Statistical and computational methods for bioinformatics and social network analysis

or how did I learn to stop worrying and love the bomb

George G Vega Yon

University of Southern California, Department of Preventive Medicine

October 4, 2019

Contents

Paper 1: Exponential Random Graph Models for Small Networks

Paper 2: On the prediction of gene functions using phylogenetic trees

Future directions

Things that are very interesting but I most probably won't have any time to discuss with the attendees

References

Contents

Paper 1: Exponential Random Graph Models for Small Networks

Paper 2: On the prediction of gene functions using phylogenetic trees

Future directions

Things that are very interesting but I most probably won't have any time to discuss with the attendees

References

What are Exponential Random Graph Models

Keck School of Medicine of USC

Exponential Family Random Graph Models, aka ERGMs are:

What are Exponential Random Graph Models

Keck School of Medicine of USC

Exponential Family Random Graph Models, aka ERGMs are:

- Statistical models of (social) networks

What are Exponential Random Graph Models

Keck School of Medicine of USC

Exponential Family Random Graph Models, aka ERGMs are:

- Statistical models of (social) networks
- In simple terms: statistical inference on what network patterns/structures/motifs govern the data-generating process

What are Exponential Random Graph Models

Keck School of Medicine of USC

Exponential Family Random Graph Models, aka ERGMs are:

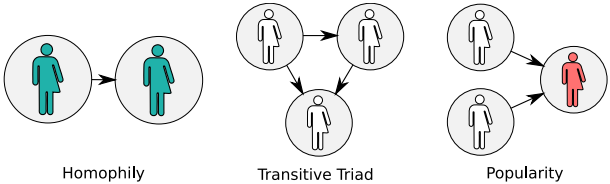
- Statistical models of (social) networks
- In simple terms: statistical inference on what network patterns/structures/motifs govern the data-generating process

What are Exponential Random Graph Models

Keck School of Medicine of USC

Exponential Family Random Graph Models, aka ERGMs are:

- Statistical models of (social) networks
- In simple terms: statistical inference on what network patterns/structures/motifs govern the data-generating process



ERGMs (cont'd)

Keck School of Medicine of USC

A vector of model parameters

A vector of sufficient statistics

$$\Pr(\mathbf{Y} = \mathbf{y} \mid \theta, \mathbf{X}) = \frac{\exp\{\theta^{\mathbf{t}} s(\mathbf{y}, \mathbf{X})\}}{\sum_{\mathbf{y}' \in \mathcal{Y}} \exp\{\theta^{\mathbf{t}} s(\mathbf{y}', \mathbf{X})\}}, \quad \forall \mathbf{y} \in \mathcal{Y}$$

Observed data

The normalizing constant

All possible networks

more on terms

ERGMs: State of the Art

Keck School of Medicine of USC

ERGMs: State of the Art

Keck School of Medicine of USC

Small-large (dozens to a couple of thousand vertices) networks

- Markov Chain Monte Carlo (MCMC) based approaches like MC-MLE or Robbins-Monro Stochastic approximation. details
- Maximum Pseudo Likelihood (MPLE)

ERGMs: State of the Art

Keck School of Medicine of USC

Small-large (dozens to a couple of thousand vertices) networks

- Markov Chain Monte Carlo (MCMC) based approaches like MC-MLE or Robbins-Monro Stochastic approximation. details
- Maximum Pseudo Likelihood (MPLE)

large-huge networks (up to the millions of vertices)

- Semi-parametric bootstrap
- Conditional joint estimation (like snowball sampling, a.k.a. divide and conquer)
- Equilibrium Expectation Algorithm (millions of vertices)

ERGMs: State of the Art

Keck School of Medicine of USC

Small-large (dozens to a couple of thousand vertices) networks

- Markov Chain Monte Carlo (MCMC) based approaches like MC-MLE or Robbins-Monro Stochastic approximation. details
- Maximum Pseudo Likelihood (MPLE)

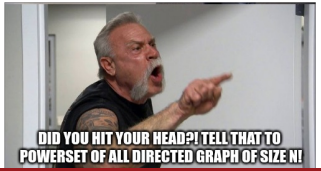
large-huge networks (up to the millions of vertices)

- Semi-parametric bootstrap
- Conditional joint estimation (like snowball sampling, a.k.a. divide and conquer)
- Equilibrium Expectation Algorithm (millions of vertices)

What about tiny to small networks?

A frame with examples of small networks...

Keck School of Medicine of USC



ERGMs for small networks

Keck School of Medicine of USC

MC-MLE works great (we have some simulations showing this), but it has some problems:

ERGMs for small networks

Keck School of Medicine of USC

MC-MLE works great (we have some simulations showing this), but it has some problems:

- Possible accuracy issues (error rates)

What shall we do then?

ERGMs for small networks

Keck School of Medicine of USC

MC-MLE works great (we have some simulations showing this), but it has some problems:

- Possible accuracy issues (error rates)
- Prone to degeneracy problems (sampling and existance of MLE)

What shall we do then?

ERGMs for small networks

Keck School of Medicine of USC

MC-MLE works great (we have some simulations showing this), but it has some problems:

- Possible accuracy issues (error rates)
- Prone to degeneracy problems (sampling and existance of MLE)
- It is not MLE,

What shall we do then?

ERGMs for Small Networks

Keck School of Medicine of USC

ERGMs for Small Networks

Keck School of Medicine of USC

- In the case of small-enough networks, computation of the likelihood becomes computationally feasible.

ERGMs for Small Networks

Keck School of Medicine of USC

- In the case of small-enough networks, computation of the likelihood becomes computationally feasible.
- For example, a network with 5 nodes has 1,048,576 unique configurations.

ERGMs for Small Networks

Keck School of Medicine of USC

- In the case of small-enough networks, computation of the likelihood becomes computationally feasible.
- For example, a network with 5 nodes has 1,048,576 unique configurations.
- This allow us to directly compute **the normalizing constant**.

ERGMs for Small Networks

Keck School of Medicine of USC

- In the case of small-enough networks, computation of the likelihood becomes computationally feasible.
- For example, a network with 5 nodes has 1,048,576 unique configurations.
- This allow us to directly compute **the normalizing constant**.
- Using the exact likelihood opens a huge window of methodological-possibilities.

ERGMs for Small Networks

Keck School of Medicine of USC

- In the case of small-enough networks, computation of the likelihood becomes computationally feasible.
- For example, a network with 5 nodes has 1,048,576 unique configurations.
- This allow us to directly compute **the normalizing constant**.
- Using the exact likelihood opens a huge window of methodological-possibilities.

ERGMs for Small Networks

Keck School of Medicine of USC

- In the case of small-enough networks, computation of the likelihood becomes computationally feasible.
- For example, a network with 5 nodes has 1,048,576 unique configurations.
- This allow us to directly compute **the normalizing constant**.
- Using the exact likelihood opens a huge window of methodological-possibilities.

We implemented this and more in the `ergmito` R package more

Paper 1 Simulation Studies

Keck School of Medicine of USC

In order to compare the MLE with the MC-MLE estimation method, we performed a simulation study with the following features:

Paper 1 Simulation Studies

Keck School of Medicine of USC

In order to compare the MLE with the MC-MLE estimation method, we performed a simulation study with the following features:

- Draw 20,000 samples of groups of small networks

Paper 1 Simulation Studies

Keck School of Medicine of USC

In order to compare the MLE with the MC-MLE estimation method, we performed a simulation study with the following features:

- Draw 20,000 samples of groups of small networks
- Each group had prescribed: (model parameters, number of networks, sizes of the networks)

Paper 1 Simulation Studies

Keck School of Medicine of USC

In order to compare the MLE with the MC-MLE estimation method, we performed a simulation study with the following features:

- Draw 20,000 samples of groups of small networks
- Each group had prescribed: (model parameters, number of networks, sizes of the networks)
- Each group could have from 5 to 300 small networks

Paper 1 Simulation Studies

Keck School of Medicine of USC

In order to compare the MLE with the MC-MLE estimation method, we performed a simulation study with the following features:

- Draw 20,000 samples of groups of small networks
- Each group had prescribed: (model parameters, number of networks, sizes of the networks)
- Each group could have from 5 to 300 small networks
- We estimated the models using MC-MLE and MLE.

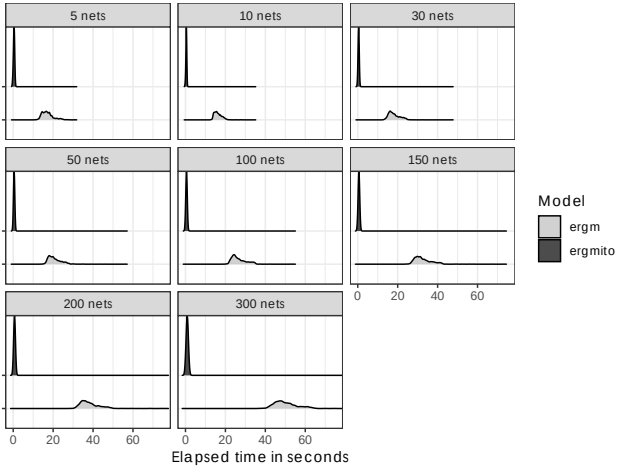
Paper 1 Simulation Studies: Empirical Type I error

Keck School of Medicine of USC

		P(Type I error)		
Sample size	N. Simulations	MC-MLE	MLE	chi2
5	2,189	0.084	0.057	11.71 ***
10	2,330	0.070	0.045	12.46 ***
15	2,395	0.084	0.066	5.55 *
20	2,430	0.074	0.060	3.58
30	2,460	0.057	0.052	0.67
50	2,495	0.046	0.044	0.17
100	2,499	0.048	0.048	0.00

Table 1 Empirical Type I error rates. The χ^2 statistic is from a 2-sample test for equality of proportions, and the significance levels are given by *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$. The lack of fitted samples in some levels is due to failure of the estimation method.

Paper 1 Simulation Studies: Elapsed time



more results

Paper 1: Key takeaway

Keck School of Medicine of USC

Paper 1: Key takeaway

Keck School of Medicine of USC

- Developed a new extension of ERGMs using exact statistics for small networks (families, teams, ego-centered, etc.)

Paper 1: Key takeaway

Keck School of Medicine of USC

- Developed a new extension of ERGMs using exact statistics for small networks (families, teams, ego-centered, etc.)
- Performance: Same (un)bias, Lower Type I error rates, (way) faster.

Paper 1: Key takeaway

Keck School of Medicine of USC

- Developed a new extension of ERGMs using exact statistics for small networks (families, teams, ego-centered, etc.)
- Performance: Same (un)bias, Lower Type I error rates, (way) faster.
- Opens the door the new methods.

Contents

Paper 1: Exponential Random Graph Models for Small Networks

Paper 2: On the prediction of gene functions using phylogenetic trees

Future directions

Things that are very interesting but I most probably won't have any time to discuss with the attendees

References

Phylogenetic Trees

Keck School of Medicine of USC

Keck School of Medicine of USC

- It can be very general: think of the tree of life

Phylogenetic Trees

Keck School of Medicine of USC

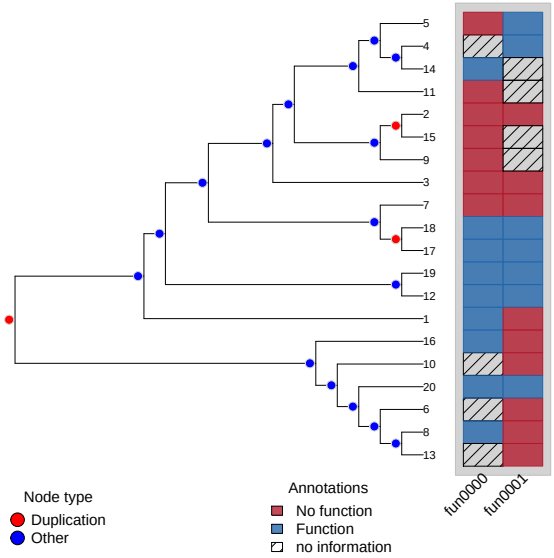
- It can be very general: think of the tree of life
- Nowadays, thanks to gene-sequencing techniques, we are building trees at the gene level (using sequence-alignment methods, i.e. comparing gene sequences to see how much similar/different two genes are between and within species (whattt!)).

Phylogenetic Trees

Keck School of Medicine of USC

- It can be very general: think of the tree of life
- Nowadays, thanks to gene-sequencing techniques, we are building trees at the gene level (using sequence-alignment methods, i.e. comparing gene sequences to see how much similar/different two genes are between and within species (whattt!)).
- A single phylogenetic tree can host multiple species

Annotated Phylogenetic Tree



Keck School of Medicine of USC

Accession	GO:0060047
Name	heart contraction
Ontology	biological_process
Synonyms	heart beating, cardiac contraction, hemolymph circulation
Alternate IDs	None
Definition	The multicellular organismal process in which the heart decreases in volume in a characteristic way to propel blood through the body. Source: GOC:dph

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

pthr10037

pthr11521

pthr11521

pthr24356

Speciation

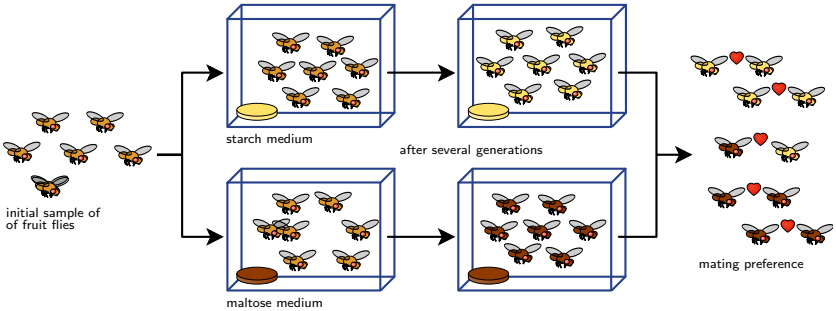


Figure 1 Dodd (1989): After one year of isolation, flies showed a significant level or assortativity in mating (wikimedia)

Duplication

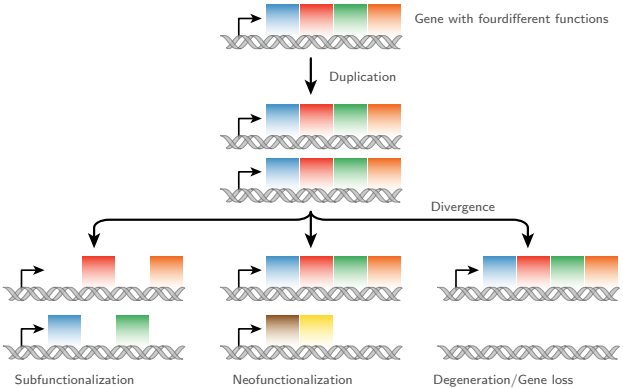


Figure 2 A key part of molecular innovation, gene duplication provides opportunity for new functions to emerge (wikimedia)

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

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;

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;

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**

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. $(1 - \mu_1)$;

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;

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

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;

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**

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;

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;

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

An evolutionary model of gene functions: Formal statement

Keck School of Medicine of USC

- The whole is based on the markov-assumption: The current state of the gene can be fully explained by its parent(s).

An evolutionary model of gene functions: Formal statement

Keck School of Medicine of USC

- The whole is based on the markov-assumption: The current state of the gene can be fully explained by its parent(s).
- For this we use Felsensteins' pruning algorithm (also known as...)
  Formally

$$P(x = 1) = P(x = 1|x_p = 0)P(\text{Gain}) + P(x = 1|x_p = 1)P(\text{No loss})$$

Paper 2: Estimation of the model

Keck School of Medicine of USC

The model has (so far) 7 fixed parameters

ψ_0, ψ_1	Probability of making a mistake (mislabel)
μ_{d0}, μ_{d1}	Probability of functional gain/loss (duplication nodes)
μ_{s0}, μ_{s1}	Probability of functional gain/loss (other nodes)
π	Probability that the root has the function

Paper 2: Estimation of the model (cont'd)

Keck School of Medicine of USC

- We developed a full set of tools (C++ library + R pacakge) for this framework more
- Estimation is done via: MLE, MAP, and MCMC (using an adaptive kernel)
- Posterior probabilities are estimated using the conditional on the observed data.
- To evaluate performance, we used two datasets: manually (fully) annotated (inferred) trees, and experimentally annotated trees

Paper 2: Pooled estimation

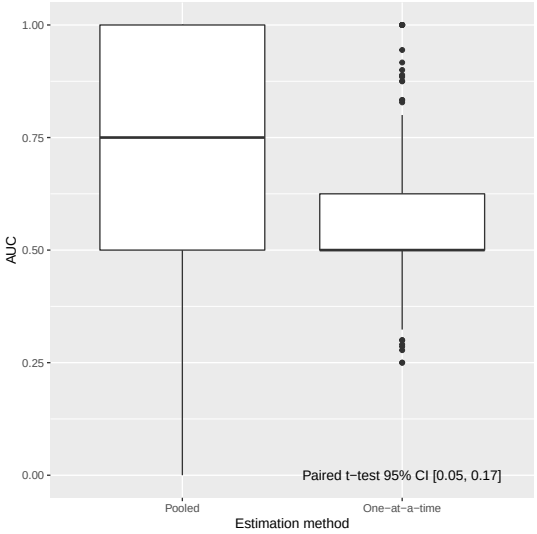
Keck School of Medicine of USC

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ψ_0	0.00	0.00	0.23	0.25	0.00	0.00	0.21	0.25
ψ_1	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.01
μ_{d0}	0.01	0.01	0.97	0.96	1.00	0.01	1.00	0.98
μ_{d1}	0.01	0.02	0.52	0.58	0.25	0.02	0.51	0.58
μ_{s0}	0.00	0.00	0.05	0.06	0.07	0.00	0.05	0.06
μ_{s1}	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.02
π	0.81	0.91	0.78	0.45	0.82	0.91	0.83	0.49
Tree count	88	88	141	141	88	88	141	141
Method	MCMC	MCMC	MCMC	MCMC	MLE	MLE	MLE	MLE
Prior	Uniform	Beta	Uniform	Beta	Uniform	Beta	Uniform	Beta
Inferred	Yes	Yes	No	No	Yes	Yes	No	No
AUC	1.00	1.00	0.69	0.67	0.98	1.00	0.70	0.67
P. Score (obs)	1.00	1.00	0.81	0.81	0.92	1.00	0.81	0.81
P. Score (random)	0.71	0.71	0.61	0.61	0.71	0.71	0.61	0.61

Table 3 Parameter estimates using different estimation methods, priors, and types of annotations.

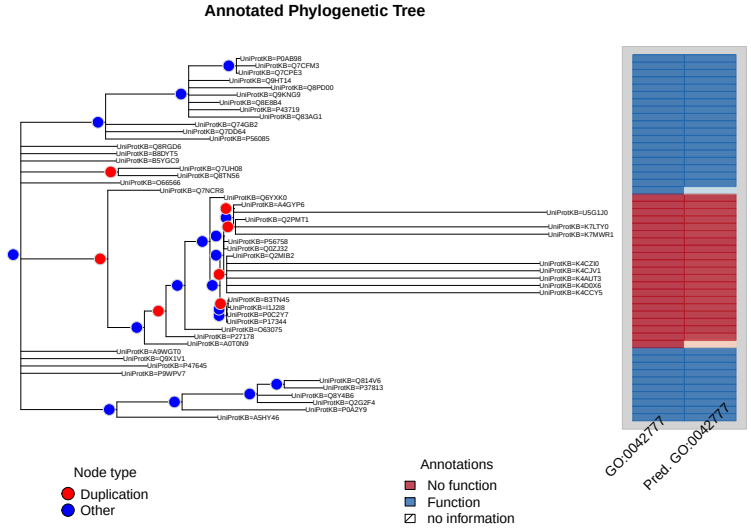
Paper 2: Pooled estimation (worth it?)

Keck School of Medicine of USC



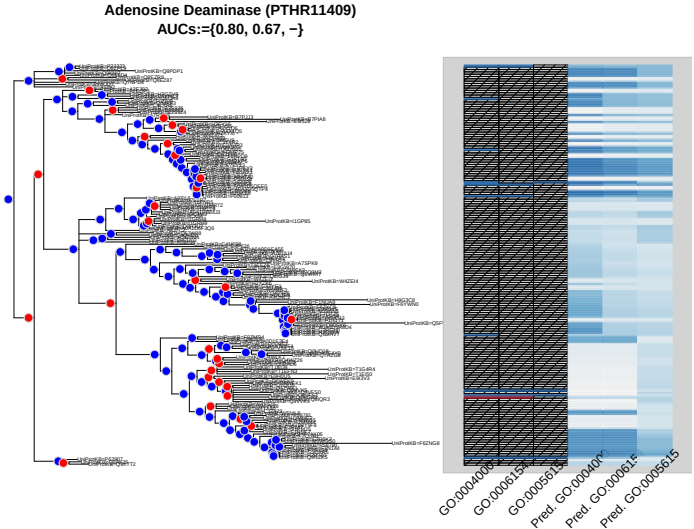
Paper 2: Leave-one-out predictions

Keck School of Medicine of USC



Paper 2: Out-of-sample-predictions

Keck School of Medicine of USC



Paper 2: Key takeaways

Keck School of Medicine of USC

Paper 2: Key takeaways

Keck School of Medicine of USC

- (Yet another) model for predicting gene functions using phylogenetics

Paper 2: Key takeaways

Keck School of Medicine of USC

- (Yet another) model for predicting gene functions using phylogenetics
- Big difference... computationally scalable

Paper 2: Key takeaways

Keck School of Medicine of USC

- (Yet another) model for predicting gene functions using phylogenetics
- Big difference... computationally scalable
- Meaningful biological results

Paper 2: Key takeaways

Keck School of Medicine of USC

- (Yet another) model for predicting gene functions using phylogenetics
- Big difference... computationally scalable
- Meaningful biological results
- Preliminary accuracy results comparable to state-of-the-art phylo-based models

Future directions: ERGMitos

Keck School of Medicine of USC

Possible venues to continue

Future directions: ERGMitos

Keck School of Medicine of USC

Possible venues to continue

- Identify an adequate test for goodness-of-fit assesment

Future directions: ERGMitos

Keck School of Medicine of USC

Possible venues to continue

- Identify an adequate test for goodness-of-fit assesment
- Extend to estimation of large graphs by splitting the networks in induced-subgraphs

Future directions: Gene functional prediction

Keck School of Medicine of USC

Possible venues to continue

Future directions: Gene functional prediction

Keck School of Medicine of USC

Possible venues to continue

- Incorporate more external information using leaf(and node?) level features.

Future directions: Gene functional prediction

Keck School of Medicine of USC

Possible venues to continue

- Incorporate more external information using leaf(and node?) level features.
- Adapt the model to incorporate joint estimation of functions using pseudo-likelihood.

$$P(a, b, c) \approx P(a, b)P(b, c)P(a, c)$$

Future directions: Gene functional prediction

Keck School of Medicine of USC

Possible venues to continue

- Incorporate more external information using leaf(and node?) level features.
- Adapt the model to incorporate joint estimation of functions using pseudo-likelihood.

$$P(a, b, c) \approx P(a, b)P(b, c)P(a, c)$$

- Make the model hierarchical when pooling trees: different mutation rates.

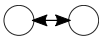
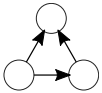
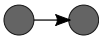
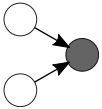
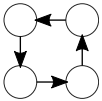
Here are some by-products of my research here at USC

- The slurmR R package
- The pruner C++ library
- The fmcmc R package

References I

Dodd, D. M. B. (1989). Reproductive isolation as a consequence of adaptive divergence in drosophila pseudoobscura. *Evolution*, *43*(6), 1308–1311. Retrieved from http://www.jstor.org/stable/2409365

Sufficient statistics have various forms

Representation	Description
	Mutual Ties (Reciprocity) $\sum_{i \neq j} y_{ij} y_{ji}$
	Transitive Triad (Balance) $\sum_{i \neq j \neq k} y_{ij} y_{jk} y_{ik}$
	Homophily $\sum_{i \neq j} y_{ij} \mathbf{1}(x_i = x_j)$
	Covariate Effect for Incoming Ties $\sum_{i \neq j} y_{ij} x_j$
	Four Cycle $\sum_{i \neq j \neq k \neq l} y_{ij} y_{jk} y_{kl} y_{li}$

go back

ERGMs: The MC-MLE approach

Keck School of Medicine of USC

One of the most popular methods for estimating ERGMs is the MC-MLE approach (citations here)

This consists on the folling steps

1. Start from a sensible guess on what should be the population parameters (usually done using pseudo-MLE esimtation)
2. While the algorithm doesn't converge, do:
   - 2.1 Simulate a stream of networks with the current state of the parameter, θ_t
   - 2.2 Using the law of large numbers, approximate the ratio of likelihoods based on the parameter θ_t, this is the objective function
   - 2.3 Update the parameter by a Newton-Raphson step
   - 2.4 Next iteration

go back

The ergmito

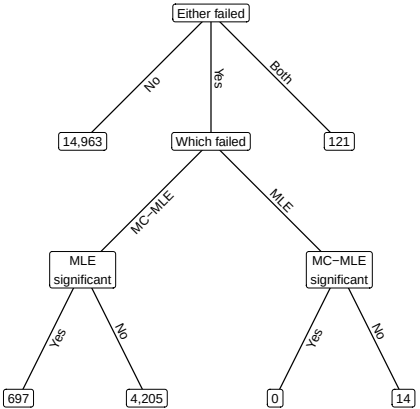
Keck School of Medicine of USC

- Implements estimation of ERGMs using exact statistics for small networks
- Metaprogramming allows specifying likelihood (and gradient) functions for joint models
- Includes tools for simulating, and postestimation checks
- Getting ready for CRAN!

go back

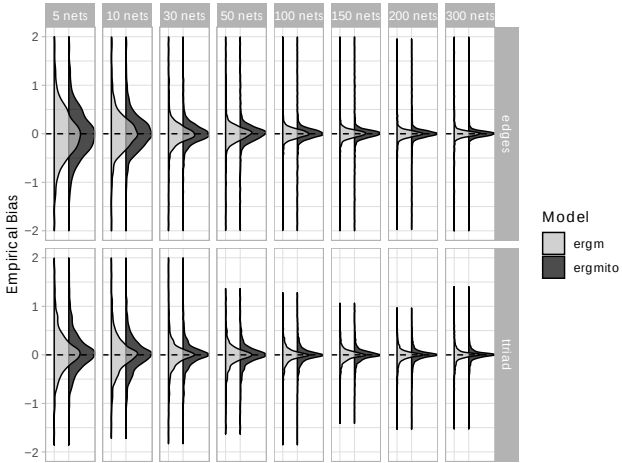
Paper 1 Simulation Studies: Error rate

Keck School of Medicine of USC



go back

Paper 1 Simulation Studies: Empirical Bias



go back

The aphylo

Keck School of Medicine of USC

- TBF

go back