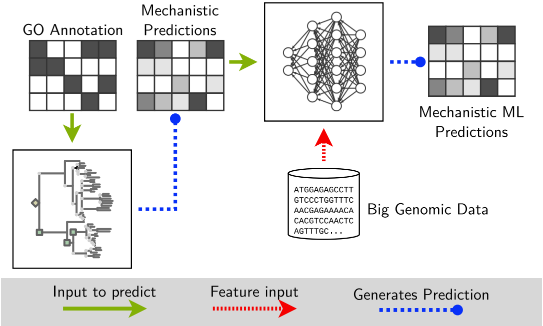
1. Fit the mechanistic model using GEESE



Three strategies

a. ML Correction

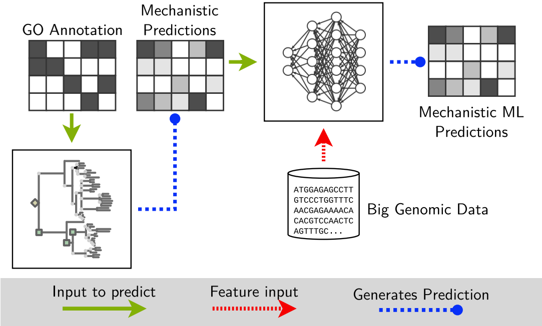
1. Fit the mechanistic model using GEESE
2. Generate the mechanistic-based predictions, $\hat{\mathbf{y}}^{GEESE}$,



Three strategies

a. ML Correction

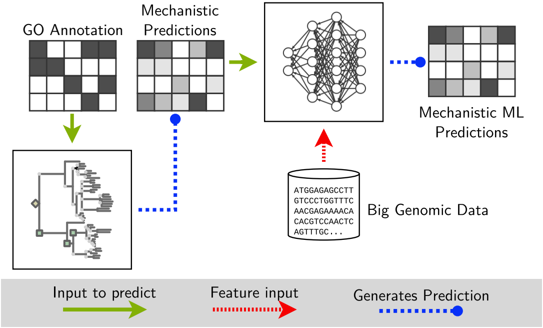
1. Fit the mechanistic model using GEESE
2. Generate the mechanistic-based predictions, $\hat{\mathbf{y}}^{GEESE}$,
3. fit an ML model $f(X, \Omega)$ to predict $\varepsilon \equiv (\mathbf{y} - \hat{\mathbf{y}}^{GEESE})$,



Three strategies

a. ML Correction

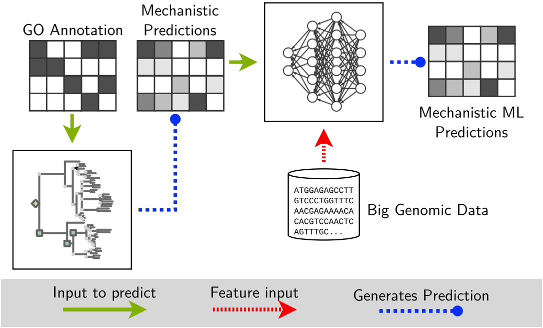
1. Fit the mechanistic model using GEESE
2. Generate the mechanistic-based predictions, $\hat{\mathbf{y}}^{GEESE}$,
3. fit an ML model $f(X, \Omega)$ to predict $\varepsilon \equiv (\mathbf{y} - \hat{\mathbf{y}}^{GEESE})$,
4. generate the predictions of $\hat{\varepsilon}$, and



Three strategies

a. ML Correction

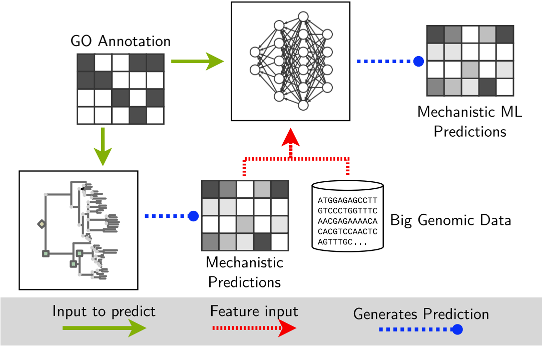
1. Fit the mechanistic model using GEESE
2. Generate the mechanistic-based predictions, $\hat{\mathbf{y}}^{GEESE}$,
3. fit an ML model $f(X, \Omega)$ to predict $\varepsilon \equiv (\mathbf{y} - \hat{\mathbf{y}}^{GEESE})$,
4. generate the predictions of $\hat{\varepsilon}$, and
5. Compute the Mechanistic-ML predictions as $\hat{\mathbf{y}}^{MML1} \equiv \hat{\mathbf{y}}^{GEESE} + \hat{\varepsilon}$



Three strategies

b. Mechanistic Feature

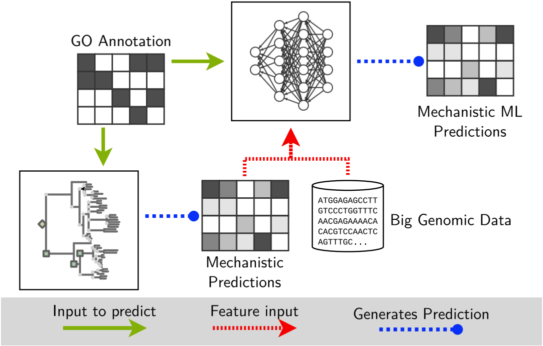
1. Fit the mechanistic model using GEESE,



Three strategies

b. Mechanistic Feature

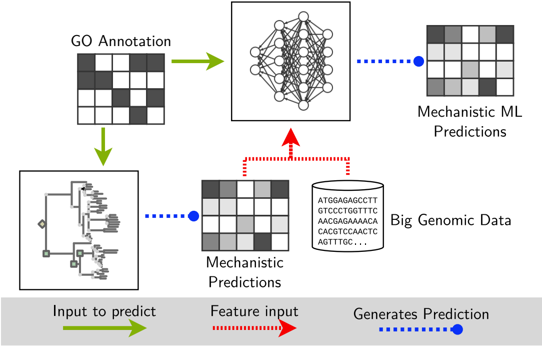
1. Fit the mechanistic model using GEESE,
2. generate the mechanistic-based predictions, $\hat{\mathbf{y}}^{GEESE}$,



Three strategies

b. Mechanistic Feature

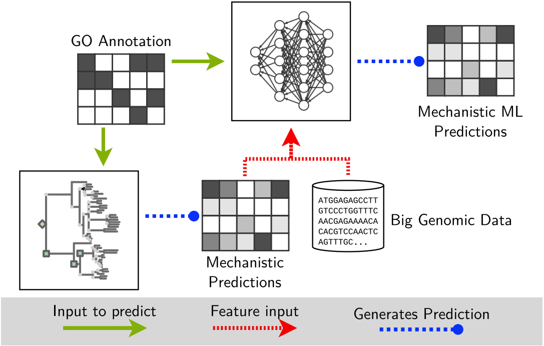
1. Fit the mechanistic model using GEESE,
2. generate the mechanistic-based predictions, $\hat{\mathbf{y}}^{GEESE}$,
3. fit an ML model that uses the mechanistic predictions as features, $f(X, \Omega, \hat{\mathbf{y}}^{GEESE})$, and



Three strategies

b. Mechanistic Feature

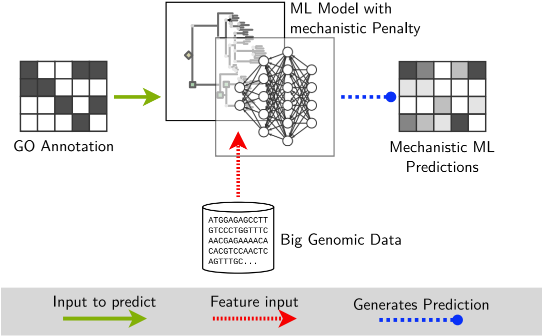
1. Fit the mechanistic model using GEESE,
2. generate the mechanistic-based predictions, $\hat{\mathbf{y}}^{GEESE}$,
3. fit an ML model that uses the mechanistic predictions as features, $f(X, \Omega, \hat{\mathbf{y}}^{GEESE})$, and
4. Compute the Mechanistic-ML predictions as $\hat{\mathbf{y}}^{MML2} \equiv f(X, \Omega, \hat{\mathbf{y}}^{GEESE})$



Three strategies

c. Mechanistic Penalty

1. Fit the mechanistic model using GEESE and store the parameter estimates $\hat{\theta}$,



Three strategies

c. Mechanistic Penalty

1. Fit the mechanistic model using GEESE and store the parameter estimates $\hat{\theta}$,
2. minimize the following loss function:

$$L(y^{obs}_{,} X, \Omega) - \mathcal{L}\left(f(y^{obs}_{,} X, \Omega)\right)_{GEESE},$$

where $\mathcal{L}(\cdot)$ is the likelihood function under GEESE.

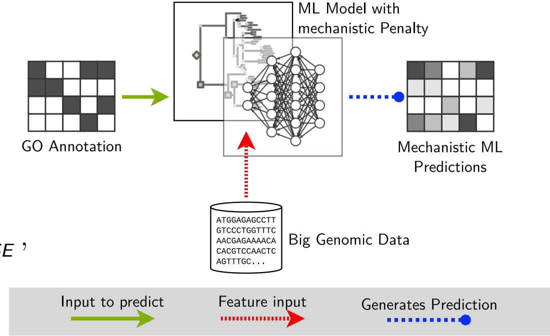


Table of Contents

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

Beyond GO and Trees... Bgee

The Bgee project "is a **database** for retrieval and **comparison of gene expression** patterns **across multiple animal species**. It provides an intuitive answer to the question 'where is a gene expressed?' and supports research in cancer and agriculture as well as evolutionary biology." – Bastian et al. [4]

- Raw expression annotations.
- Standardized expression scores (so can compare across species/tissues).
- And also yes/no expression annotations based on the standardized scores.

Beyond GO and Trees... Bgee (cont.)

Divergence across species in gene expression levels has been linked to evolutionary events,[13,9] *i.e.*, expression levels clustered phylogenies.

Thinking of different ways to use it as:

- As an additional feature for our model: "Given the phylo, observed annotations, **and expression levels in *n* tissues**, ..."
- As 0/1 variable (expression is present/absent) to predict in our model: "Model the evolution of gene function **and expression**."
- As part of a prediction model in, say, a Machine Learning Model.

What went into the blender

What went into the blender

Data Feats

- Bgee 15 dataset: approx 7 billion annotations for 1.5 million genes.
- Our dataset: 1,484 predictions for 1,318 genes.
- Search by Gene name: 9,923,427 Bgee annotations.

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

What went into the blender

Data Feats

- Bgee 15 dataset: approx 7 billion annotations for 1.5 million genes.
- Our dataset: 1,484 predictions for 1,318 genes.
- Search by Gene name: 9,923,427 Bgee annotations.

Final model

- 10 GO terms (in a full-Markov model, this is 1 MM params).
- 278 annotations for 256 genes.
- 10 GEESE predictions for each gene.
- 46 Bgee score for gene expression computed as **mean expression score by gene-genus**

What went into the blender

Data Feats

- Bgee 15 dataset: approx 7 billion annotations for 1.5 million genes.
- Our dataset: 1,484 predictions for 1,318 genes.
- Search by Gene name: 9,923,427 Bgee annotations.

Final model

- 10 GO terms (in a full-Markov model, this is 1 MM params).
- 278 annotations for 256 genes.
- 10 GEESE predictions for each gene.
- 46 Bgee score for gene expression computed as **mean expression score by gene-genus**

GO terms: GO:0004672, GO:0004713, GO:0004867, GO:0005730, GO:0005829, GO:0005886, GO:0006468, GO:0009408, GO:0015020, GO:0060070

Genus: Anguilla, Anolis, Astatotilapia, Astyanax, Bos, Branchiostoma, Caenorhabditis, Callithrix, Canis, Capra, Cavia, Cercocebus, Chlorocebus, Danio, Drosophila, Equus, Esox, Felis, Gadus, Gallus, Gasterosteus, Gorilla, Heterocephalus, Homo, Latimeria, Lepisosteus, Macaca, Manis, Meleagris, Microcebus, Monodelphis, Mus, Neolamprologus, Nothobranchius, Ornithorhynchus, Oryctolagus, Oryzias, Ovis, Pan, Papio, Poecilia, Rattus, Salmo, Scophthalmus, Sus, Xenopus

Mechanistic ML

We are comparing three models:

$\boxed{\text{GEESE}}$

Phylogenetic based predictions (evolution of gene function)

Mechanistic ML

We are comparing three models:

GEESE

Phylogenetic based predictions (evolution of gene function)

Bgee

Linear Prob. model using expression as predictors.

Mechanistic ML

We are comparing three models:

GEESE

Phylogenetic based predictions (evolution of gene function)

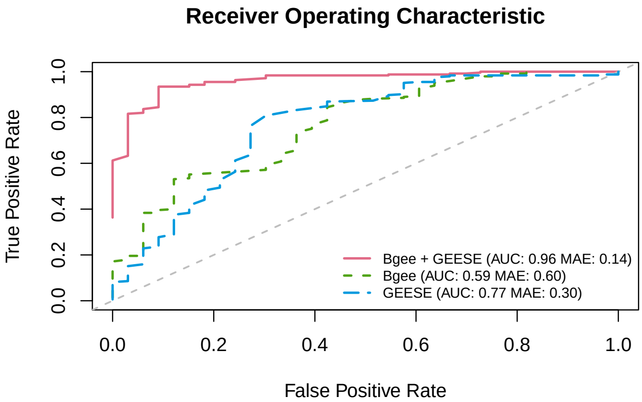
Bgee

Linear Prob. model using expression as predictors.

GEESE + Bgee

Linear Prob. model using expression as predictors **and** predictions made by GEESE.

Mechanistic ML (prelim res.)



Linear probability models predicting 278 GO annotations.

Both AUC and MAE were computed only using predictions for which we knew the true value.

Discussion

Discussion

Gene function

- We are racing to discover what genes do.
- Experimental assessment is expensive (money and time,) $\rightarrow$ automatic annotations.
- Many ways to do it (seq. homology, evolutionary theory, ML, etc.)
- The best methods use ML (pattern discovery)... but none (AFAIK) are based on bio. theory.

Discussion

Gene function

- We are racing to discover what genes do.
- Experimental assessment is expensive (money and time,) $\rightarrow$ automatic annotations.
- Many ways to do it (seq. homology, evolutionary theory, ML, etc.)
- The best methods use ML (pattern discovery)... but none (AFAIK) are based on bio. theory.

Evol. Model

- We proposed an Evolutionary model of Gene Function.
- This new model, GEESE, uses sufficiency to reduce "Markov complexity."
- We showed it really helps.

Discussion

Preliminaries
Evolution of Gene Function
Mechanistic Machine Learning
Proof of Concept

Gene function

- We are racing to discover what genes do.
- Experimental assessment is expensive (money and time,) → automatic annotations.
- Many ways to do it (seq. homology, evolutionary theory, ML, etc.)
- The best methods use ML (pattern discovery)... but none (AFAIK) are based on bio. theory.

Evol. Model

- We proposed an Evolutionary model of Gene Function.
- This new model, GEESE, uses sufficiency to reduce "Markov complexity."
- We showed it really helps.

Mechanistic ML

- Mechanistic Machine Learning (mixing theory-based models with ML) promises improved predictions.
- I showed an application using gene expression (Bgee).
- Adding our mechanistic predictions (based on GEESE) boosted quality

Thank you!

Predicting of Gene Functions by Leveraging Biological Insights with Mechanistic Machine Learning

George G. Vega Yon, Ph.D.
george.vegayon@utah.edu
Division of Epidemiology @ University of Utah

Jan 11th, 2023 @ Data Science Seminar

Preliminaries
Evolution of Gene Function
Mechanistic Machine Learning
Proof of Concept

References I

[1] Nada Al taweraqi and Ross D. King. "Improved Prediction of Gene Expression through Integrating Cell Signalling Models with Machine Learning". In: BMC Bioinformatics 23.1 (Aug. 6, 2022), p. 323. ISSN: 1471-2105. DOI: 10.1186/s12859-022-04787-8. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04787-8 (visited on 12/07/2022).

[2] Stephen F. Altschul et al. "Basic Local Alignment Search Tool". In: Journal of Molecular Biology 215.3 (Oct. 1990), pp. 403–410. ISSN: 00222836. DOI: 10.1016/S0022-2836(05)80360-2. URL: https://linkinghub.elsevier.com/retrieve/pii/S0022283605803602 (visited on 12/08/2022).

[3] Ruth E. Baker et al. "Mechanistic Models versus Machine Learning, a Fight Worth Fighting for the Biological Community?" In: Biology Letters 14.5 (May 2018), p. 20170660. ISSN: 1744-9561, 1744-957X. DOI: 10.1098/rsbl.2017.0660. URL: https://royalsocietypublishing.org/doi/10.1098/rsbl.2017.0660 (visited on 12/01/2022).

References II

[4] Frederic B Bastian et al. "The Bgee Suite: Integrated Curated Expression Atlas and Comparative Transcriptomics in Animals". In: Nucleic Acids Research 49.D1 (Jan. 8, 2021), pp. D831–D847. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkaa793. URL: https://academic.oup.com/nar/article/49/D1/D831/5920517 (visited on 12/06/2022).

[5] B.E. Engelhardt et al. "Genome-scale phylogenetic function annotation of large and diverse protein families". In: Genome research 21.11 (2011), pp. 1969–1980. DOI: 10.1101/gr.104687.109.

[6] B.E. Engelhardt et al. "Protein Molecular Function Prediction by Bayesian Phylogenomics". In: PLOS Computational Biology 1.5 (2005). DOI: 10.1371/journal.pcbi.0010045.

References III

[7] Nathan Gaw et al. “Integration of Machine Learning and Mechanistic Models Accurately Predicts Variation in Cell Density of Glioblastoma Using Multiparametric MRI”. In: Scientific Reports 9.1 (Dec. 2019), p. 10063. ISSN: 2045-2322. DOI: 10.1038/s41598-019-46296-4. URL: http://www.nature.com/articles/s41598-019-46296-4 (visited on 12/09/2022).

[8] Vladimir Gligorijević et al. “Structure-Based Protein Function Prediction Using Graph Convolutional Networks”. In: Nature Communications 12.1 (May 26, 2021), p. 3168. ISSN: 2041-1723. DOI: 10.1038/s41467-021-23303-9. URL: https://www.nature.com/articles/s41467-021-23303-9 (visited on 12/08/2022).

[9] Andrea Hodgins-Davis and Jeffrey P. Townsend. “Evolving Gene Expression: From G to E to G×E”. In: Trends in Ecology & Evolution 24.12 (Dec. 2009), pp. 649–658. ISSN: 01695347. DOI: 10.1016/j.tree.2009.06.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169534709002213 (visited on 11/16/2022).

References IV

[10] Aashish Jain and Daisuke Kihara. "Phylo-PFP: improved automated protein function prediction using phylogenetic distance of distantly related sequences". In: Bioinformatics 35 (5 Mar. 2019), pp. 753–759. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty704.

[11] Kjell Jorner et al. "Machine Learning Meets Mechanistic Modelling for Accurate Prediction of Experimental Activation Energies". In: Chemical Science 12.3 (2021), pp. 1163–1175. ISSN: 2041-6520, 2041-6539. DOI: 10.1039/D0SC04896H. URL: http://xlink.rsc.org/?DOI=D0SC04896H (visited on 12/07/2022).

[12] Maxat Kulmanov and Robert Hoehndorf. "DeepGOPlus: Improved Protein Function Prediction from Sequence". In: Bioinformatics (July 27, 2019). Ed. by Lenore Cowen, btz595. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btz595. URL: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz595/5539866 (visited on 11/17/2022).

References V

[13] B. Nabholz, H. Ellegren, and J. B. W. Wolf. "High Levels of Gene Expression Explain the Strong Evolutionary Constraint of Mitochondrial Protein-Coding Genes". In: Molecular Biology and Evolution 30.2 (Feb. 1, 2013), pp. 272–284. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/mss238. URL: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mss238 (visited on 11/16/2022).

[14] Judea Pearl. "The Seven Tools of Causal Inference, with Reflections on Machine Learning". In: Communications of the ACM 62.3 (Feb. 21, 2019), pp. 54–60. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3241036. URL: https://dl.acm.org/doi/10.1145/3241036 (visited on 11/18/2022).

[15] G.G. Vega Yon et al. "Bayesian parameter estimation for automatic annotation of gene functions using observational data and phylogenetic trees". In: PLoS Comput Biol 17 (2021), e1007948. DOI: 10.1371/journal.pcbi.1007948.

References VI

[16] Jared Willard et al. "Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems". In: ACM Computing Surveys (Mar. 25, 2022), p. 3514228. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3514228. URL: https://dl.acm.org/doi/10.1145/3514228 (visited on 11/11/2022).

[17] Ronghui You et al. "GOLabeler: Improving Sequence-Based Large-Scale Protein Function Prediction by Learning to Rank". In: Bioinformatics 34.14 (July 15, 2018). Ed. by Jonathan Wren, pp. 2465–2473. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bty130. URL: https://academic.oup.com/bioinformatics/article/34/14/2465/4924212 (visited on 11/17/2022).

[18] Naihui Zhou et al. "The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens". In: Genome Biology 20 (1 2019), pp. 1–23. ISSN: 1474760X. DOI: 10.1186/s13059-019-1835-8.

References VII

[19] Naihui Zhou et al. "The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens". In: Genome Biology 20 (1 2019), pp. 1–23. ISSN: 1474760X. DOI: 10.1186/s13059-019-1835-8.

Tree likelihoods: Felsenstein's Pruning algorithm

All possible transitions from $\mathbf{x}_n$

Transition Probability (ERGM)

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

Tree likelihoods: Felsenstein's Pruning algorithm

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

All possible transitions from $\mathbf{x}_n$

Transition Probability (ERGM)

Model Parameters

Vector of Sufficient Statistics

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

Normalizing Constant

the ***lingua franca*** of SNA

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

Some computational features of **barry**