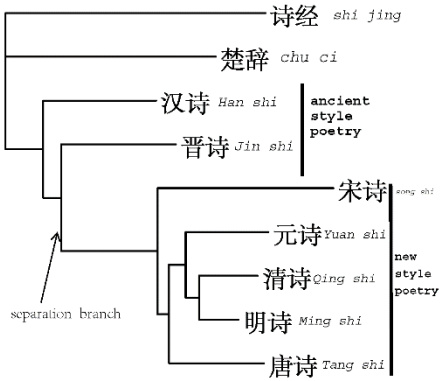
A probabilistic model for gene family evolution

Liang Liu

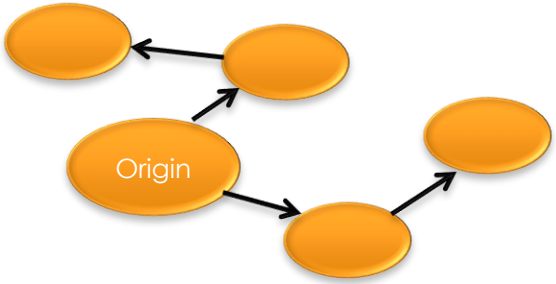
Department of Statistics
Institute of Bioinformatics
University of Georgia

Evolution

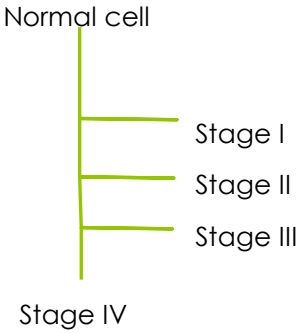
- Biology
- Evolutionary linguistics
- Cancer
- Infectious disease



The evolution of classical Chinese poetry



Pathogens evolve along different paths



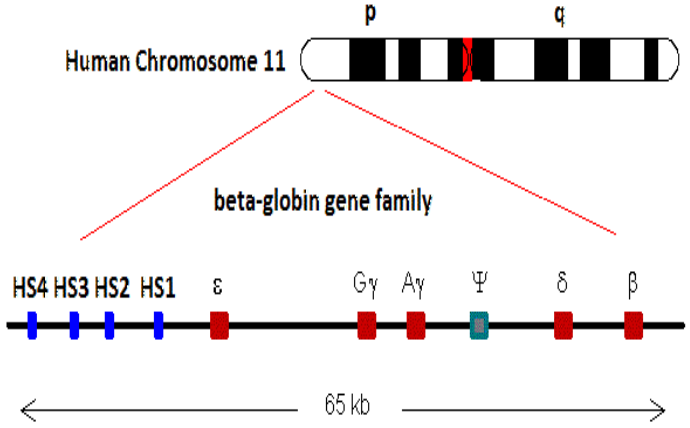
Cancer evolution

Outline

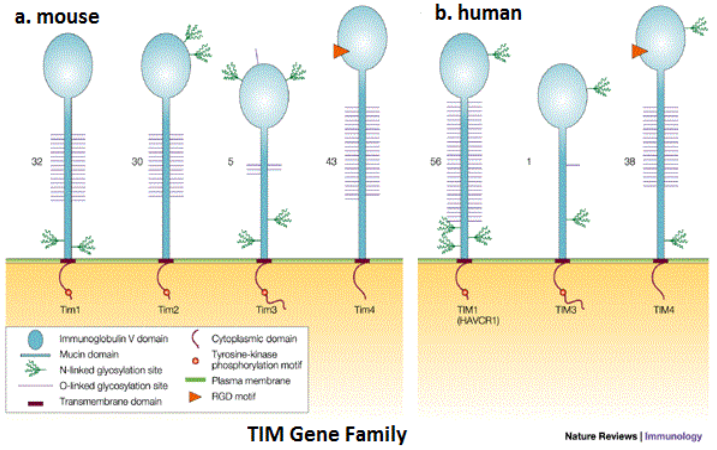
- Biological process
- Stochastic process
- Likelihood function
- Statistical inference

A gene family

- A gene family is a group of genes that share important characteristics (sequence, structure, or function)



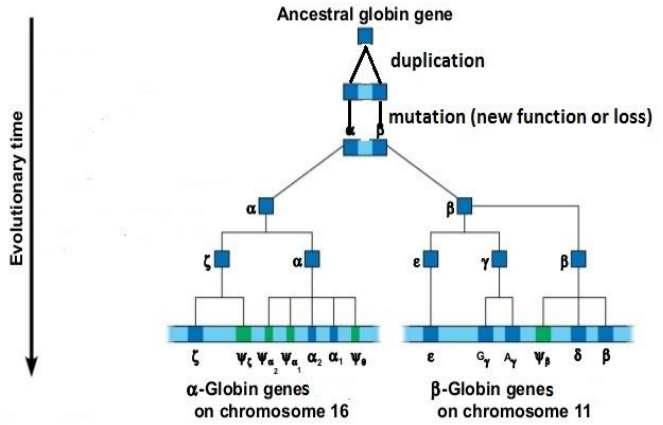
from http://www.web-books.com/MoBio/Free/Ch3F2.htm

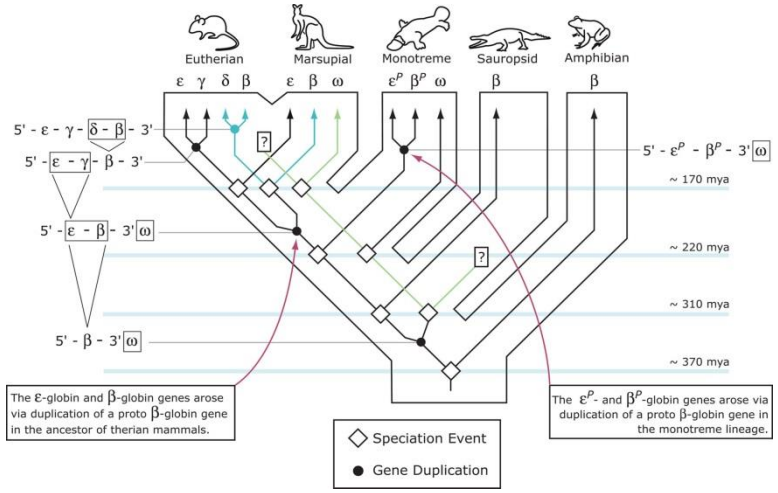


Vijay et al. Nature Reviews Immunology 2003

The evolution of a gene family

- Genes in the same gene family were formed by duplication of a single original gene





http://www.slideshare.net/veneethmathew/21-lecture-genomeandevolution

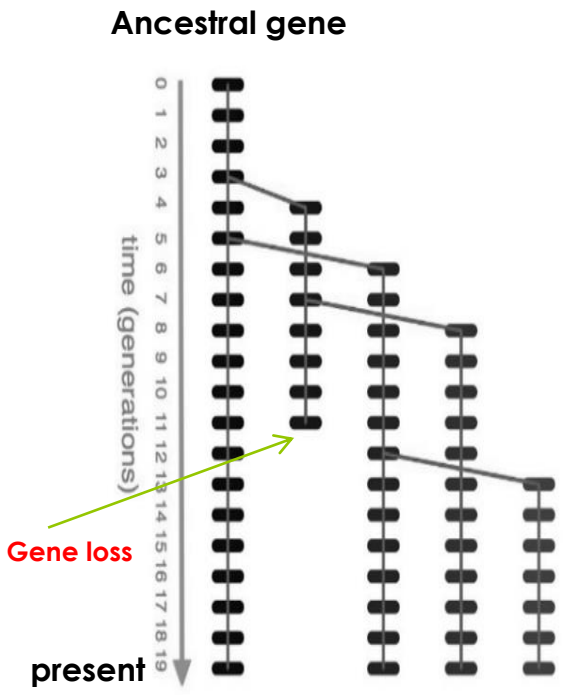
Opazo et al. PNAS February 5, 2008 vol. 105

Importance

- Gene duplications are an essential source of genetic novelty
- Gene duplications/losses are related to diseases
- Inference of the evolutionary forces in shaping the gene family evolution, e.g. natural selection

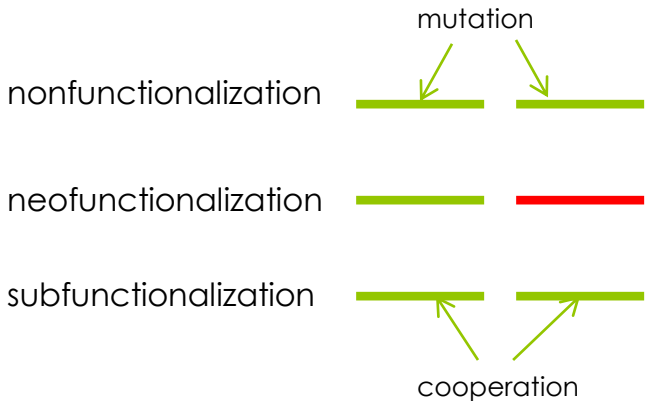
Modeling gene family evolution for a single species

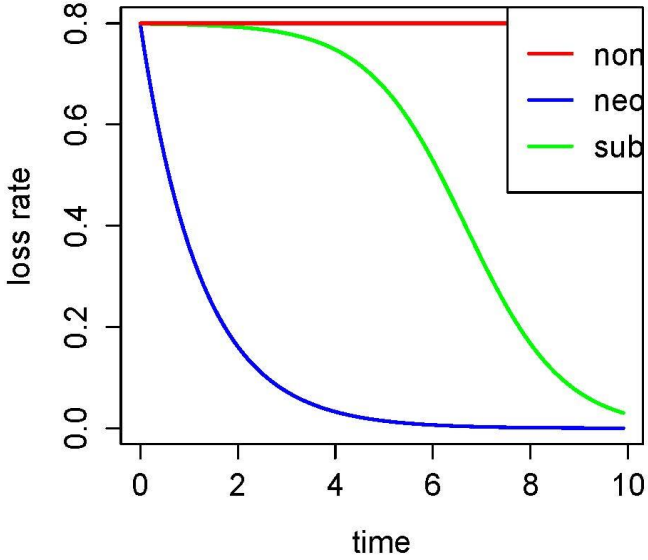
- Duplication produces new gene copies
- Mutation: silence (gene loss) or new function (retained)
- A birth and death process with a constant birth rate λ and a time-dependent death rate μ_t



Modeling death rates

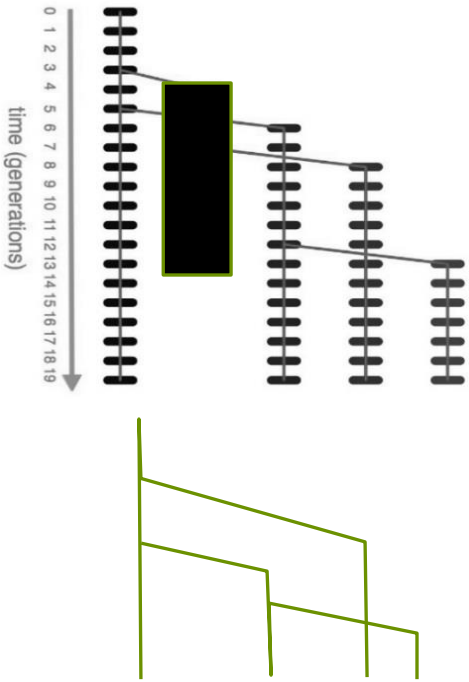
- Fates of a gene duplicate: loss, new function, cooperating with the original copy, etc.
- Nonfunctionalization
- Neofunctionalization
- Subfunctionalization





Modeling the evolution of gene copies in a gene family

- We know that those gene copies have survived to the present
- The evolutionary process of gene copies observed at the present time
- The reconstructed evolutionary process (Nee 1994)



history of gene copies observed at the present time

The reconstructed evolutionary process

- Derived from the birth and death process
- A pure birth process with a time-dependent birth rate λP(t,T)

  λ: the birth rate

  t: duplication time

  T: the present time

$$P(t_i,T)=\left[1+\int_{t_i}^{T}\mu_t e^{\rho(t_i,t)}dt\right]^{-1}$$

$$\rho(t_i,t)=\int_{t_i}^{t}(\mu_s-\lambda)ds$$

The likelihood function of duplication times (Nee 1994)

Likelihood = $(N-1)!\lambda^{N-2} \times$

$$\left\{\prod_{i=3}^{N} P(t_i, T)\right\}(1-u_{x_2})^2 \prod_{i=3}^{N}(1-u_{x_i}).$$

Conditional density function

- The conditional density function of t_i ($i > 2$), given its previous duplication time t_{i-1}, T, and n_T

$$f(t_i \mid t_{i-1}, n_T, T) = \frac{f(t_i \mid t_{i-1}) P(n_T \mid n_{t_i})}{P(n_T \mid n_{t_{i-1}})}$$

- The joint density function of $t = \{t_i \mid i = n_0+1, \ldots, n_T\}$

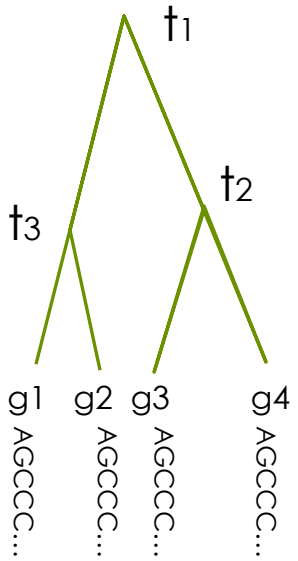
$$f(t \mid n_T, n_0, T) = \frac{\prod_{i=n_0+1}^{n_T} (i-1)\lambda P(t_i, T)(1-\eta_{t_{i-1}, t_i})^{i-1}}{P(n_T \mid n_0)}$$

Simulation and inference

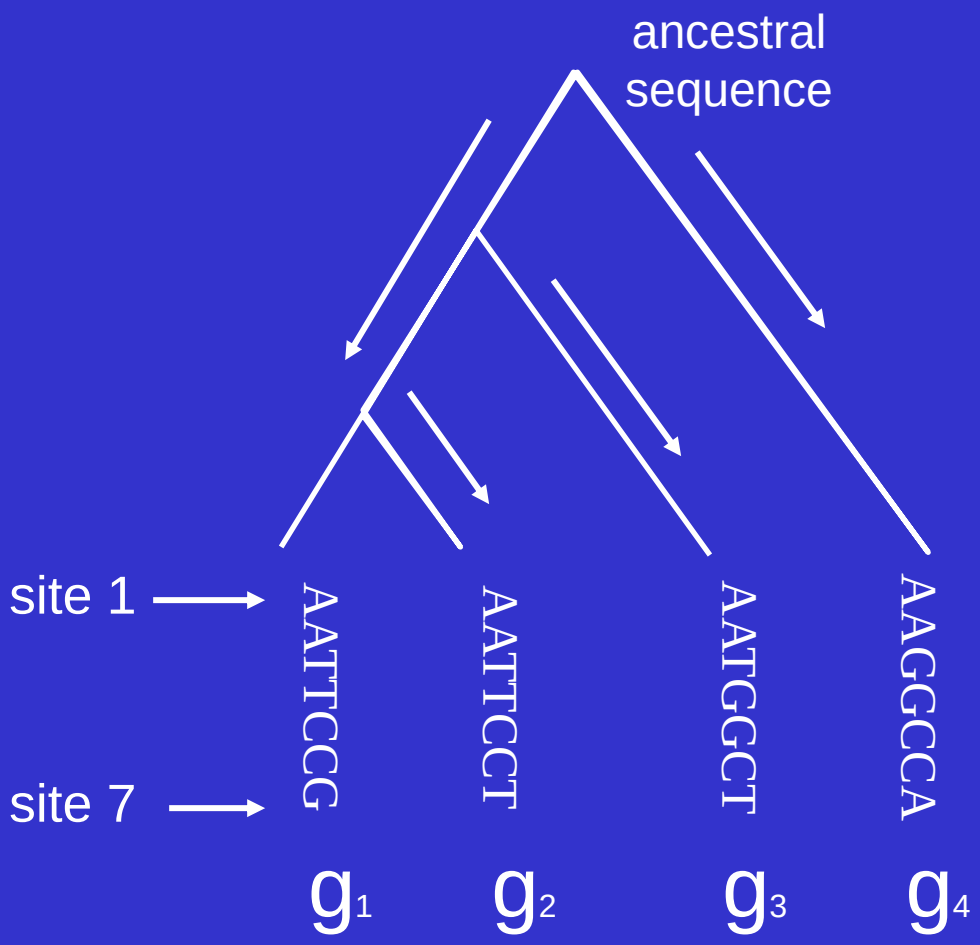
- Simulating data: conditional density function
- Inference: joint probability density function of duplication times
- Given duplication times, we can estimate model parameters (duplication and loss rates)

A Bayesian hierarchical model for gene family evolution

- Data: sequences
- Parameters: (1) duplication and loss rates, (2) the phylogenetic tree representing the evolutionary history of gene copies, (3) parameters in the substitution model
- likelihood = f(D | Tree)
- Prior f(t) is the joint density function of duplication times
- Uniform prior for the tree topology



P(D|Tree)

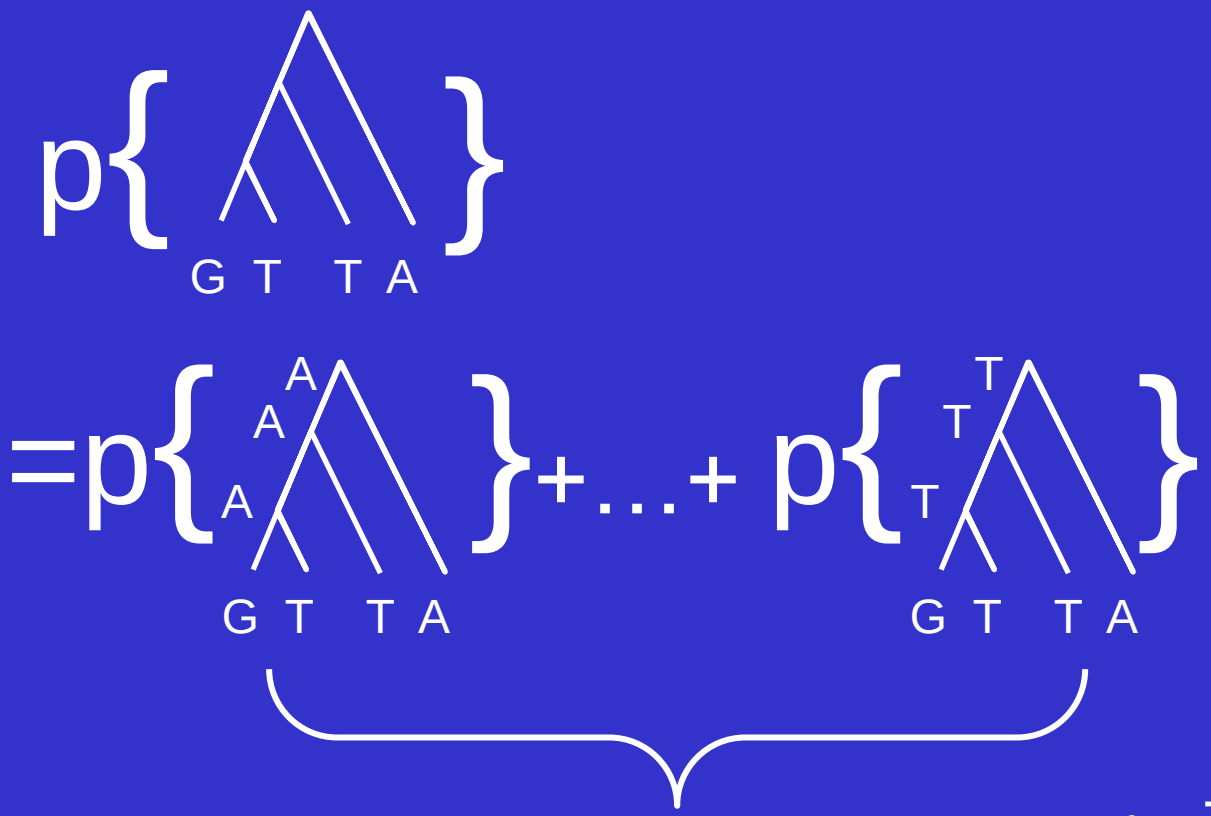


Assumption: nucleotides at different sites evolve independently

$P(D|G) = \Pi p_i$

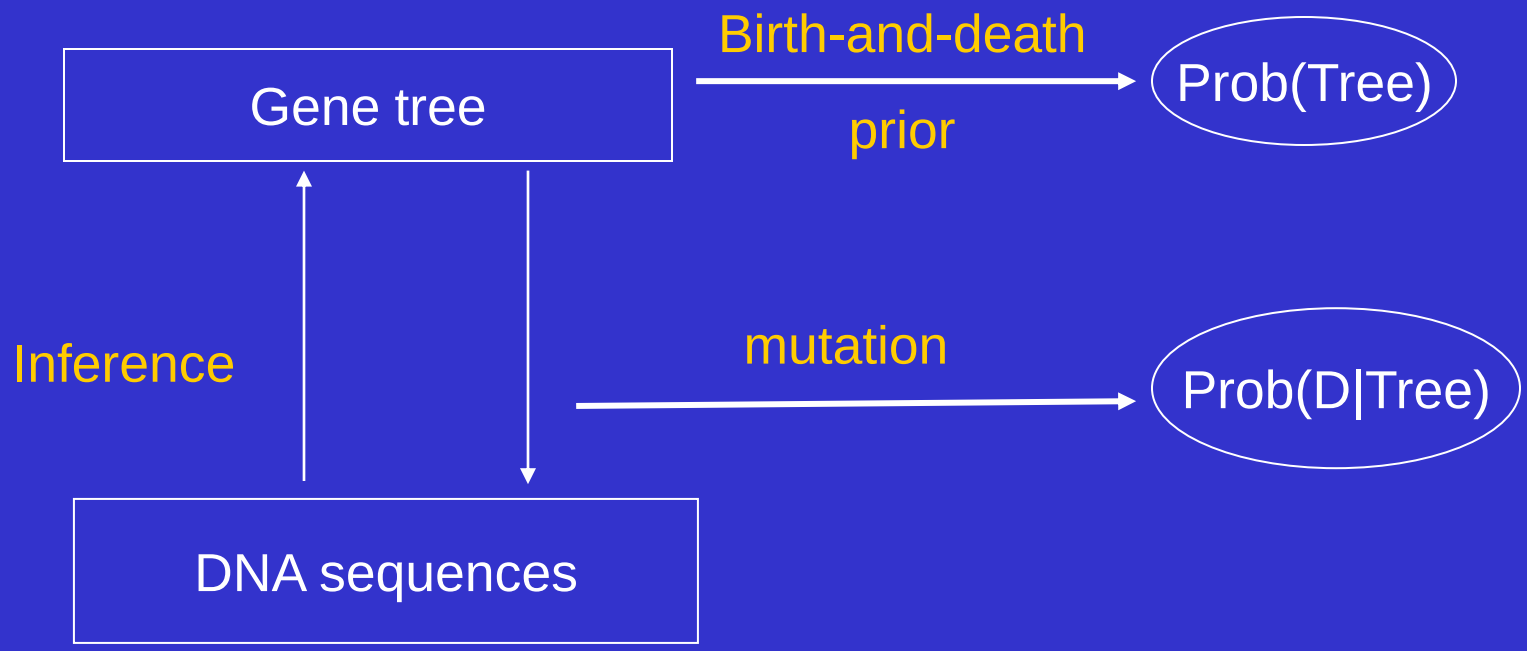
p_i: the probability of the nucleotides at site i.

probability of the nucleotides at a single site given the gene tree G



- Jukes-cantor model
- HKY model.
- GTR model.

Model


Gene tree
Birth-and-death
prior
Prob(Tree)
Inference
mutation
Prob(D|Tree)
DNA sequences

Bayesian inference of gene family evolution

- Testing the homogeneous birth and death rate model

  H_0: homogeneous death rate

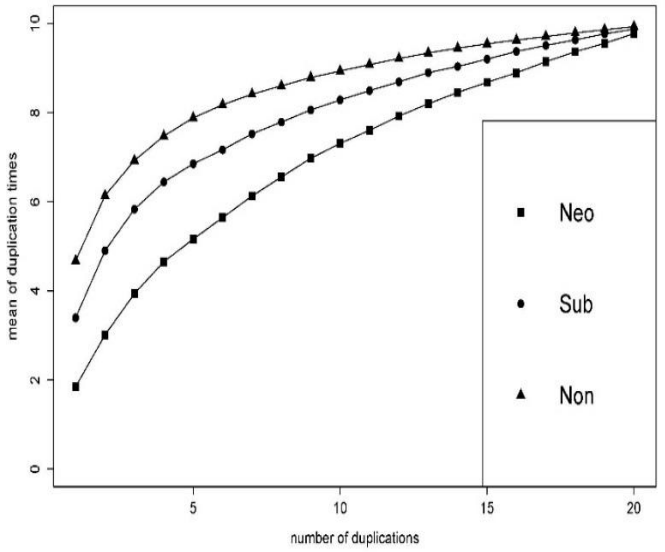
  H_1: heterogeneous death rate

- Model selection: Nonfunctionalization, neofunctionalization, or subfunctionalization

Simulation results

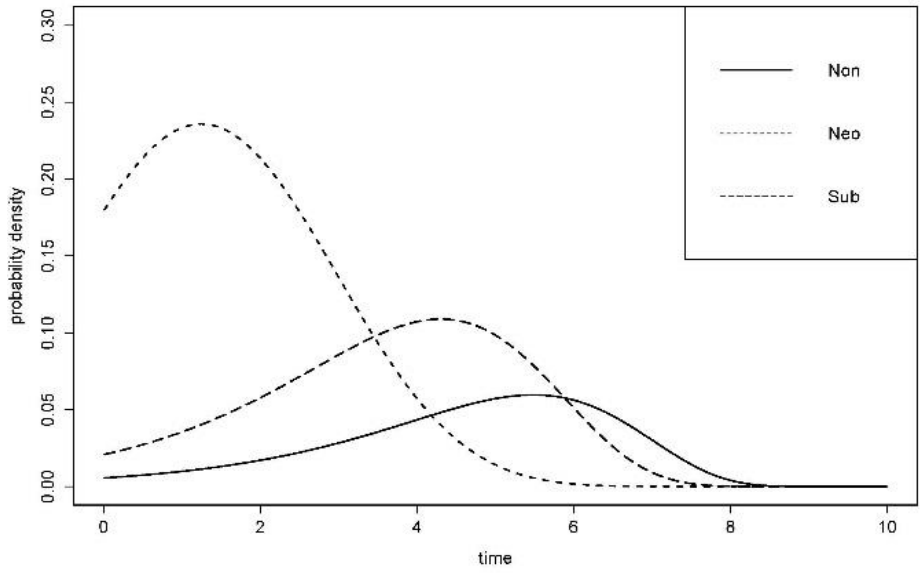
	λ	μ	α
Nonfunctionalization	0.2	0.8	
Neofunctionalization	0.2		0.8
Subfunctionalization	0.2		0.8

Different mechanisms produce different patterns for duplication times



The means of simulated duplication times

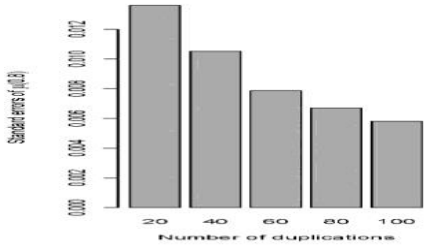
Simulation results

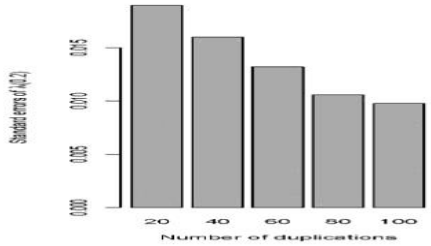


The density curves of the first duplication time

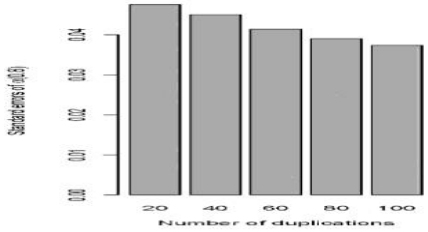
Parameter estimation

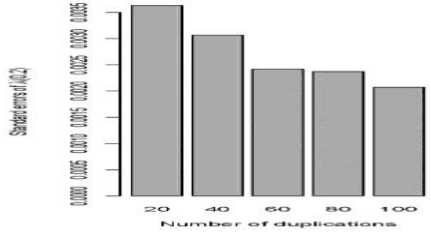
a) nonfunctionalization



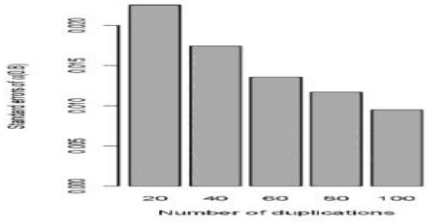


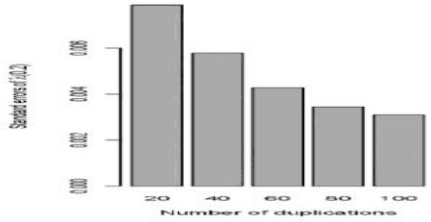
b) neofunctionalization





c) subfunctionalization



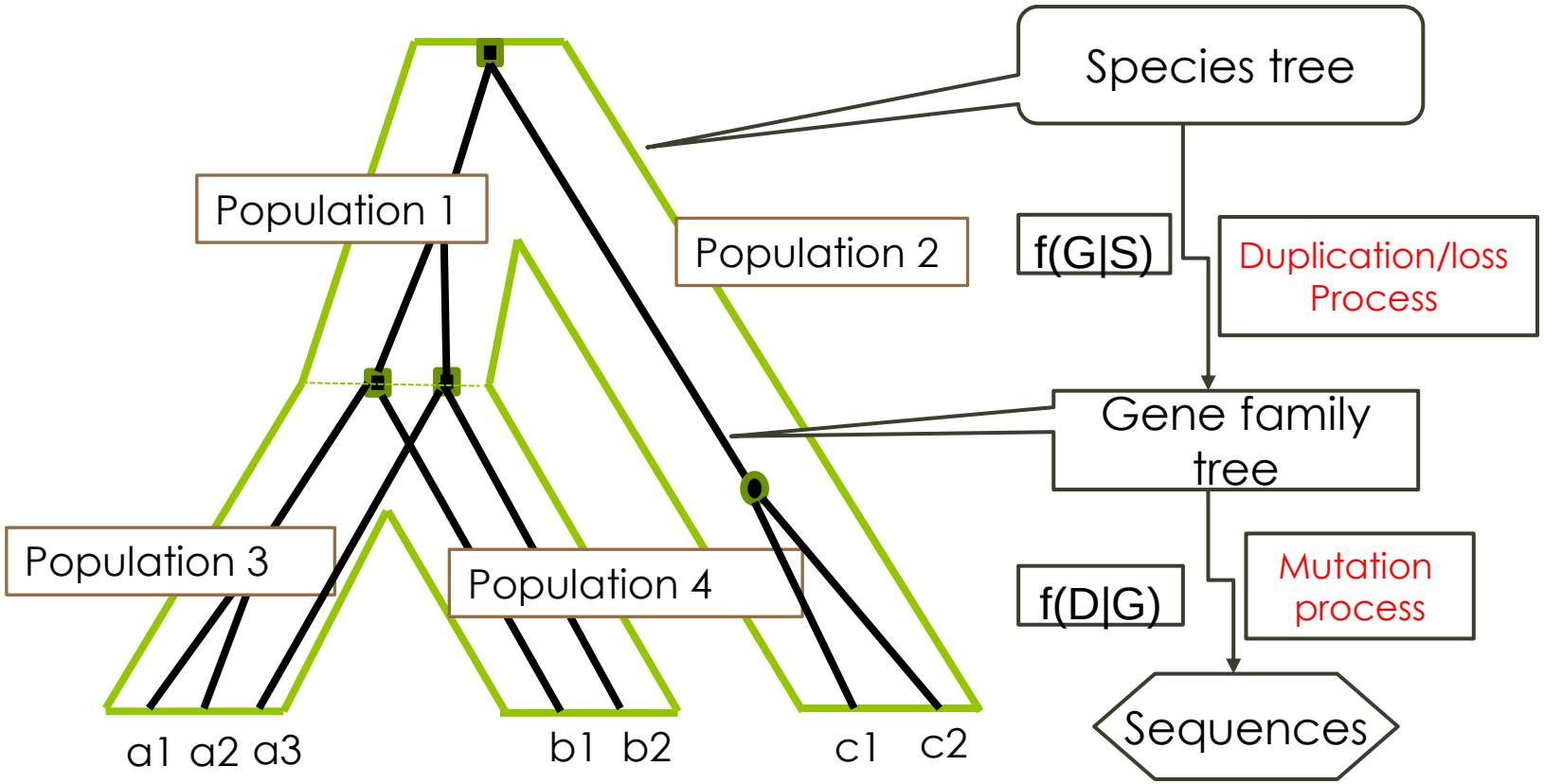


Modeling gene family evolution for multiple species

- The probabilistic density function of a gene family tree given the species tree

$$f(G|S) = \prod_{i=1}^{2n-1} f_i = \prod_{i=1}^{n-1} \frac{f(t_{ij}, j = 1, \ldots, (n_i - m_i)|n_i, m_i, t_{il} - t_{iu})}{\binom{n_i}{2}\binom{n_i-1}{2} \ldots \binom{m_i+1}{2}}$$

The Hierarchical model



Bayesian inference

- The probability density functions, P(G | S) and P(D | G)
- Prior distribution of model parameters
  - (1) Topology of the species tree is fixed
  - (2) Branch length of the species tree
  - (3) other model parameters

Future directions

- Orthologs vs paralogs
- The evolution of orthologs follows a coalescence process
- The evolution of paralogs follows a duplication process
- Incorporate coalescence process into the model

Acknowledgements

- Dr. David Liberles
- Jing Zhao
- Ashley Teufel
- NSF

Questions