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 9, 2019

Keck School of Medicine of USC

Statistical and computational methods for bioinformatics and social network analysis

- We live in a non-*IID* world.
- Some times, looking the whole helps understanding the parts.
- We have the computational tools to do such.

Keck School of Medicine of USC

Paper 1: Exponential Random Graph Models for Small Networks

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

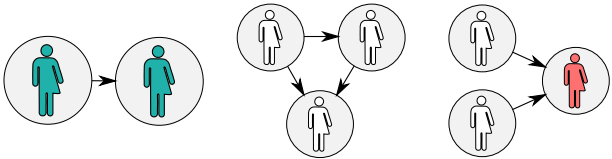
Things that are very interesting but I most probably won't have any time to discuss with the attendees

Keck School of Medicine of USC

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



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

The normalizing constant has $2^{n(n-1)}$ terms!

more on terms

Keck School of Medicine of USC

Medium-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 small networks?

Do we care about small networks?

Keck School of Medicine of USC

We see small networks everywhere

- Families and friends
- Small teams
- Egocentric networks
- Online networks (sometimes)
- etc.

Keck School of Medicine of USC

From the methodological point of view, current methods are great, but:

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

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

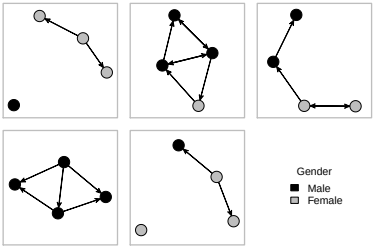


Figure 1 Random sample of 5 networks simulated using the ergmito package

	Edgecount	Full model
Edgecount	-0.69^{*}	-1.70^{**}
	(0.27)	(0.54)
Homophily (on Gender)		1.59^{*}
		(0.64)
AIC	78.38	73.34
BIC	80.48	77.53
Log Likelihood	-38.19	-34.67
Num. networks	5	5

Standard errors in parenthesis. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1 Fitted ERGMitos using the fivenets dataset.

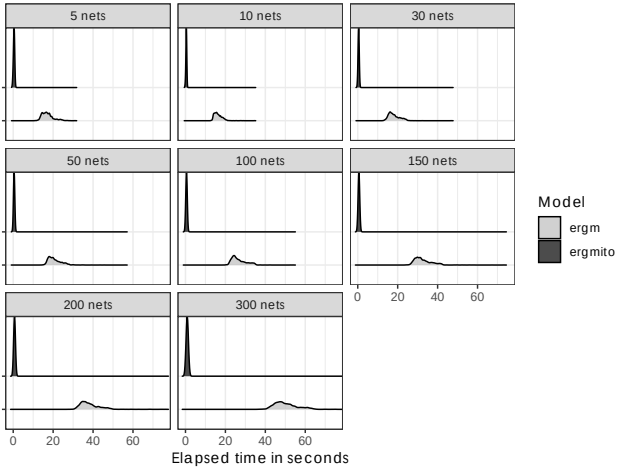
We performed a large simulation study ▸ more comparing MC-MLE (ergm) with MLE (ergmito).

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 2 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

Keck School of Medicine of USC

Key takeaways

- 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.

Next steps

- Revisit measurment of goodness-of-fit.
- Explore extending this method for (very) large networks.

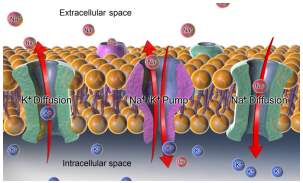
Keck School of Medicine of USC

Keck School of Medicine of USC

How we organize the information about genes (According to the Gene Ontology)

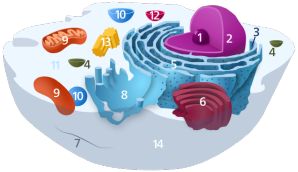
Molecular function

Active transport GO:0005215



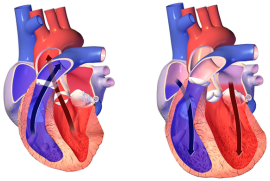
Cellular component

Mitochondrian GO:0004016



Biological process

Heart contraction GO:0060047



Keck School of Medicine of USC

- Currently, the Gene Ontology Project has: 44,945 validated terms, $\sim$ 6,400,000 annotations on $\sim$ 1,150,000 species.
- Of all annotations, about $\sim$ 500,000 are on human genes.
- Knowledge about gene functions can accelerate bio-medical research.

Keck School of Medicine of USC

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 3 Heart Contraction Function. source: amigo.geneontology.org

You know what is interesting about this function?

These four species have a gene with that function... and two of these are part of the same evolutionary tree!

pthr10037

pthr11521

pthr11521

pthr24356

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.
- A single phylogenetic tree can host multiple species

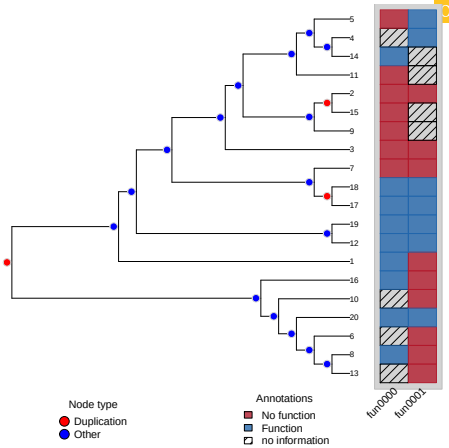


Figure 2 Random annotated phylogenetic tree.

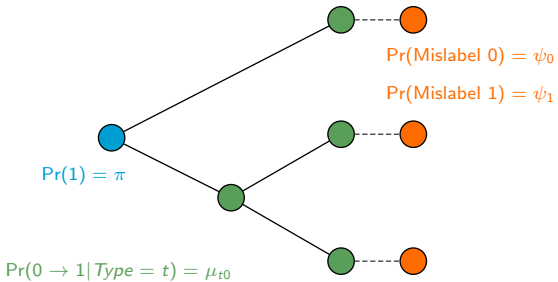
We can use **the evolutionary tree**

to infer presence/absence of

gene functions (annotations)!

Keck School of Medicine of USC

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



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

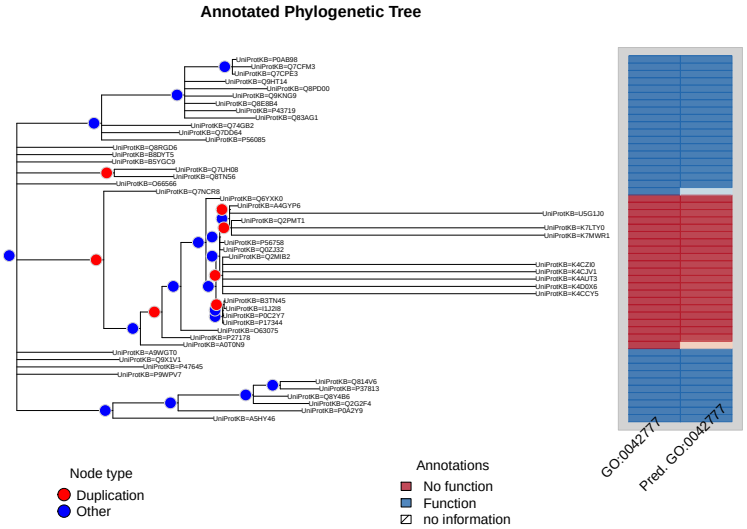
Keck School of Medicine of USC

	(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 4 Parameter estimates using different estimation methods, priors, and types of annotations.

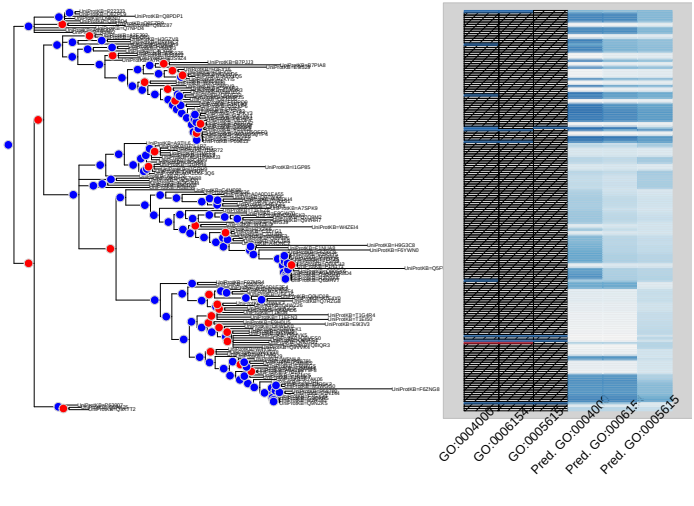
Prediction with real data: Leave-one-out

Keck School of Medicine of USC



Keck School of Medicine of USC

Adenosine Deaminase (PTHR11409)
AUCs:={0.80, 0.67, −}



Keck School of Medicine of USC

Key takeaways

- (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.

Next steps

- 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.

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 9, 2019

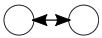
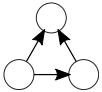
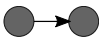
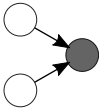
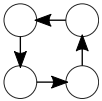
Thanks!

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

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

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

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

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

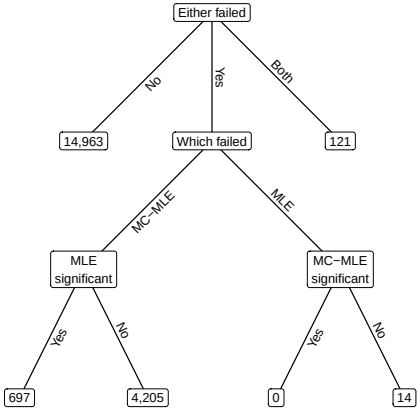
Keck School of Medicine of USC

We performed a simulation study with the following features:

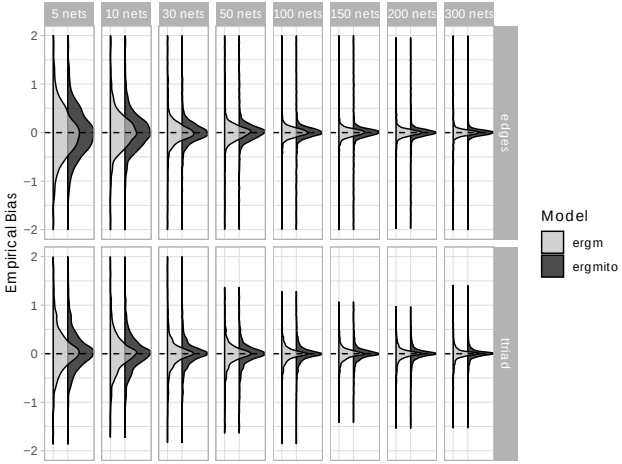
- Draw 20,000 samples of groups of small networks
- Each group had prescribed: (model parameters, number of networks, sizes of the networks)
- Each group could have from 5 to 300 small networks
- We estimated the models using MC-MLE and MLE.

go back

Keck School of Medicine of USC



go back


5 nets
10 nets
30 nets
50 nets
100 nets
150 nets
200 nets
300 nets
edges
ttriad
Empirical Bias
2
1
0
−1
−2
Model
ergm
ergmito

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

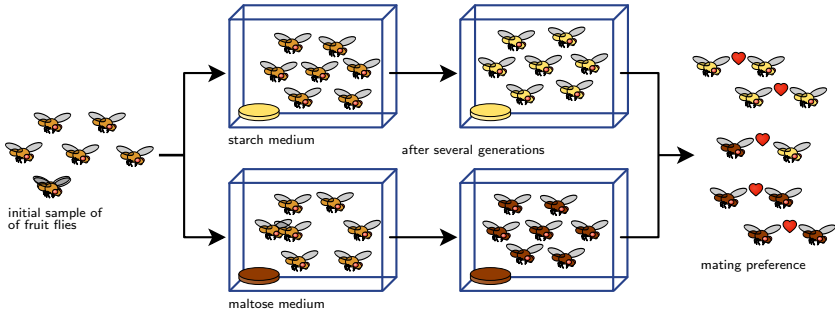


Figure 3 11989DoddDodd (): After one year of isolation, flies showed a significant level or assortativity in mating (wikimedia)

go back

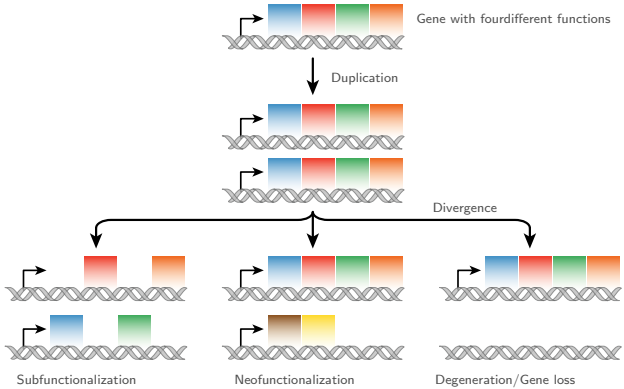


Figure 4 A key part of molecular innovation, gene duplication provides opportunity for new functions to emerge (wikimedia)

Keck School of Medicine of USC

- Pruning algorithm implemeted in C++ using the `pruner` template library (implemeted in this project).
- The estimation is done using either Maximum Likelihood, Maximum A Posteriory, or MCMC.
- The MCMC estimation is done via the `fmcmc` R package using adaptive MCMC (also implemeted as part of this project)

go back