Mechanistic Models in Comparative Genomics

David A. Liberles

University of Wyoming

From the Beginning…

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Message from ISCB

The Roots of Bioinformatics in ISMB

Todd A. Gibson*

Computer Science Department, California State University, Chico, California, United States of America

"When I first began this, there was a very common response, especially among senior biologists, that: "computational biology is just a faster way to do theoretical biology, and we all know that theoretical biology doesn't work. And so computational biology is just a way to do something that doesn't work even faster.""

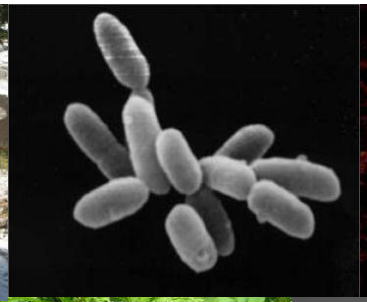
"The biologists now accept the need for computation, but I think they tend to think of the people who do this, the computer scientists, the engineers, mathematicians, as people who are very useful for producing tools that the biologists can use.
And the computer scientists, engineers, etc., sometimes are quite naive about the complexity of biologic problems. "

Building an interdisciplinary bridge from biophysical chemistry to evolutionary biology for the functional analysis of comparative genomic data

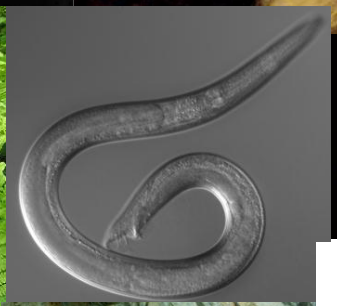
- TAED: A comparative genomic study of chordates
- **Moving from informatics to theory rooted in biochemistry and evolutionary biology in bioinformatics**
  - What is the right level of mechanism for biological inference?
  - Evolutionary/Functional models for the retention of gene duplicates
  - A population genetic model for inter-specific amino acid substitution patterns

Explaining the Functional Genomic Basis of Biodiversity

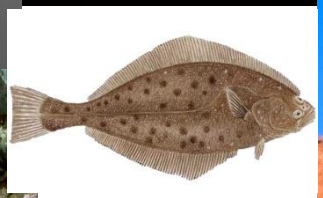






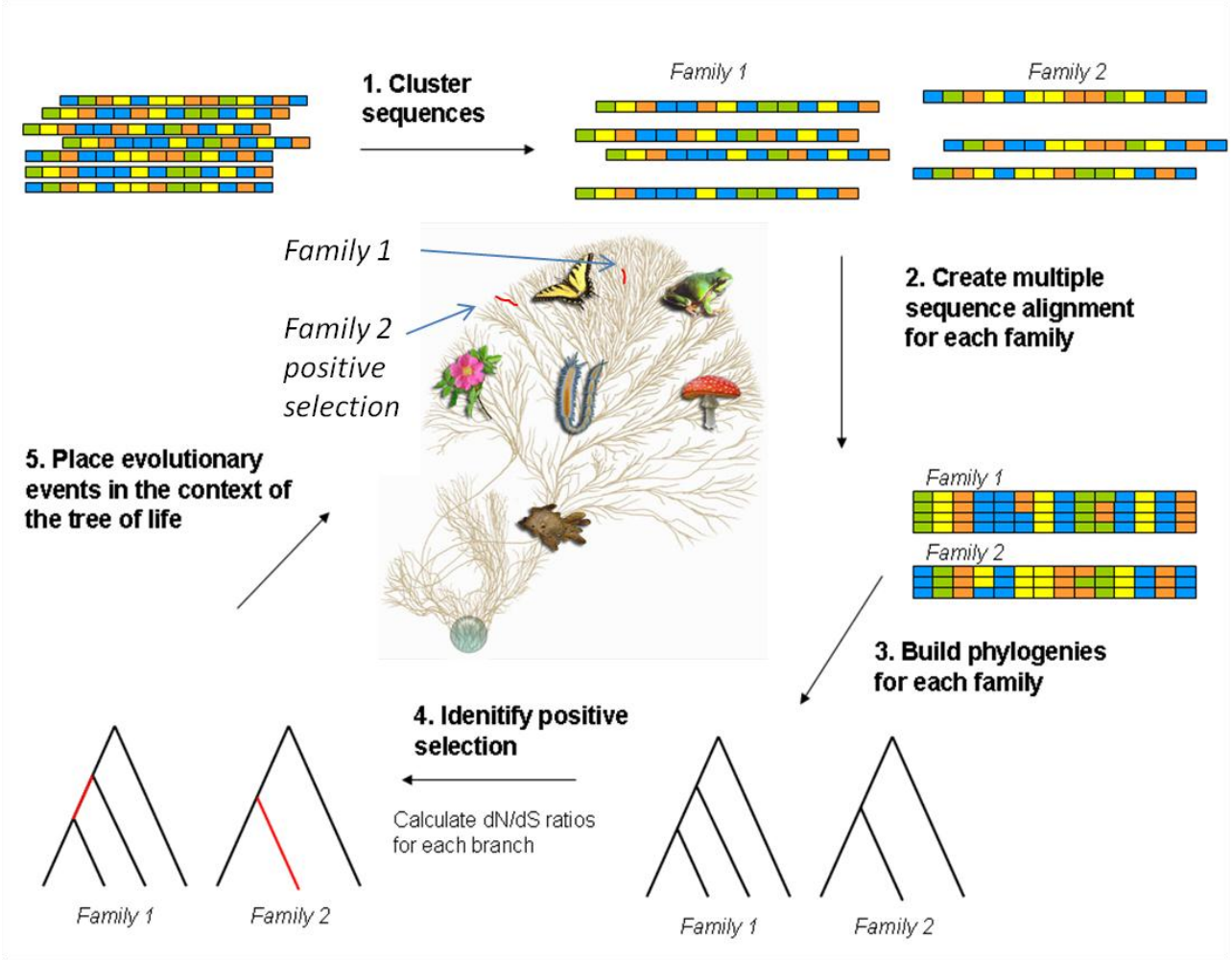




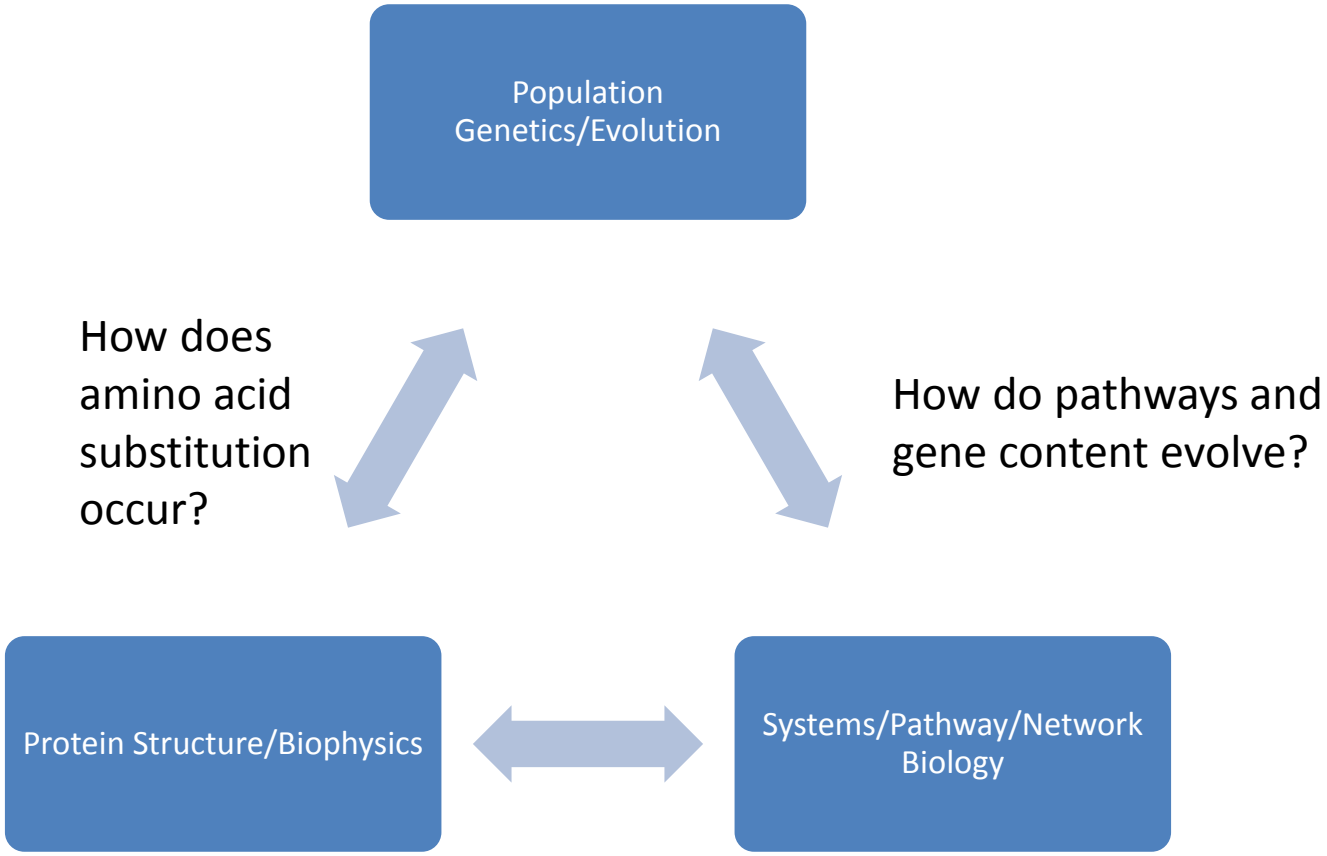





The Adaptive Evolution Database Pipeline



New Models For Comparative Genomics

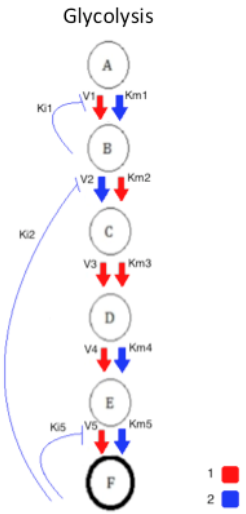


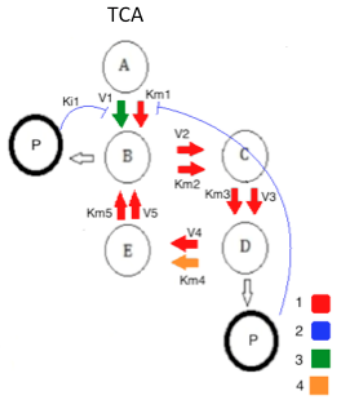
Some additional examples of projects in the lab (I)

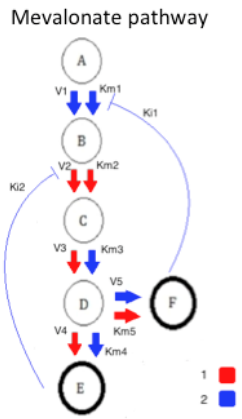
- Given a mutation in a protein, what is its probability of fixation
  - When a protein must fold into a stable structure to properly orient key residues
    - How to account for alternative conformations that a protein might adopt upon mutation?
  - Bind specific other proteins
  - Not bind specific other proteins
  - What other selective constraints govern a protein that we are mis-specifying?
  - Models and methods for simulation and for inference over a phylogeny

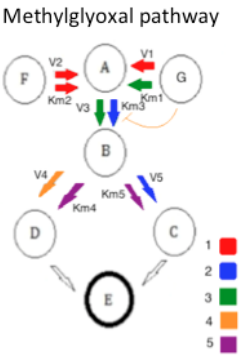
Some additional examples of projects in the lab (II)

- How do metabolic pathways evolve with selective constraints for:
  - Flux
  - Against wasteful mRNA and protein synthesis
  - Against the production of deleterious intermediates
  - With duplication and the emergence of promiscuous activities (according to the patchwork and retrograde models)
- What is the role of mutation-selection balance? And are there/why are there rate limiting steps?
- More practically, can we differentiate between inter-molecular (functional) compensatory covariation and functional shifts?



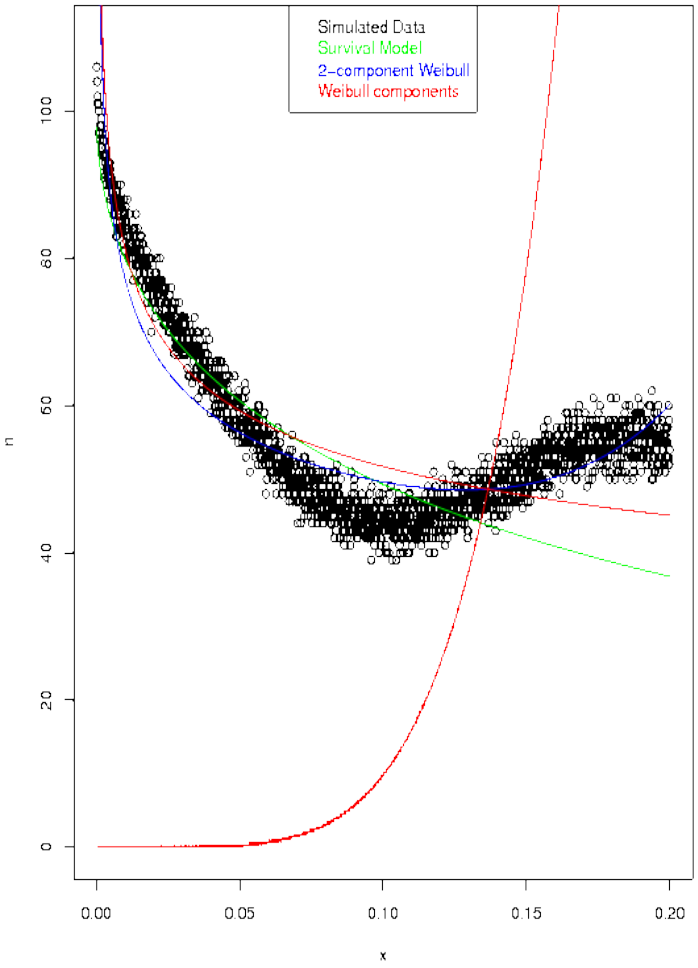






Some Thoughts From a Recent Review With Liang Liu and Tanja Stadler

- Model identification
  - Is there a natural bias when comparing phenomenological models vs. constrained mechanistic models in terms of likelihood vs. # parameters?
- Model validation:
  - Statistical identifiability vs. Mechanistic identifiability
  - Describing a process vs. fitting the data



And now for a focus on gene duplication…

Understanding how duplicate genes contribute to changing genome function

Types of Gene Duplication

- Whole genome duplication
  - duplicates identical
- Other large scale duplication (eg whole chromosome)
  - duplicates identical
- Tandem duplication (through replication or recombination)
  - coding sequences likely identical, may be missing expression elements in some cases
- Transposition
  - coding sequences may be identical, expression elements likely different
- Retrotransposition
  - coding sequence identical, but without introns, expression elements likely different

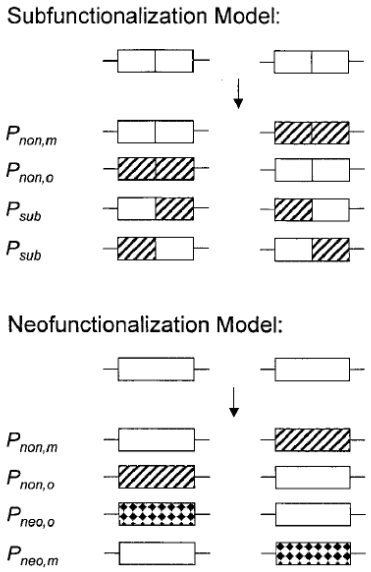
What matters in duplicate gene retention

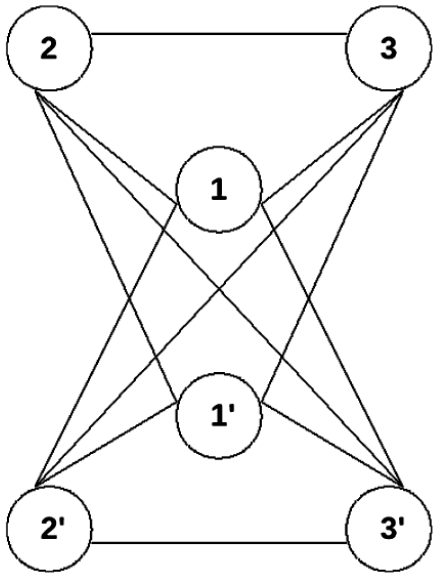
- Gene expression (timing, localization, level)
- Coding sequence function (e.g. intermolecular interactions)

- Changes in these governed by mutations of different types in different locations within a gene (upstream, coding sequence, splice site, …)
- Population genetic processes acting upon the mutation

Mechanisms of Duplicate Gene Retention

- Evolutionary Processes Considered
  - Nonfunctionalization
  - Neofunctionalization
  - Subfunctionalization
  - Dosage balance (stoichiometry-driven)

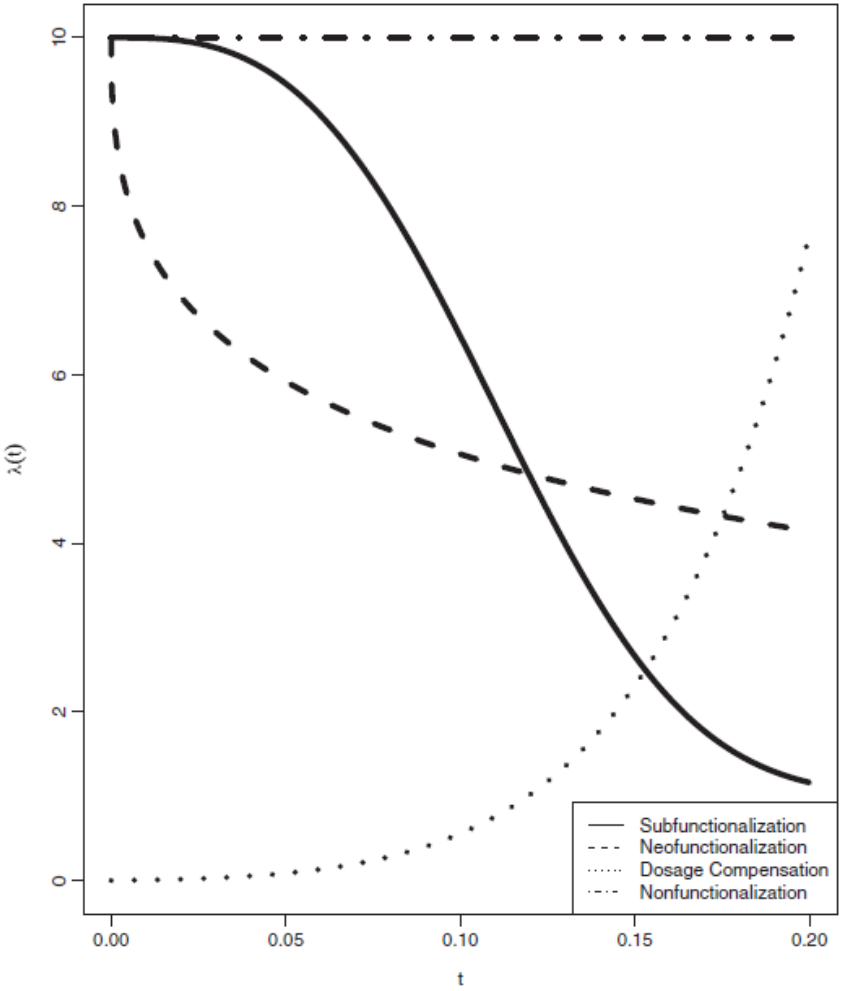


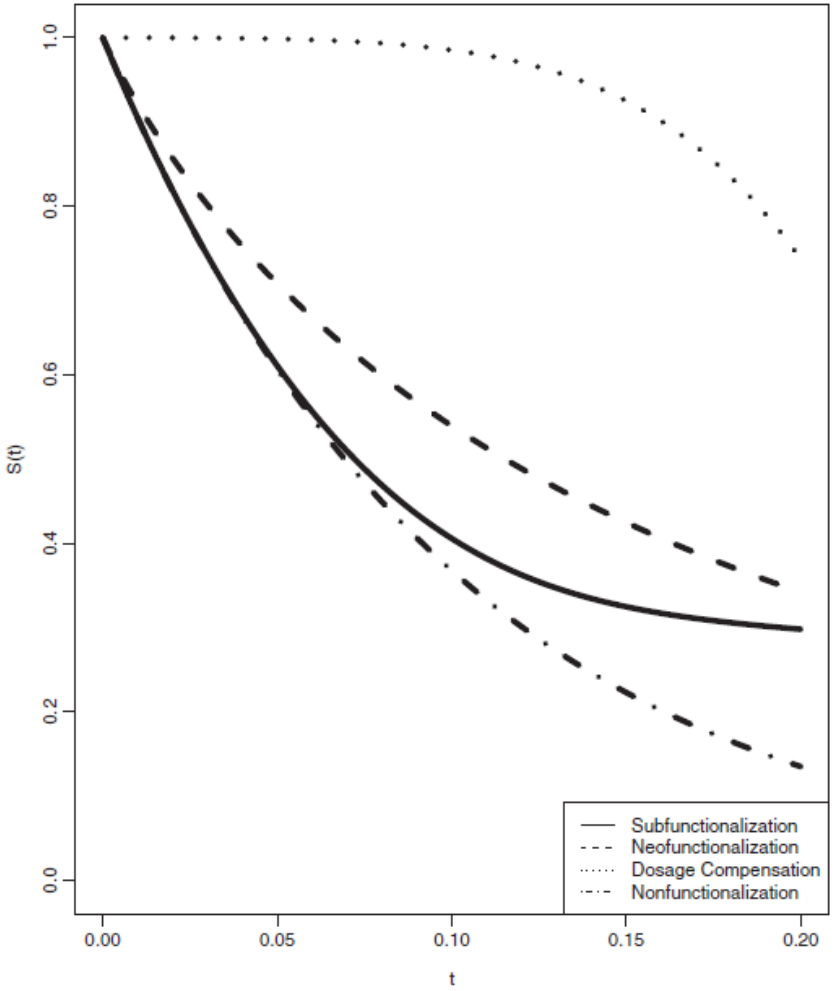


- Goal: Develop models to differentiate between duplicate gene fates
  - Intra-genomic analysis (dS plots)
  - Gene tree /Species Tree Reconciliation

(Figures from Lynch et al., 2001 and Konrad et al., 2011)

Theoretical Hazard and Survival Functions



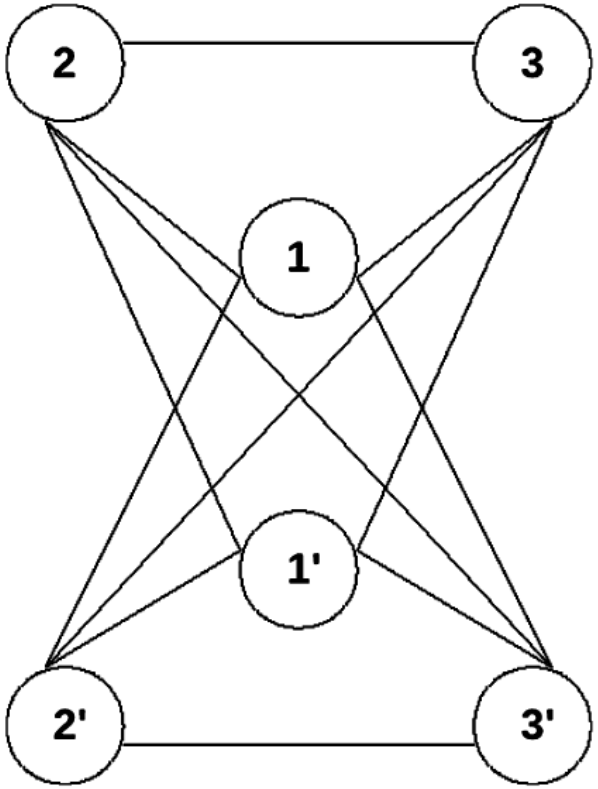


A General Death Model

- Hazard: $\lambda(t) = ge^{-bt^c} + d$
- Survival: $S(t) = N_0 e^{-dt - g\sum_{n=0}^{\infty} \frac{(-b)^n t^{cn+1}}{cn(n!)+n!}}$

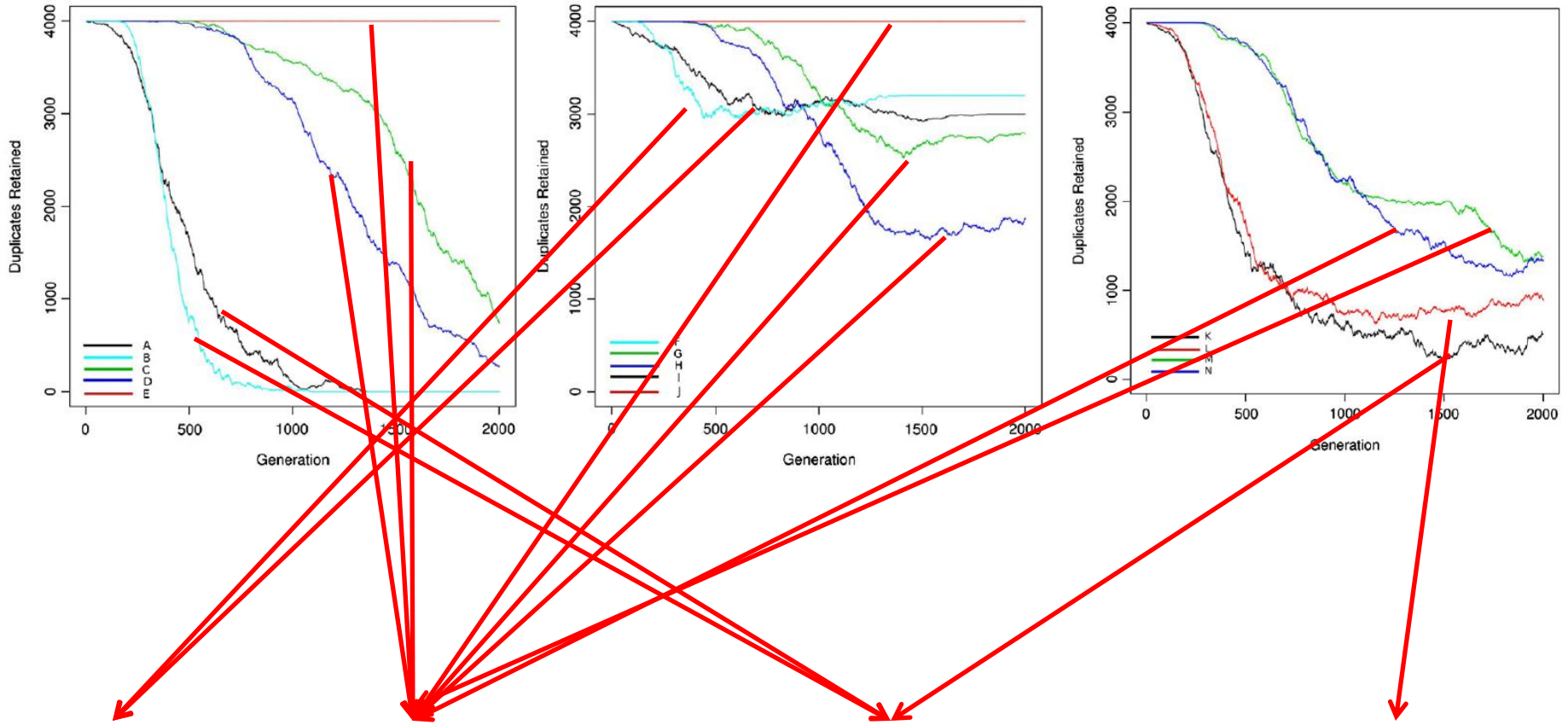
- For all, g > 0
- Non: g = 0, d> 0 (d>10)
- Neo: b > 0, 0 < c <1, d > 0, g>0
- Sub: b > 0, c > 1, d > 0, g>0
- Dos: b < 0, 0 < c < 1, d = -g, ($\lambda(t)_{0.02}$<0.1)

A simulation scheme for gene duplication



Simulation run with and without subfunctionalization allowed (regulatory network vs. protein complex) with probabilities of gene loss and link loss in a population genetic framework.

Simulated Data for Model Comparison



Subfunction. Dosage Balance Nonfunction. Neofunction.

Ongoing work…

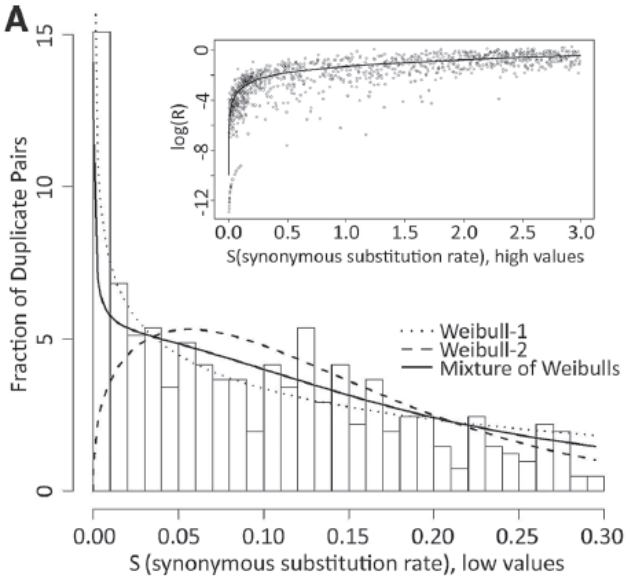
- Hybrid process parameterization (dosage+neo; dosage+sub)
- Models for larger scale duplication, duplication rate variation
- Evaluation of assumptions about population genetics
- Use of the birth-death model and migration to gene tree/species tree reconciliation in a Bayesian framework
- Plus simulation of data under more complex genetic and population genetic regimes

What happens in real genomes?

- This is a figure from a 2010 paper involving a model that is not ours. There has been critique of our models and modeling, but everyone comes to the same conclusion that comes with our models, that there is support in all genomes analyzed for a declining hazard function consistent with neofunctionalization according to the framework presented.

Plasticity of Animal Genome Architecture Unmasked by Rapid Evolution of a Pelagic Tunicate

France Denoeud,[1,2,3] Simon Henriet,[4*] Sutada Mungpakdee,[4*] Jean-Marc Aury,[1,2,3*] Corinne Da Silva,[1,2,3*] Henner Brinkmann,[5] Jana Mikhaleva,[4] Lisbeth Charlotte Olsen,[4] Claire Jubin,[1,2,3] Cristian Cañestro,[6,24] Jean-Marie Bouquet,[4] Gemma Danks,[4,7] Julie Poulain,[1,2,3] Coen Campsteijn,[4] Marcin Adamski,[4] Ismael Cross,[8] Fekadu Yadetie,[4] Matthieu Muffato,[9] Alexandra Louis,[9] Stephen Butcher,[10] Georgia Tsagkogeorga,[11] Anke Konrad,[22] Sarabdeep Singh,[12] Marit Flo Jensen,[4] Evelyne Huynh Cong,[4] Helen Eikeseth-Otteraa,[4] Benjamin Noel,[1,2,3] Véronique Anthouard,[1,2,3] Betina M. Porcel,[1,2,3] Rym Kachouri-Lafond,[13] Atsuo Nishino,[14] Matteo Ugolini,[4] Pascal Chourrout,[15] Hiroki Nishida,[14] Rein Aasland,[16] Snehalata Huzurbazar,[12] Eric Westhof,[13] Frédéric Delsuc,[11] Hans Lehrach,[17] Richard Reinhardt,[17] Jean Weissenbach,[1,2,3] Scott W. Roy,[18] François Artiguenave,[1,2,3] John H. Postlethwait,[6] J. Robert Manak,[10] Eric M. Thompson,[4,19] Olivier Jaillon,[1,2,3] Louis Du Pasquier,[20] Pierre Boudinot,[21] David A. Liberles,[22] Jean-Nicolas Volff,[23] Hervé Philippe,[5] Boris Lenhard,[4,7,19] Hugues Roest Crollius,[9] Patrick Wincker,[1,2,3]† Daniel Chourrout[4]†



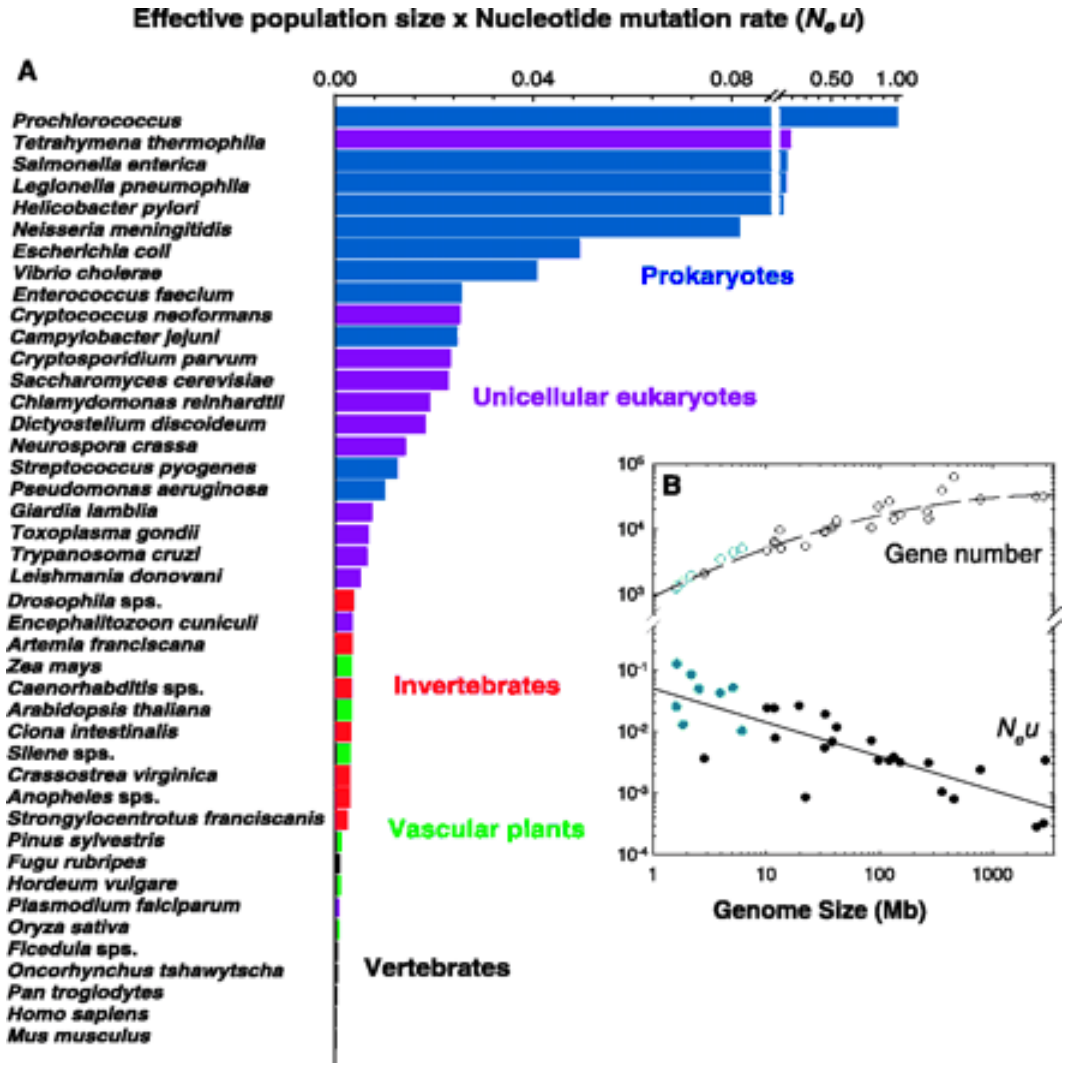
- Further controls are needed to validate the biological conclusion of widespread neofunctionalization.

How do homologous protein-coding genes diverge?...

About the interplay between thermodynamics and population size….

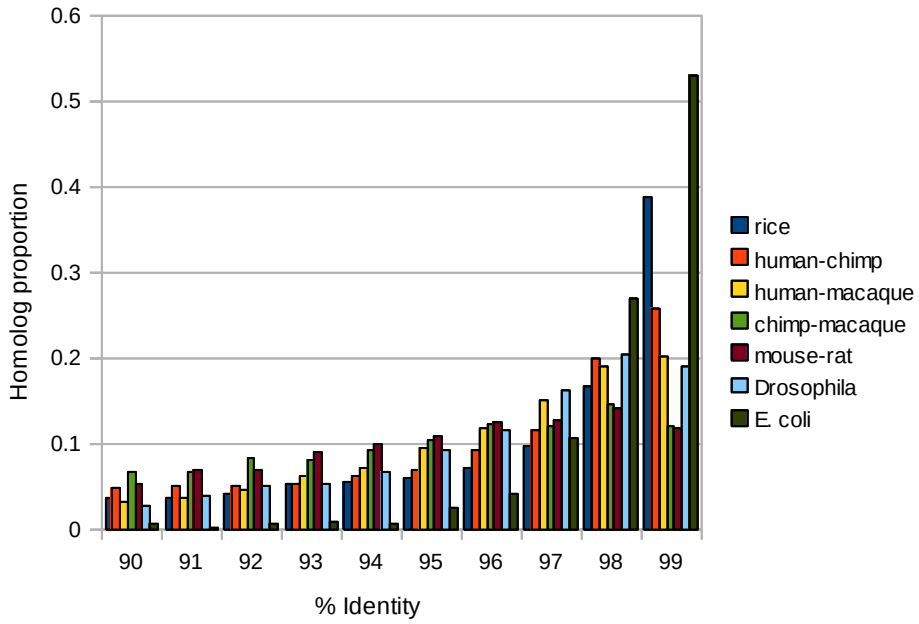
- Contrary to some thought in the protein structure community, one does not necessarily expect the thermodynamics of protein structure to be the only signal in amino acid substitution data
- Population genetic theory predicts that the strength of selection (thermodynamic constraint) on a protein sequence will be guided by the effective population size. The larger the effective population size, the more power to select and the less random observed changes are expected to be….
- Does effective population size modulate the relative probabilities of amino acid substitution?
- And can we build a model with Ne and s for amino acids that is useful in characterizing lineage-specific change?

Some organismal effective population sizes…



Lynch and Conery, Science 302:1401-1404.

Generating Genome-Specific PAM Matrices



Identifying genome pairs across effective population size ranges with similar orthologous sequence similarity profiles (>97% amino acid identity)

Building a Model for Probabilities of Amino Acid Transitions

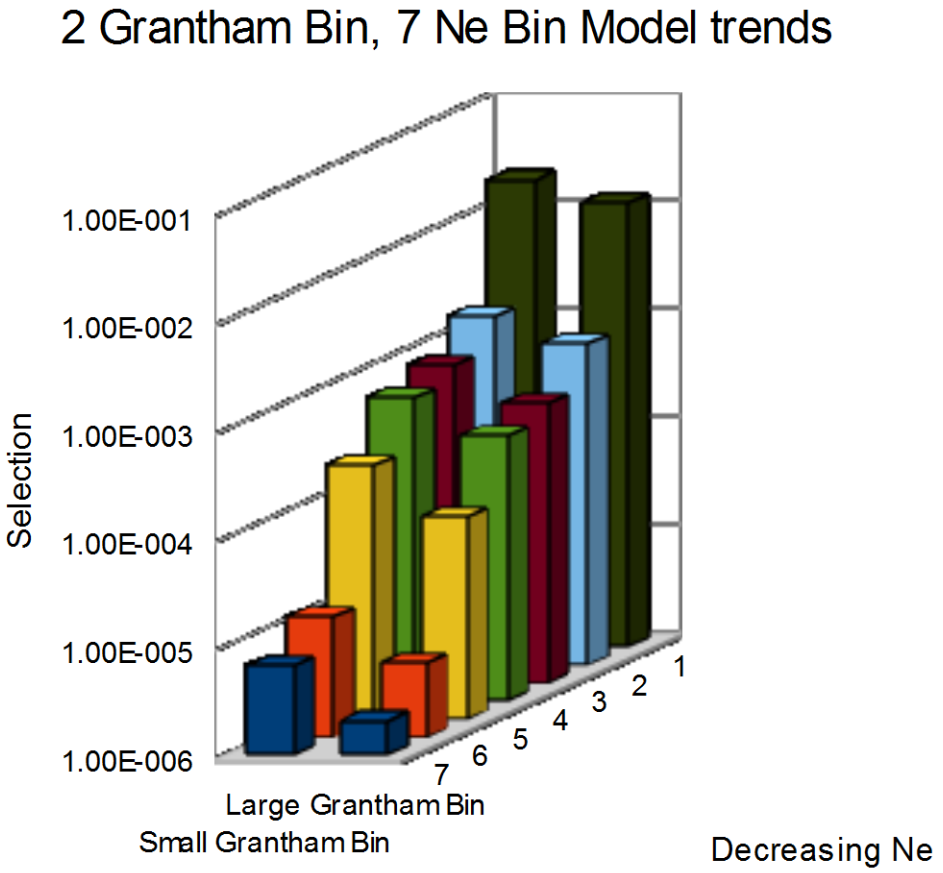
- Kimura Fixation Probabilities for Amino Acids, relating strength of selection and effective population size to probability of fixation:

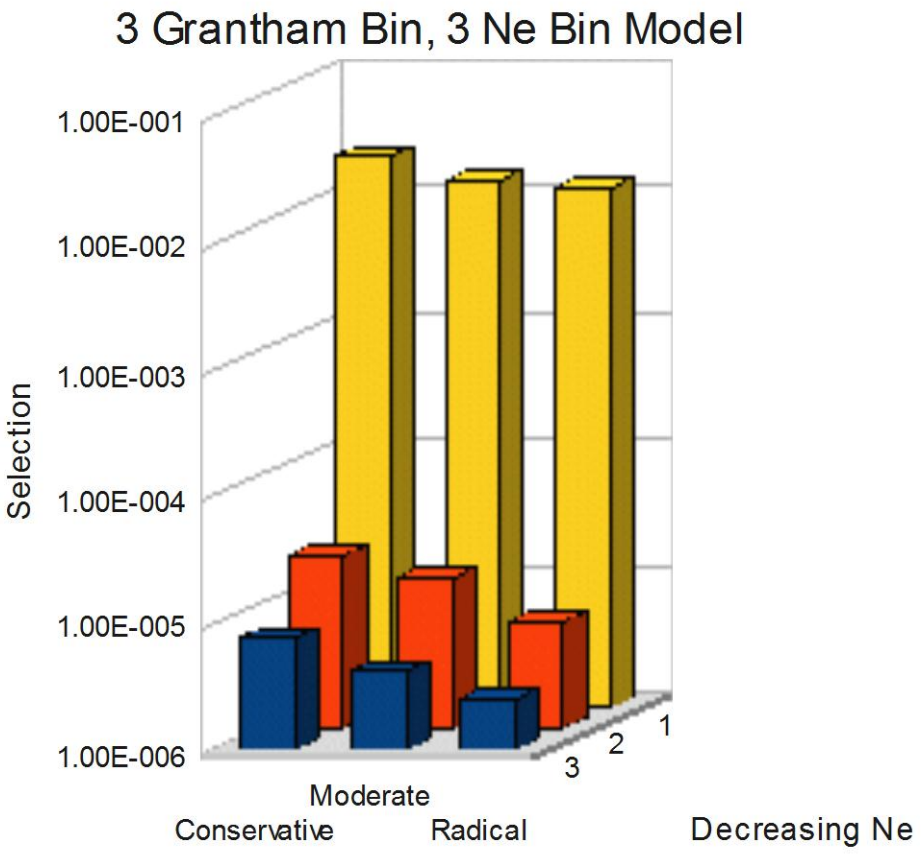
$$F = (1- e^{-2S}) / (1- e^{-4 Ne S})$$

- When different amino acid transitions are considered separately, the differential probabilities of transition between amino acids dictated by the genetic code must be considered as part of the mutational opportunity, as shown on the next slide.
- Some assumptions:
  - Each amino acid position segregates independently
  - Fixed, constant population size separating species
  - Changes observed are fixed rather than segregating
  - Transitions in a Grantham Matrix category are under similar selective pressures
  - Constant, equal equilibrium frequencies of amino acids
- Extending the model:

$$RP_{i=} \frac{\mu_i \frac{1 - e^{-2s_i}}{1 - e^{-2Ns_i}}}{\sum_j \mu_j \frac{1 - e^{-2s_j}}{1 - e^{-2Ns_j}}}$$

Trends of Measured Selection





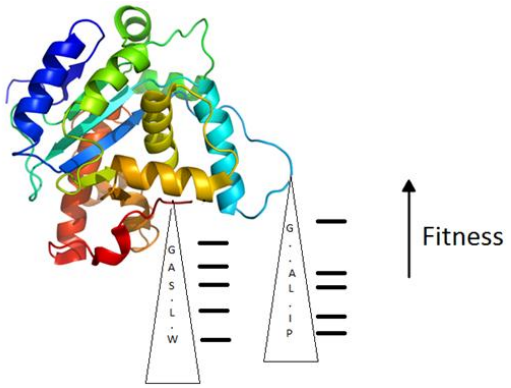
- **Models with more Ne bins, fewer Grantham bins show support**
- **Selection coefficient decreases with Ne**
- **Selection coefficient decreases with Grantham value**

Patterns of Selection

- Decreasing selection with increasing Grantham
  - Are radical and conservative changes equally solvent exposed?
- Support for multiple bins of Ne
  - Is Ne mis-specified?
- Decreasing selection with increasing population size at constant Grantham
  - Mis-specification of π?
  - Nevo et al. (1997) suggests that the interplay between linkage and population size can explain much more diversity and substitution in small effective population size organisms than is expected by the type of modeling done here
  - In larger populations, there will be more segregating variation that averages together with the fixed changes and is more likely to be slightly deleterious
  - Something else? (e.g. Goldstein (2013)?)

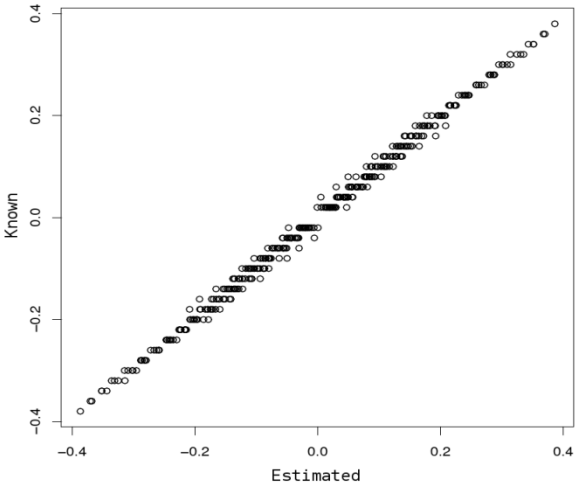
Further And Future Considerations

- Linkage (Hill-Robertson Effects)
  - Selective sweeps
  - Background selection
- Ne as a free parameter
- Accounting for the expectation of segregating variation based upon Ne
- Accounting for protein fold and position solvent accessible surface area
- A structure-based biophysical model (we have one, not presented today)

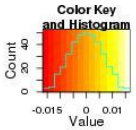


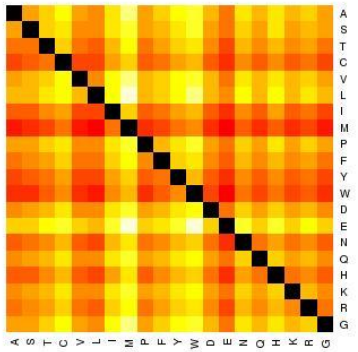
Establishing the identifiability and behavior of extended models

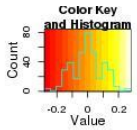
$$RP_{i=} \frac{\pi_i \mu_i \frac{1 - e^{-2s_i}}{1 - e^{-2Ns_i}}}{\sum_j \pi_j \mu_j \frac{1 - e^{-2s_j}}{1 - e^{-2Ns_j}}}$$

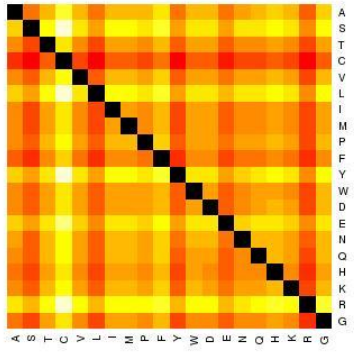


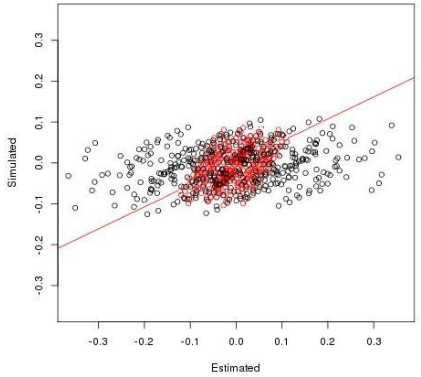
$S_{tot} = S_{AL} + S_{DE} + S_{YF}$





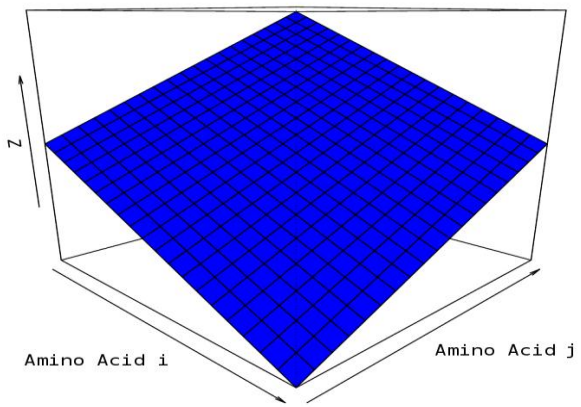


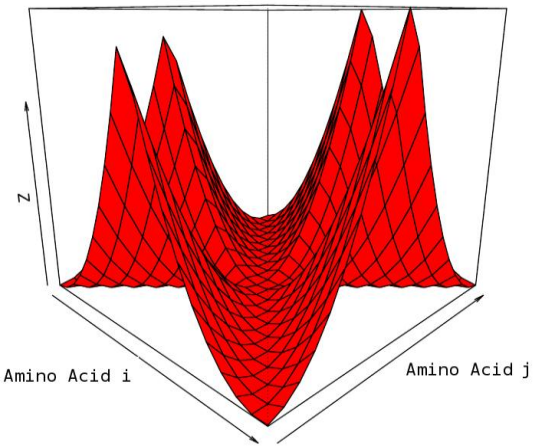


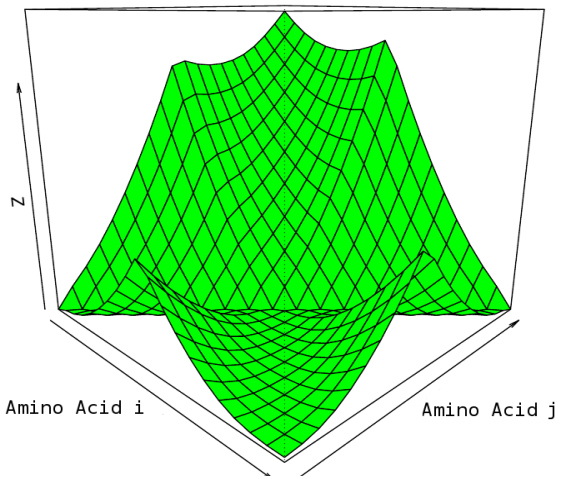


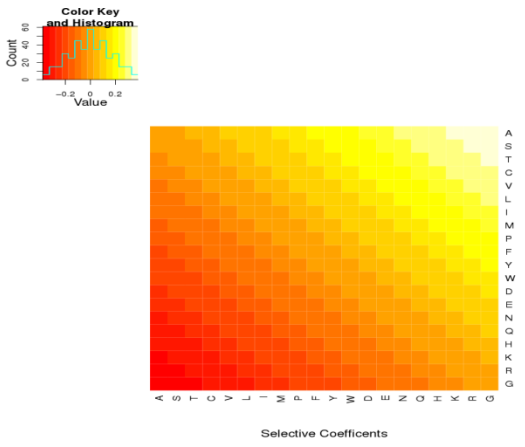
Preliminary data, Ashley Teufel

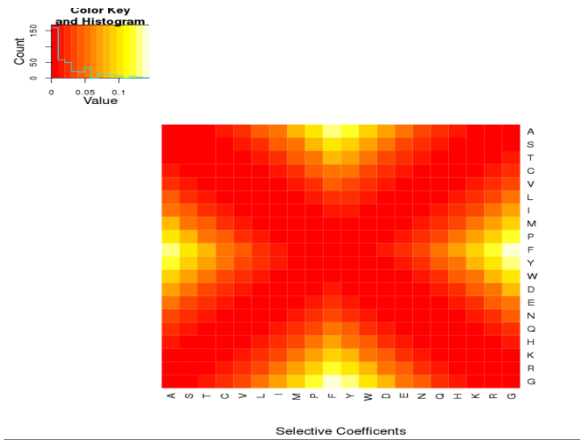
A mixture of site-specific processes

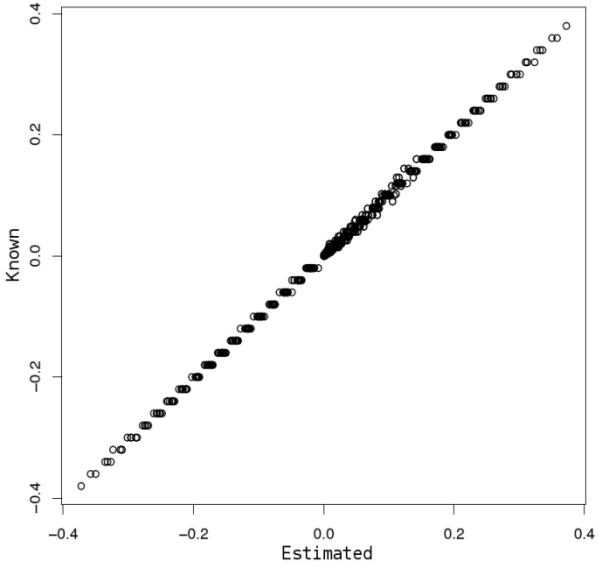










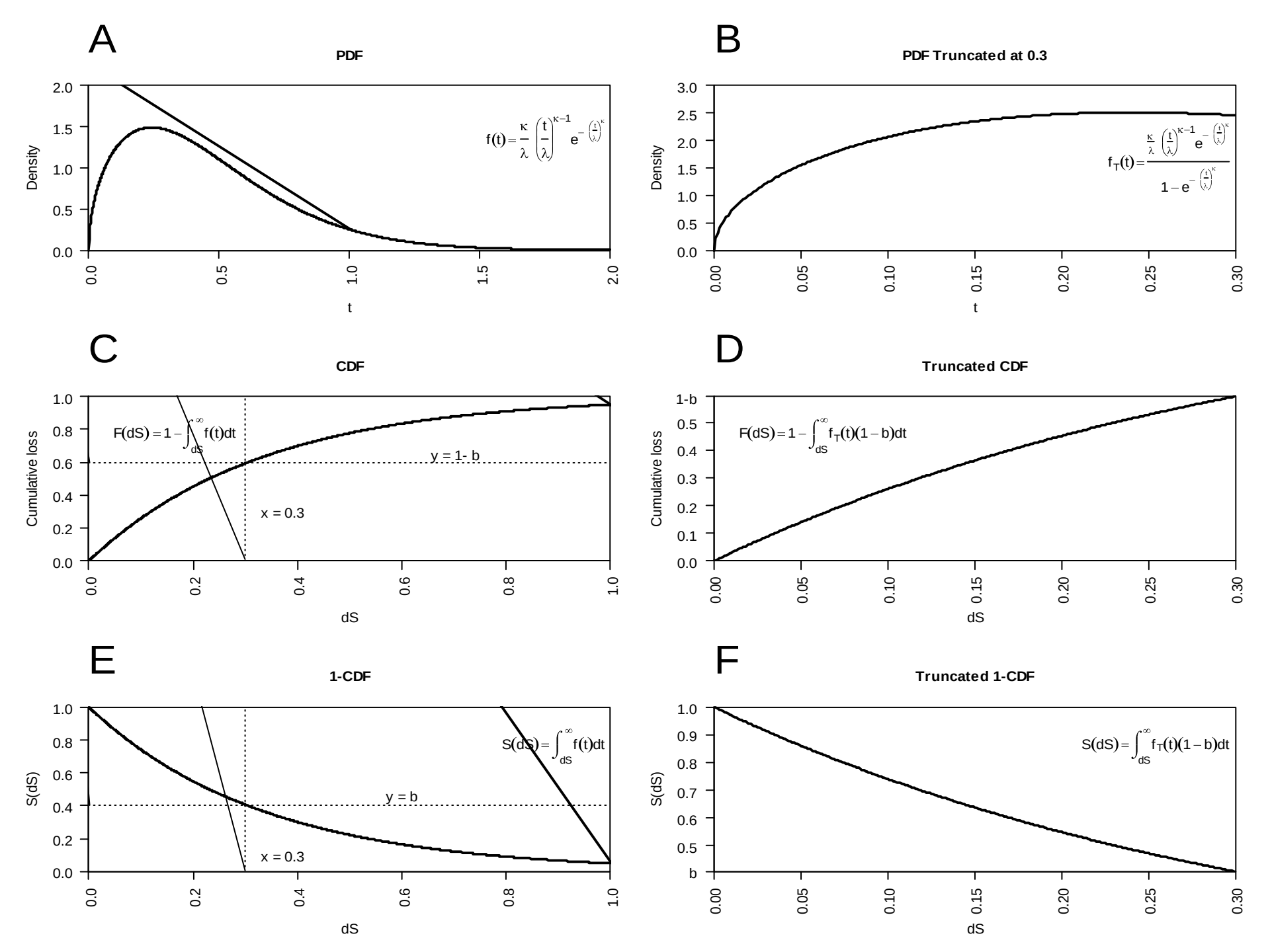


Group Members and Funding

Key Collaborator on This Work:
Liang Liu (U. Georgia Statistics)

Current Lab Members:
Russell Hermansen- Ph.D. student
Dohyup Kim- Ph.D. student
Anke Konrad- Ph.D. student
Jason Lai- Ph.D. student
Alena Orlenko- Ph.D. student
Juan Felipe Ortiz- Ph.D. student
Ashley Teufel- Ph.D. student

Funding:
NSF (DBI and DMS)
NIH (MSFD R21)
NIH-INBRE


A
PDF
Density
t
$f(t)=\frac{\kappa}{\lambda}\left(\frac{t}{\lambda}\right)^{\kappa-1}e^{-\left(\frac{t}{\lambda}\right)^{\kappa}}$
B
PDF Truncated at 0.3
Density
t
$f_T(t)=\frac{\frac{\kappa}{\lambda}\left(\frac{t}{\lambda}\right)^{\kappa-1}e^{-\left(\frac{t}{\lambda}\right)^{\kappa}}}{1-e^{-\left(\frac{t}{\lambda}\right)^{\kappa}}}$
C
CDF
Cumulative loss
dS
$F(dS)=1-\int_{dS}^{\infty}f(t)dt$
y = 1- b
x = 0.3
D
Truncated CDF
Cumulative loss
dS
$F(dS)=1-\int_{dS}^{\infty}f_T(t)(1-b)dt$
E
1-CDF
S(dS)
dS
$S(dS)=\int_{dS}^{\infty}f(t)dt$
y = b
x = 0.3
F
Truncated 1-CDF
S(dS)
dS
$S(dS)=\int_{dS}^{\infty}f_T(t)(1-b)dt$