A PROBABILISTIC MODEL FOR GENE FAMILY EVOLUTION

by

JING ZHAO

(Under the Direction of Liang Liu)

ABSTRACT

Studies in gene family evolution have revealed invaluable information about the evolutionary relationships among genes in a gene family and the underlying gene retention mechanisms that help shape the gene family across species. A gene family is formed by gene duplication and loss events during the evolutionary history of species. More importantly, gene duplication is the major source of novelties (i.e. raw materials) on which evolutionary forces may have acted. However, the probabilistic models of gene duplication and loss in the context of phylogenetic trees are still limited in the current literature, wherein no model has taken into account the effect of gene retention mechanisms such as neofunctionalization and subfunctionalization. Thus, it is essential to build a probabilistic frame work to understand gene family evolution.

In this dissertation, we are focusing on building a Bayesian hierarchical model for gene family evolution, in which different gene retention mechanisms are incorporated through the non-homogeneous birth and death process of gene copies. We first develop a birth-death age model for gene family evolution in a single population, in which the loss rates of duplicated genes are functions of the ages of genes. From the birth-death age model, we have derived the probability density function of a gene family tree given the species tree. The probability distribution can be used to estimate model parameters and to simulate gene family data. Moreover, we extend the age-

dependent birth and death model to multiple populations in the context of phylogenetic trees, where the joint probability density function of duplication times and number of gene copies at the internal nodes are given. Finally, we propose a Bayesian hierarchical model for gene family evolution, which involves two stochastic processes -mutation process of DNA sequences and the birth and death process of genes.

INDEX WORDS: Gene Duplication; Phylogenetic Methods; Probabilistic Models; Birth-Death Processes; Stochastic Processes

A PROBABILISTIC MODEL FOR GENE FAMILY EVOLUTION

by

JING ZHAO

BS, Shandong University, China, 2005

MS, Shandong University, China, 2010

MS, The University of Georgia, 2015

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2016

A PROBABILISTIC MODEL FOR GENE FAMILY EVOLUTION

by

JING ZHAO

Major Professor:	Liang Liu
Committee:	Jonathan Arnold
	James Leebens-Mack
	William McCormick
	Wenxuan Zhong

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2016

ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my advisor, Dr. Liang Liu, for his strong support and continuous guidance of my PhD study and research. His patience, motivation, enthusiasm, and confidence will always encourage me in my professional life.

I would like to thank the other members of my committee: Dr. William P. McCormick, Dr. Jonathan Arnold, Dr. James Leebens-Mack, and Dr. Wenxuan Zhong for their insightful suggestions and valuable comments for my proposal. My sincere appreciation also goes to all the professors whose courses I took for their devoted instructions during my graduate years.

I also want to thank my dear classmates, Xianyan Chen, Xinyi li, Zhen Yan, Cristian Caranica, Yuanwen Wang, Soyeon Jung and many others, for their friendship, encouragement, and all the fun we have had in the last four years.

Finally, and most importantly, I would like to thank my beloved parents, Kun Zhao and Li Chai, my husband, Qin Ma, and my two lovely children, Crystal and Ethan. Without their invaluable love and constant support, I could never have accomplished the most important goal in my life.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 INTRODUCTION OF GENE FAMILY EVOLUTION

A gene family is a set of genes that share essential characteristics. Since genes in a family have similar DNA sequences [1-3], the products that guided by these genes would have a similar structure or function. There are also occasions that genes with dissimilar sequences are grouped in one family since the proteins they yield are working together to perform one function [4]. Thus the classification of genes into a gene family can provide researchers with important information about the relationship among genes. The fact that there are a large number of genes per family also indicates that the newly arisen gene duplicates are possibly major sources of evolutionary novelties [5-8]. With the advancement of sequencing techniques, new genes are identified increasingly, for which the functional annotations of new genes are always challenging. With clustering of genes into gene families, researchers are able to predict the functions of newly identified genes through their similarity to known genes. In addition, genes that are related to a specific disease may be identified based on information of the related gene families.

Gene families are formed by multiple duplications of an ancestral gene that occur within a lineage or through speciation events [9]. Although a large portion of the duplicate genes may become pseudogenes by degenerative mutations or lost due to population dynamics, some are able to evolve novel functions that can be preserved in the population permanently [10, 11]. Thus estimation of the timing and mode of gene duplications along the evolutionary history of species

provide important information about driven mechanisms by which the genomes of organisms evolved and the genes with novel functions arose [12].

With the availability of "Omics" data, such as genomics and transcriptomics data, phylogenetic study has entered the phylogenomics era Pennisi [13]. Especially, gene family evolution is an important component in phylogenomics, which integrates evolutionary analysis and genomics study to understand the relationships between genes [14]. Many of the phylogenetic approaches can be used in analyzing gene families due to the fact that evolutionary information of gene duplication and divergence is revealed through the hierarchical aligned genes in gene families[15]. Thus we will review major approaches that are related in modeling gene family.

1.2 LITERATURE REVIEW ON MODELS OF GENE FAMILY EVOLUTION

Research in modeling gene families has attracted extensive attention due to its importance in molecular evolution. Genes in a family evolve within the branches of a species tree through gene duplication and speciation. Each gene tree reflects a duplication and divergence process of genes inside a species phylogeny, yet often incongruent with it. The presence of incongruence is the result of three major biological processes; those are horizontal gene transfer, lineage sorting (deep coalescence) and gene duplication and extinction [16, 17]. Thus it is desirable to integrate this information in order to provide a better estimation of gene family phylogeny. Extensive literature appears which depicts the relationship between gene family tree and species tree by incorporating one or some of these factors. Among these methods, parsimony based approach and probabilistic framework are two major strategies.

1.2.1 *Parsimony methods*

Parsimony methods attempt to reconcile the relationship between gene tree and species tree by minimizing the total number of specific events. Goodman et al. [18, 19] firstly introduced the

concept of reconciled tree in the context of discordance between the mammalian haemoglobin gene trees and the mammal phylogeny that was accepted previously. In this paper, a parsimony strategy was developed to reconcile the gene lineages with the species phylogeny by minimizing the total number of nucleotide substitutions, gene duplications and gene expression events. However, Page [20, 21] formalized the definition of reconciled tree which combines a species tree and its corresponding gene tree into a single summary of the historical association by assuming that no horizontal gene transfer occurred. Page and Charleston [19] developed a method to estimate unknown species tree from the known gene trees regarding gene duplications and losses, along with a heuristic algorithm realizing the method and a program implementing the algorithm. In a subsequent paper, Page [22] applied this technique to the vertebrate phylogenies with nine genes. Page and Cotton [23] discussed the development of the software "GENETREE" in accommodation with the uncertainty in gene phylogenies through resampling gene families.

There are methods minimizing different criterion other than nucleotide substitutions, gene duplications and gene expression events. It is assumed in [24] that a single duplication episode is the causation of the duplications at the internodes of a species tree and thus be minimized to establish the gene phylogeny and species phylogeny simultaneously. Another method proposed by Simmons, Bailey and Nixon [25] incorporates duplicated and unduplicated genes through uninode coding such that the effect of gene duplications is excluded in species tree estimation.

Since the parsimony criterion being minimized is the number of certain events, it is easy to implement and produces a clear interpretation. But it is a difficult job in assigning weights to associated events due to their diverged characters [26]. Another problem with parsimony approach appears in the generated solutions, in which there is no summary over the large number of possible solutions [26].

1.2.2 *Probabilistic framework*

The probabilistic models of gene family evolution, though computationally expensive, can overcome weaknesses encountered by parsimony methods. It provides a probability distribution of a gene family phylogeny within a species tree and the most likely species phylogeny can be reconstructed from the known gene trees. In addition, parameters can be defined and incorporated into different biological processes under a probabilistic framework and thus reveal the information of the underlying evolutionary forces.

The first probabilistic model of gene evolution within a species phylogeny was proposed by Arvestad et al. [27] based on a birth-death process and a tool is developed to perform orthology analysis as well as gene tree/species tree reconciliation. Later this model was extended to incorporate the sequence evolutionary model and a MCMC algorithm is provided for orthology analysis and reconstruction of gene tree and species tree by Arvestadet al. [28]. Akerborg et al. [29] further developed a more complicated model incorporating gene evolution, sequence evolution, and a substitution rates model which was implemented through a Bayesian analysis tool. Sjostrand [30] from the same group developed a phylogenetic program "PrIME-DLRS" in the context of gene duplication and loss with a relaxed molecular clock which is applied in the area of homologous gene families. Rasmussen and Kellis [31] pointed out the inaccuracies of traditional phylogenetic reconstruction methods and improved the accuracy by integrating multiple parameters in evolution including gene duplication and loss rates, speciation times, and varied substitution rate over both species and loci in a Bayesian framework named "SPIMAP". These works are aiming to estimate gene tree given a fixed species tree with constant gene duplication and loss rates and assumption of molecular clock. Advances have been made in a recent model of genome scale coestimation of gene and species tree given in Boussau [32]. In gene tree

reconciliation, the branch lengths of gene trees are ignored and the rates of sequence evolution for each gene family are constant for the purpose of fast computation. While the duplication and loss rates for each branch are assumed to be different for an improved accuracy of estimation. Konrad et al.[33] have established a maximum likelihood framework based on a modified Weibull hazard function under different duplicate gene loss/retention mechanisms.

Next, I will present a general designation of probabilistic framework including the establishment of the likelihood function and the implement of the Bayesian estimation methods. The first formulation of the likelihood of a given species tree is proposed by Maddison [34] in the context of deep coalescence, which is the product of the probability distribution of the sequences over all loci given the species tree:

$$f(sequences|species\ tree)$$
$$= \prod_{loci} \sum_{gene\ trees} [P(sequences|gene\ tree)P(gene\ tree|species\ tree)] \quad (1.1)$$

The first term in (1.1) is the probability of the observed sequences given the known gene phylogeny which comes from the sequence evolutionary model [34-36]. The second term is the probability distribution of a gene tree given a species phylogeny, which can be directly obtained in coalescent theory. The computation of this likelihood of species tree is expensive because not only all possible species trees should be visited but also the potential gene trees should be considered. Progresses have been made to the second term in this formulation in the background of gene duplication and loss [27-29], [33, 37, 38].

The Bayesian method provides an alternative way to do statistical inference that treats parameters as random variables with given prior distributions. The prior represents the initial guess about the distributions of parameters without the data. The initial guess will be improved by the data using Bayes theorem, which is

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{f(D)} \quad (1.2)$$

Estimation on the parameter θ relies on the posterior probability distribution $f(\theta|D)$ which is a product of likelihood function $f(D|\theta)$and prior $f(\theta)$. The likelihood function of sequence evolution given the species tree is given in Equation 1.1 whereas the selection of a good prior is an open question. If the information about parameters is abundant from the data, the priors will not influence the posterior too much. In this respect, the Bayesian estimation is similar to the maximum likelihood method and a non-informative prior may be chosen if there is no information about the prior. When the parameters are correlated, a non-informative prior for one parameter may affect the prior for another parameter such that the prior of the second parameter is no longer non-informative. In addition, different users may choose different priors. Thus phylogenetic software like MrBayes[39] provides great options of priors for users. And the results from different priors could be compared to check the sensitivity of the model to the priors.

Bayesian inference is made through the posterior distribution $f(\theta|D)$. However, the normalization constant $f(D)$ in the posterior probability distribution is analytically intractable. In this case, numerical methods such as Markov Chain Monte Carlo (MCMC) are adopted to estimate the posterior distribution of model parameters. It has been shown that the MCMC algorithm converges to the posterior distribution of interest under certain regular conditions. But it may take a long time for the algorithm to converge as the number of parameters increases. Because the MCMC algorithm may converge to a local maximum, it is important to monitor the convergence of MCMC.

1.2.3 Birth and death process (BDP) in gene family evolution

The process of gene family evolution can be seen as ancestral genes that evolve inside a species tree, which is similar to the process of lineage speciation and extinction. In this respect, most

methods inferring changes of lineage diversification and gene family evolution over time are based on the BDP. The theory of this process was first introduced by Kendall [40], in which each lineage divide into two at a duplication rate or dies at a loss rate. Thompson [41] obtained the joint density function of the number of lineages and the branching times by investigating the BDP of these lineages surviving to the current time. Nee et al. [42] further defined and examined a reconstructed evolutionary process (Figure 1.1), in which one of the lineages being investigated survived until being sampled. The likelihood function of a reconstructed phylogeny is derived in this paper based on the BDP conditional on the survival of at least one lineage. Rannala and Yang [43] developed a birth-death phylogenetic model conditional on the number of lineages at sampling time for estimating species tree from molecular sequence data. Aldous and Popovic [44] proposed a continuous-time critical branching process conditioned on the number of species at present, with the assumption that the birth and death rates are identical in macroevolution. Later Gernhard [45, 46] relaxed previous assumption and allow variable birth and death rates. Stadler [47] derived the probability density function of a phylogenetic tree with the assumption of constant birth and death rates.

Recently, researchers become interested in time-dependent BDP which serve as a model of performing hypothesis-driven research. Rabosky [48] distinguished rate-variable models of diversification from rate-constant models by fitting birth and death models using likelihood methods. Hohna [49, 50] and Hallinan [51] studied the reconstructed process with time-dependent rates in a more general setting by relaxing the assumptions about the number of species and the time of the process.

Next I will present a brief introduction to the theory of generalized BDP.

The BDP is a continuous-time Markov chain that models the number of members in a population, where each member can give birth to a new one or die at some time [52]. I will describe a general BDP in the context of gene family evolution. Let n be the number of genes in a gene family at time t, which is a stochastic variable taking values 1,2, ... $P_n(t)$ is the probability that there are n genes at time t in the population. The birth rate λ_t and death rate μ_t at any given time t per gene is a functions of time t for generalized BDP. And let δ be a small time interval. When δ is very small, the probability of an event occurs during $(t, t + \delta)$ is approximately $\lambda_t\delta$ or $\mu_t\delta$.Therefore, the probability of a birth in the interval $(t, t + \delta)$ is $P_{n+1}(t + \delta) = \lambda_t n\delta + o(\delta)$ and the probability of a death in the interval $(t, t + \delta)$ is $P_{n-1}(t + \delta) = \mu_t n\delta + o(\delta)$. The probability of no births or deaths occurring during $(t, t + \delta)$ is $P_n(t + \delta) = (1 - \lambda_t n\delta - \mu_t n\delta) + o(\delta)$. we assume the initial condition of a gene family with one common ancestor, so that $P_1(0) = 1$ and $P_n(0) = 0, n > 1$. Letting $\delta \to 0$, the differential equations for $P_n(t)$ is:

$$\begin{cases} P_n'(t) = (n+1)\mu_t P_{n+1}(t) + (n-1)\lambda_t P_{n-1}(t) - n(\lambda_t + \mu_t)P_n(t), n \geq 1 \\ P_0'(t) = \mu_t P_1(t) \end{cases} \quad (1.3)$$

By solving the above equations through the generating function, the solution for $P_n(t)$ is given in Kendall [52], that is

$$\begin{cases} P_0(t) = 1 - \dfrac{1}{1 + \int_0^t e^{\rho(0,x)}\mu(x)dx} \\ P_n(t) = \big(1 - P_0(t)\big)(1 - V_t)V_t^{n-1}, n \geq 1 \end{cases}$$

$\rho(t_1, t_2) = \int_{t_1}^{t_2}(\mu(x) - \lambda(x))dx$ and $V_t = 1 - \frac{e^{\rho}}{1+\int_0^t e^{\rho(0,x)}\mu(x)dx}$.

In addition, the probability of a single lineage at time t does not extinct at a later time T is calculated by Kendall[52]

$$P(t, T) = \frac{1}{1 + \int_t^T e^{\rho(t,x)}\mu(x)dx} \quad (1.4)$$

For the reconstructed phylogeny, Nee et al. [42] derived the probability mass function of the number of reconstructed lineages (n_T) at time T given the number of reconstructed lineage ($\mathrm{n_t}$) at an earlier time t , $\mathrm{P}(\mathrm{n}_T = \mathrm{n}|\mathrm{n_t} = 1)$, based on the above equations

$$\mathrm{P}(\mathrm{n}_T = \mathrm{n}|\mathrm{n_t} = 1) = \left(1 - u_t \frac{P(0,T)}{P(0,t)}\right)\left(u_t \frac{P(0,T)}{P(0,t)}\right)^{n-1} \quad (1.5)$$

where $u_t = 1 - \mathrm{P}(0,\mathrm{t})\mathrm{e}^{\rho(\mathrm{o,t})}$.

In a recent literature, Hallinan [51] gave a set of generalized results, in which the probability mass function of the number of reconstructed lineages at some time conditional on the number of lineages at any other time along the process (see Hallinan [51] for more details).

In this dissertation, we develop a probabilistic model for gene family evolution with age-dependent loss rate, wherein those results in generalized BDP will be adopted and modified accordingly.

1.3 OUTLINE OF THIS DISSERTATION

In Chapter 2, we consider the gene family evolution in a single population and establish a probabilistic model for this process. More specifically, we describe a generalized birth-death process for modeling the fates of gene duplication. Starting with a single population corresponding to the branch of a species tree and with assumption of a clock that starts ticking for each duplicate at its birth, an age-dependent birth and death process is developed by extending the results from the time-dependent birth and death process. The implementation of such models in a full phylogenetic framework is expected to enable large scale probabilistic analysis of duplicates in comparative genomic studies. See chapter 2 for details.

In Chapter 3, we extend our generalized birth and death process in one population to a multi-population scenario. Especially, we develop a probabilistic framework to model gene family evolution in the context of species tree, in which three evolutionary mechanisms of gene retention

are incorporated. This study is accomplished by developing the joint distribution of the gene counts in the internal nodes and duplication times in all branches of the species tree. A simulation study is then performed to examine the accuracy of parameter estimation through the Bayesian methods.

In Chapter 4, we introduce a conceptual hierarchical Bayesian framework of gene family evolution. This model involves two stochastic process: a mutation process of sequence data within gene family tree and a birth and death process of gene family tree in a species tree. Further, a posteriori probabilistic model is constructed and possible applications is discussed.

Note that each chapter is self-contained in terms of development and assessment of the above methods, but we give an overall conclusion for all of the chapters in Chapter five.

1.4 REFERENCES

1. Ohta, T., *Simulating evolution by gene duplication.* Genetics, 1987. **115**(1): p. 207-13.
2. Fortna, A., et al., *Lineage-specific gene duplication and loss in human and great ape evolution.* PLoS Biol, 2004. **2**(7): p. E207.
3. Nei, M. and A.P. Rooney, *Concerted and birth-and-death evolution of multigene families.* Annu Rev Genet, 2005. **39**: p. 121-52.
4. [Internet]., N.L.o.M.U.G.H.R., *Bethesda (MD): The Library; 2013 Sep 16. What is a gene family?; [cited 2013 Sep 19]; [about 3 screens].* Available from: http://ghr.nlm.nih.gov/condition/cystic-fibrosis.
5. Lynch, M., et al., *The probability of preservation of a newly arisen gene duplicate.* Genetics, 2001. **159**(4): p. 1789-804.
6. Hurles, M., *Gene duplication: the genomic trade in spare parts.* PLoS Biol, 2004. **2**(7): p. E206.
7. Ohta, T., *Role of gene duplication in evolution.* Genome, 1989. **31**(1): p. 304-10.

8. Zhang, J.Z., *Evolution by gene duplication: an update.* Trends in Ecology & Evolution, 2003. **18**(6): p. 292-298.

9. Silver, L., et al., *Genetics: From Genes to Genomes*. 2010: McGraw-Hill Education.

10. Lynch, M., *Genomics. Gene duplication and evolution.* Science, 2002. **297**(5583): p. 945-7.

11. Teufel, A.I., J. Masel, and D.A. Liberles, *What Fraction of Duplicates Observed in Recently Sequenced Genomes Is Segregating and Destined to Fail to Fix?* Genome Biol Evol, 2015.

12. Hahn, M.W., et al., *Estimating the tempo and mode of gene family evolution from comparative genomic data.* Genome Res, 2005. **15**(8): p. 1153-60.

13. Pennisi, E., *Evolution. Building the tree of life, genome by genome.* Science, 2008. **320**(5884): p. 1716-7.

14. Eisen, J.A. and C.M. Fraser, *Phylogenomics: intersection of evolution and genomics.* Science, 2003. **300**(5626): p. 1706-7.

15. Thornton, J.W. and R. DeSalle, *Gene family evolution and homology: genomics meets phylogenetics.* Annu Rev Genomics Hum Genet, 2000. **1**: p. 41-73.

16. Salichos, L. and A. Rokas, *Inferring ancient divergences requires genes with strong phylogenetic signals.* Nature, 2013. **497**(7449): p. 327-31.

17. Rokas, A., et al., *Genome-scale approaches to resolving incongruence in molecular phylogenies.* Nature, 2003. **425**(6960): p. 798-804.

18. Goodman, M., et al., *Fitting the Gene Lineage into Its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences.* Systematic Zoology, 1979. **28**(2): p. 132-163.

19. Page, R.D.M. and M.A. Charleston, *From gene to organismal phylogeny: Reconciled trees and the gene tree species tree problem.* Molecular Phylogenetics and Evolution, 1997. **7**(2): p. 231-240.

20. Page, R.D.M., *Maps between Trees and Cladistic-Analysis of Historical Associations among Genes, Organisms, and Areas.* Systematic Biology, 1994. **43**(1): p. 58-77.

21. Page, R.D.M., *Component Analysis - a Valiant Failure.* Cladistics-the International Journal of the Willi Hennig Society, 1990. **6**(2): p. 119-136.

22. Page, R.D.M., *Extracting species trees from complex gene trees: Reconciled trees and vertebrate phylogeny.* Molecular Phylogenetics and Evolution, 2000. **14**(1): p. 89-106.

23. Page, R.D.M. and J.A. Cotton, *Genetree: A Tool for Exploring Gene Family Evolution*, in *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, D. Sankoff and J.H. Nadeau, Editors. 2000, Springer Netherlands: Dordrecht. p. 525-536.

24. Guigo, R., I. Muchnik, and T.F. Smith, *Reconstruction of ancient molecular phylogeny.* Molecular Phylogenetics and Evolution, 1996. **6**(2): p. 189-213.

25. Simmons, M.P., C.D. Bailey, and K.C. Nixon, *Phylogeny reconstruction using duplicate genes.* Molecular Biology and Evolution, 2000. **17**(4): p. 469-473.

26. Szollosi, G.J., et al., *The inference of gene trees with species trees.* Syst Biol, 2015. **64**(1): p. e42-62.

27. Arvestad, L., et al., *Bayesian gene/species tree reconciliation and orthology analysis using MCMC.* Bioinformatics, 2003. **19 Suppl 1**: p. i7-15.

28. Arvestad, L., et al., *Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution*, in *Proceedings of the eighth*

annual international conference on Resaerch in computational molecular biology. 2004, ACM: San Diego, California, USA. p. 326-335.

29. Akerborg, O., et al., *Simultaneous Bayesian gene tree reconstruction and reconciliation analysis.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(14): p. 5714-5719.

30. Sjöstrand, J., et al., *DLRS: gene tree evolution in light of a species tree.* Bioinformatics, 2012. **28**(22): p. 2994-2995.

31. Rasmussen, M.D. and M. Kellis, *Accurate gene-tree reconstruction by learning gene-and species-specific substitution rates across multiple complete genomes.* Genome research, 2007. **17**(12): p. 1932-1942.

32. Boussau, B., et al., *Genome-scale coestimation of species and gene trees.* Genome research, 2013. **23**(2): p. 323-330.

33. Konrad, A., et al., *Toward a general model for the evolutionary dynamics of gene duplicates.* Genome Biol Evol, 2011. **3**: p. 1197-209.

34. Maddison, W.P., *Gene trees in species trees.* Systematic Biology, 1997. **46**(3): p. 523-536.

35. Felsenstein, J., *A likelihood approach to character weighting and what it tells us about parsimony and compatibility.* Biological Journal of the Linnean Society, 1981. **16**(3): p. 183-196.

36. Nakhleh, L., *Computational approaches to species phylogeny inference and gene tree reconciliation.* Trends in Ecology & Evolution, 2013. **28**(12): p. 719-728.

37. Górecki, P. *Reconciliation problems for duplication, loss and horizontal gene transfer*. in *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*. 2004. ACM.

38. Rasmussen, M.D. and M. Kellis, *A Bayesian approach for fast and accurate gene tree reconstruction.* Molecular Biology and Evolution, 2011. **28**(1): p. 273-290.

39. Ronquist, F. and J.P. Huelsenbeck, *MrBayes 3: Bayesian phylogenetic inference under mixed models.* Bioinformatics, 2003. **19**(12): p. 1572-1574.

40. DG, K., *On the generalized birth-and-death process.* Ann Math Stat, 1948. **19(1):1–15**.

41. Thompson, *The likelihood approach. In: Human evolutionary trees.* 1975.

42. Nee, S., R.M. May, and P.H. Harvey, *The reconstructed evolutionary process.* Philos Trans R Soc Lond B Biol Sci, 1994. **344**(1309): p. 305-11.

43. Rannala, B. and Z. Yang, *Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference.* J Mol Evol, 1996. **43**(3): p. 304-11.

44. Aldous, D. and L. Popovic, *A critical branching process model for biodiversity.* Advances in Applied Probability, 2005. **37**(4): p. 1094-1115.

45. Gernhard, T., *The conditioned reconstructed process.* Journal of Theoretical Biology, 2008. **253**(4): p. 769-778.

46. Gernhard, T., *New Analytic Results for Speciation Times in Neutral Models.* Bulletin of Mathematical Biology, 2008. **70**(4): p. 1082-1097.

47. Stadler, T., *Sampling-through-time in birth-death trees.* J Theor Biol, 2010. **267**(3): p. 396-404.

48. Rabosky, D.L., *Likelihood methods for detecting temporal shifts in diversification rates.* Evolution, 2006. **60**(6): p. 1152-64.

49. Hohna, S., *Fast simulation of reconstructed phylogenies under global time-dependent birth-death processes.* Bioinformatics, 2013. **29**(11): p. 1367-74.

50. Hohna, S., *The time-dependent reconstructed evolutionary process with a key-role for mass-extinction events.* J Theor Biol, 2015. **380**: p. 321-31.

51. Hallinan, N., *The generalized time variable reconstructed birth-death process.* J Theor Biol, 2012. **300**: p. 265-76.

52. Kendall, D.G., *On the Generalized Birth-and-Death Process.* Annals of Mathematical Statistics, 1948. **19**(1): p. 1-15.

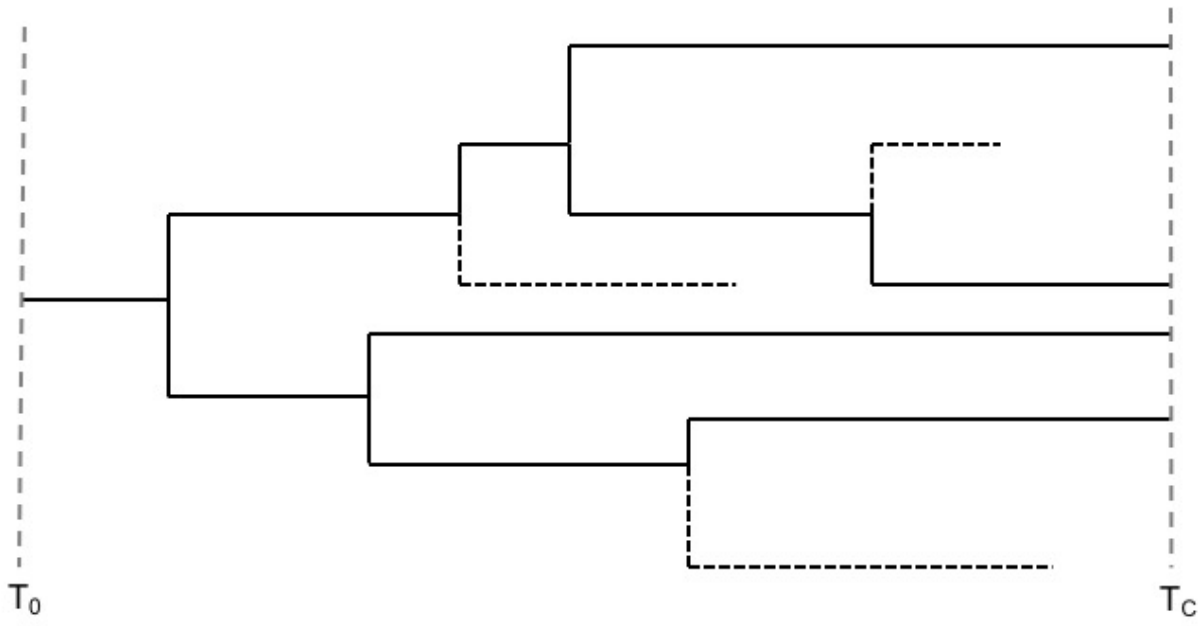

Figure 1.1 The reconstructed evolutionary process introduced by Nee et al.: T_0 is the starting time and T_C denotes present time. The solid black lines are the lineages that can survive to the current time and the dashed black lines are the lineages that distinct in the past.

CHAPTER 2

A GENERALIZED BIRTH AND DEATH PROCESS FOR MODELING THE FATES OF GENE DUPLICATION[1]

[1] Zhao, J., Teufel, A., Liberles, D. and Liu, L. *BMC Evol Biol*. (2015), 15: 275.
Reprinted here with permission of publisher.

ABSTRACT

Accurately estimating the timing and mode of gene duplications along the evolutionary history of species can provide invaluable information about underlying mechanisms by which the genomes of organisms evolved and the genes with novel functions arose. Mechanistic models have previously been introduced that allow for probabilistic inference of the evolutionary mechanism for duplicate gene retention based upon the average rate of loss over time of the duplicate. However, there is currently no probabilistic model embedded in a birth-death modeling framework that can take into account the effects of different evolutionary mechanisms of gene retention when analyzing gene family data.

In this study, we describe a generalized birth-death process for modeling the fates of gene duplication. Use of mechanistic models in a phylogenetic framework requires an age-dependent birth-death process. Starting with a single population corresponding to the lineage of a phylogenetic tree and with an assumption of a clock that starts ticking for each duplicate at its birth, an age-dependent birth-death process is developed by extending the results from the time-dependent birth-death process. The implementation of such models in a full phylogenetic framework is expected to enable large scale probabilistic analysis of duplicates in comparative genomic studies.

We develop an age-dependent birth-death model for understanding the mechanisms of gene retention, which allows a gene loss rate dependent on each duplication event. Simulation results indicate that different mechanisms of gene retentions produce distinct likelihood functions, which can be used with genomic data to quantitatively distinguish those mechanisms.

2.1 INTRODUCTION

A gene family is a group of genes with similar sequences that show evidence of descent from a common ancestor [1-3]. This includes orthologs that originate through speciation as well as paralogs (modeled here) that can be found within a species or shared between species from an older duplication event that predated speciation. The large number of genes per family suggests that the newly arisen gene duplicates are potentially major contributors to evolutionary novelties [4-7]. Gene duplication can provide raw genetic material for evolutionary forces to act on. Although a majority of duplicate genes may be silenced by degenerative mutations or lost due to population dynamics, some duplicated genes are able to evolve novel functions permanently preserved in the population [8, 9]. Accurately estimating the timing and mode of gene duplications along the evolutionary history of species can provide invaluable information about underlying mechanisms by which the genomes of organisms evolved and the genes with novel functions arose [10].

Several biological models have been proposed to depict the mechanisms that lead to different evolutionary fates of a gene duplicate [11-14]. Nonfunctionalization refers to the process in which mutations occur on one of the gene duplicates and produce a non-functional protein [11, 15]. The neofunctionalization model [16] assumes that duplication itself does not affect fitness. Although a duplicate is most likely to be pseudogenized by degenerative mutation (nonfunctionalization) or lost due to population dynamics [9], the redundant copy may occasionally acquire a new beneficial function through mutation that will be preferentially preserved in the population. While this function may subsequently be optimized and accommodated within the genome structure (assuming a coding sequence change) by an evolutionary Stokes shift [17], the initial event leading to retention is a single beneficial change.

The waiting time for this single change gives rise to a convexly decaying hazard function when modeled together with non-functionalizing changes and is referred to as the neofunctionalization model (see [15, 18, 19] for a review). The duplication-degeneration-complementation model [20] describes a so-called subfunctionalization mechanism in which two gene copies are partially damaged by degenerated mutations. Both copies must be maintained in order to perform the original function of the gene [21, 22]. This model, called subfunctionalization, involves a waiting time for multiple events to occur as deleterious substitutions accumulate in both copies before the retaining mutation can occur. This waiting time for multiple changes gives rise to a switch from a convex to a concave (sigmoidal) hazard function when modeled together with non-functionalizing mutations (again, see [15, 18, 19] for a review and engaged discussion). In addition to the processes acting on individual genes, large-scale gene duplication events (for example, whole genome duplication) may have occurred and produced multiple interacting genes together creating an additional retention mechanism. Dosage balance promotes the retention of duplicated interaction networks, as loss of interaction stoichiometry can lead to declines in fitness. This gives rise to very different retention dynamics compared to neofunctionalization or subfunctionalization (see [15, 18, 19] for a review). The models described represent one of many conceivable modeling frameworks for duplicate gene retention (see [19] for an enhanced discussion). The models here are used within a single population, reflecting a lineage of a phylogenetic tree, but the ultimate aim is to extend their use into an interspecific phylogenetic framework with the population genetic assumptions that accompany this. Simpler models have already been incorporated into a fuller phylogenetic framework of this nature (see for example [23]).

Accurately reconstructing the evolution of gene families requires informative datasets, powerful mathematical models, and efficient computational algorithms. Advanced biotechnologies provide a vast amount of genetic data for understanding the evolution of gene families [24, 25]. Meanwhile, probabilistic models, describing the process of gene family evolution, significantly enhance our ability to extract useful information from genetic data [26-29]. The birth-death (BD) model [30], which has been broadly applied in analyzing species phylogenies [25, 29, 31, 32], could also be adopted in phylogenetic analysis of gene families [33]. In 1975, Thompson [34] introduced a phylogenetic model based on the birth-death process to understand the evolution of human populations. Under the generalized birth-death model, Nee et al. [35] derived a reconstructed evolutionary process [36] to estimate birth and death rates in a interspecific phylogenetic framework. Rannala and Yang [37] developed a birth-death phylogenetic model for estimating phylogenetic trees from molecular sequence data. Aldous and Popovic [38] proposed a continuous-time critical branching process conditioned on the number of species in the present, with the assumption that the birth and death rates are identical in macroevolution, which was later relaxed by Gernhard [39, 40] to allow variable birth and death rates. With the assumption of constant birth and death rates, Stadler [41] derived the probability density function of a phylogenetic tree under the birth-death model. Recently, time-dependent BD processes have attracted more attention as a mode of performing hypothesis-driven research [42-45]. Rabosky [42] distinguished rate-variable models of diversification from rate-constant models by fitting BD models using likelihood methods. Hohna [44, 46] and Hallinan [45] studied the reconstructed process with time-dependent rates in a more general setting by relaxing the assumptions about the number of species and the time of the process. The BD model was first adopted in [47] and further extended by other researchers to reconcile gene and species trees

(Arvestad et al. [48], Akerborg et al. [23], Rasmussen and Kellis [49] and Sjostrand et al.[50]). Recently, Boussau et al. [51] established a BD phylogenetic model for co-estimating gene and species trees without the need of estimation of divergence times in species trees and duplication and loss rates.

The current computational methods for analyzing gene family data (including gene duplication and loss) suffer a variety of weaknesses that need to be addressed. There is no probabilistic model embedded in a birth-death phylogenetic modeling framework that can take into account the effects of different evolutionary mechanisms of gene retention when analyzing gene family data. It is desirable to build a stochastic model as a good approximation to the real biological process of gene duplication and loss. Such probabilistic models can both add biological realism to improve the fit of the model to the data as well as enable mechanistic inference that is currently not possible. In this study, we integrate several evolutionary mechanisms of gene retention into the age-dependent BD model [42-45], in which the loss rate is a function of the ages of gene copies. Moreover, we derive the probability density function of gene duplication times for each mechanism. The conditional density function of a duplication time given the previous duplication time is derived from the reconstructed process under the generalized birth-death model [35, 52]. The conditional density function can be utilized to calculate the joint density of duplication times, and to efficiently simulate duplication times under the generalized BD model. The simulation results suggest that the maximum likelihood approach can accurately estimate the parameters in the generalized BD model for different mechanisms of gene retention, and the proposed gene-retention model can be used to detect the underlying mechanism that drives the evolutionary process of duplicates within a gene family.

2.2 MODELING THE GENE FAMILY EVOLUTION

2.2.1 *Modeling the loss rate*

For simplicity, we consider the process of gene duplication/loss in a single population. For a single population, we assume that a gene copy may duplicate or die at time t. The homogeneous birth-death model assumes that the rate of loss (hazard) of a duplicated gene is constant through time [11, 53]. This expectation is consistent with the nonfunctionalization process, but does not take into account any of the processes of neofunctionalization and subfunctionalization, which can affect the loss rate of gene duplicates. Unlike the homogeneous birth-death model, our model includes a time-dependent loss rate and a constant duplication rate λ. The time-dependent loss rates will be extended to age-dependent loss rates in the age-dependent birth-death model (see section 2.3). The process starts at time 0, and the number of gene copies at time 0 is 2. The process of gene duplication and loss occurs under the following postulates [54]: (1) the probability that a duplication will occur during an infinitesimal interval (t, t+Δt] is $n_t\lambda\Delta t + o(\Delta t)$, while the probability that no duplication will occur is $1- n_t\lambda\Delta t + o(\Delta t)$, and (2) the probability that a gene duplicate will be lost during an infinitesimal interval (t, t+Δt] is $n_t\mu_t\Delta t + o(\Delta t)$, while the probability that no loss will occur is $1- n_t\mu_t \Delta t + o(\Delta t)$, in which the loss rate μ_t is a function of time t.

We introduce three formulas for the loss rate μ_t based on the processes of nonfunctionalization, neofunctionalization, and subfunctionalization, with assumptions about these processes made in the introduction and also described in [45]. For nonfunctionalization, the loss rate μ_t is constant over time t, i.e., $\mu_t = \mu$. The neofunctionalization hazard rate (instantaneous rate of duplicate copy loss) declines with time [55]. Averaging across the probability of hitting a neofunctionalizing substitution, the nonfunctionalization probability for

duplicate genes declines, leading to the overall decline of duplicate loss over long evolutionary time periods [19]. This convexly declining loss rate has been described with a Weibull hazard function to characterize the average process (the process for a single gene with a known neofunctionalization event would be a discrete jump in the hazard rate) [18]. We use an exponential function to model the loss rate of neofunctionlization, i.e., $\mu_t = \alpha e^{-t\alpha}$ for $0 < \alpha < 1$. Further, the subfunctionalization loss rate behavior has been characterized to be concavely (sigmoidally) declining based upon theoretical expectations of a waiting time for complementary mutations [18, 20]. Konrad [15] introduced an extended exponential hazard function to describe the instantaneous rate of loss. We adopt a generalized logistic function for the loss rate μ_t of subfunctionalization, i.e., $\mu_t = \frac{\alpha e^{\gamma - t}}{1+e^{\gamma - t}}$, in which the scale parameter $0 < \alpha < 1$ and known location parameter $\gamma > 0$.

2.2.2 *The time-dependent birth-death model*

We are interested in the probability distribution of duplication times of the reconstructed lineages (the lineages that have survived to the present time), because the phylogeny reconstructed from the sequences of contemporary species does not include the extinct lineages [35]. The pure birth process of the reconstructed lineages can be derived from a generalized birth-death process [34, 36]. We use the following notations which are defined closely to Nee et al. [35] throughout this paper. Let $t_2 = 0$ be the first duplication time at the root of the tree, and T be the present time (we are looking forward in time, i.e., $T > 0$). Let n_T be the number of lineages at the present time T. Let n_i be the number of reconstructed lineages alive at t_i that survive to the present. We use $\{t_i \mid i = 2, \ldots, n_T\}$ to denote the duplication times of n_T lineages at the tips of a phylogenetic tree, and $t_2 < t_3 < t_4 < \ldots < T$. Let $P(\tau, T)$ be the probability that one lineage at time τ leaves multiple descendants at the present time T, i.e., $P(\tau, T) = P(n_T > 0 \mid n_\tau = 1)$ [34-36, 44],

$$P(\tau,T) = \left[1 + \int_{\tau}^{T} \mu_t e^{\rho(\tau,t)} dt\right]^{-1}. \qquad (2.1)$$

In Equation 2.1, $\rho(\tau,T) = \int_{\tau}^{T} (\mu_s - \lambda) ds$. Since the integral $\int_{\tau}^{T} \mu_t e^{\rho(\tau,t)}$ is analytically intractable, it is approximated by a Monte Carlo method. We define u_{ij} as the probability $P(n_j > 1 \mid n_i = 1)$ that one lineage at time t_i leaves multiple descendant reconstructed lineages at a later time t_j. This probability has been derived under the time-dependent BD model, i.e., $u_{ij} = P(n_j > 1 | n_i = 1) = 1 - P(t_i, t_j) e^{\rho(t_i,t_j)}$ (see Eq. (8) in [45]). Given the number n_T of lineages at the present time T and the number n_0 of lineages at time 0, the probability density function of the duplication times $t = \{t_i \mid i = n_0+1, \ldots, n_T\}$ is given by [45]

$$f(t|n_T, n_0, T) = \frac{\prod_{i=n_0+1}^{n_T} (i-1)\lambda P(t_i,T)\left(1-\eta_{t_{i-1},t_i}\right)^{i-1}}{\binom{n_T-1}{n_0-1}(1-\eta_{0,T})^{n_0}\eta_{0,T}^{n_T-n_0}}. \quad (2.2)$$

In (2.2), $\eta_{ij} = 1 - \frac{1-u_{iT}}{1-u_{jT}}$. The conditional probability distribution of duplication time t_i ($i > 2$), given its previous duplication time t_{i-1}, T and n_T, is given by [45]

$$f(t_i|t_{i-1}, n_T, T) = \frac{f(t_i|t_{i-1})P(n_T|n_{t_i},T)}{P(n_T|n_{t_{i-1}},T)} \quad (2.3)$$

In Equation 2.3, $f(t_i|t_{i-1}) = (i-1)\lambda P(t_i,T)\left(1-\eta_{t_{i-1},t_i}\right)^{i-1}$ (See Equations 19 and 23 in [45]). With the conditional densities $f(t_i|t_{i-1}, n_T, T)$ of duplication times, the duplication events between times 0 and T can be simulated recursively in forward direction. The conditional density in (3) differs from the density of duplication times derived by Hohna [44], in which the duplication events are treated as a random sample from a common probability distribution.

2.2.3 *The age-dependent birth-death model*

The time-dependent birth-death model described in the previous section starts with a single population corresponding to the lineage of a phylogenetic tree and assumes a molecular clock that starts ticking for all duplicates at the root. Thus, in the time-dependent birth-death model, the

loss rate μ_t of a gene copy is a function of time *t*. However, the loss rate μ_t should be a function of the ages of gene copies. In this section, the time-dependent birth-death process is extended to the age-dependent process, where the clock for each duplicate starts ticking at its birth. When the loss rate is constant (i.e., nonfunctionalization), the age-dependent model is identical with the time-dependent model. Thus, we only describe the age-dependent model for neofunctionalization and subfunctionalization. In the age-dependent model, the expressions for the loss rates of neofunctionalization and subfunctionalization remain unchanged (see section 2.2.1), except that time *t* is replaced with the age *t'* of the gene copy, i.e.,$\mu_{t'} = \alpha e^{-t'\alpha}$ for neofunctionalization and $\mu_{t'} = \frac{\alpha e^{\gamma - t'}}{1+e^{\gamma - t'}}$ for subfunctionalization. Moreover, it is assumed that the number of gene copies increases or decreases by 1 or remains the same during an infinitesimal interval (t, t+Δt) with probabilities described in (2.4a-2.4c)

$$P(n_{t+\Delta t} = n_t + 1) = n_t \lambda \Delta t + o(\Delta \mathrm{t}) \qquad (2.4a)$$

$$P(n_{t+\Delta t} = n_t - 1) = \sum_{i=1}^{n_t} \mu_{t_i'} \Delta t + o(\Delta \mathrm{t}) \qquad (2.4b)$$

$$P(n_{t+\Delta t} = n_t) = 1 - (n_t \lambda + \sum_{i=1}^{n_t} \mu_{t_i'}) \Delta t + o(\Delta \mathrm{t}) \quad (2.4c)$$

In (2.4b), $\mu_{t_i'}$ is the loss rate of gene copy *i* at the age of t_i' for *i* = 1, 2, …, n_t. Let t_i^0 be the duplication time of gene copy *i*. The age t_i' of gene copy *i* is a random variable, because it is a function of the random duplication time t_i^0 , i.e., $t_i' = t - t_i^0$. Therefore, (2.4b) and (2.4c) are integrated over all possible values of $\mu_{t_i'}$ with respect to the probability density function $f(t')$ of the age t' of a gene copy. The age-dependent loss rate $\mu_{t_i'}$ in (2.4b) and (2.4c) is replaced with its expectation $E(\mu_{t_i'})$. Since all t_i's have the same probability distribution, the loss rates of n_t gene

copies have the same expected values. Let t^0 be the most recent duplication time of a gene copy that survives to time t. Since t^0 is the most recent duplication time, it indicates that no duplication or loss events have occurred between t^0 and t on the gene copy. It has been shown that the number of duplication or loss events follows the Poisson distribution with mean $\int_0^t(\lambda+\mu_x)dx$. The probability of no duplication or loss events occurring within the time interval [0, t] is equal to $e^{-\int_0^t(\lambda+\mu_x)dx}$. Thus, the probability density of duplication time t^0 is proportional to $D_{t^0}e^{-\int_0^t(\lambda+\mu_x)dx}$ for $0 < t^0 < t$, in which D_{t^0} is the duplication rate at time t^0 and $e^{-\int_0^t(\lambda+\mu_x)dx}$ is the probability that t^0 is the most recent duplication time of the gene copy. Given that duplication occurs on a specific lineage, D_{t^0} is equal to the duplication rate λ. Thus, the probability density of the most recent duplication time t^0 is

$$f(t^0) = \frac{e^{-\int_{t^0}^t(\lambda+\mu_x)dx}}{\int_0^t\left(e^{-\int_{t^0}^t(\lambda+\mu_x)dx}\right)dt^0} \qquad (2.5)$$

Because the gene age t' is equal to $t - t^0$, the probability density of age t' for $0 < t' < t$ is given by

$$f(t') = \frac{e^{-\int_{t-t'}^t(\lambda+\mu_x)dx}}{\int_0^t\left(e^{-\int_{t-t'}^t(\lambda+\mu_x)dx}\right)dt'} \qquad (2.6)$$

Since the denominator in Equation 2.6 is intractable, it is approximated by Monte Carlo simulation. It follows that the mean loss rate at time t is $\phi_t = E\left(\mu_{t_i'}\right) = \int_0^t \mu_{t'}f(t')dt'$. The exact calculation of mean loss rate is shown in Appendix 1. Thus, the postulates in (2.4b) and (2.4c) become $P(n_{t+\Delta t} = n_t - 1) = n\phi_t\Delta\mathrm{t} + o(\Delta\mathrm{t})$ and $P(n_{t+\Delta t} = n_t) = 1 - n_t(\lambda+\phi_t)\Delta t + o(\Delta\mathrm{t})$. The loss rate in Equation 2.1 is replaced by the mean loss rate ϕ_t accordingly and $P(\tau, T)$ is modified as

$$P(\tau, T) = \left[1 + \int_{\tau}^{T} \phi_t e^{\rho(\tau,t)} dt\right]^{-1} \qquad (2.7)$$

Finally, the joint and conditional probability density of duplication times (in Equations 2.2 and 2.3) for the age dependent model remain unchanged, except that the loss rate μ_t in Equations 2.2 and 2.3 is replaced with the mean loss rate ϕ_t.

2.3 APPLICATIONS

2.3.1 *Simulation for the time-dependent model*

To evaluate the performance of the time-dependent birth-death model on simulated data where the true values of parameters are known, we generated duplication times of gene copies using the rejection-sampling algorithm with the conditional probability density function of duplication times in Equation 2.3, We found the maximum likelihood score for the conditional probability distribution using an optimization function *optim* in R. The maximum score was used as the upper bound in the rejection-sampling algorithm. Specifically, duplication times were simulated from Equation 2.3 with a fixed current time *T= 10* and a fixed number of gene copies n_T = 32 at time *T*. The first duplication time is set to 0, i.e., t_2 = 0; the second one is simulated conditional on the first one and so on so that additional 30 duplication times are generated sequentially. Duplication events were generated under each of 3 duplication mechanisms (nonfunctionalization, neofunctionalization, and subfunctionalization) with different parameterizations specified in Table 2.1. We set a constant duplication rate $\lambda = 0.2$ for all simulations (Table 2.1). The loss rates were determined by the equations described previously for nonfunctionalization, neofunctionalization, and subfunctionalization models with parameters shown in Table 2.1. The values of parameters were selected such that three mechanisms have the same starting deletion rate.

For each mechanism, simulation was repeated 100 times. The mean of simulated duplication times for each of three mechanisms are shown in Figure 2.1a. Duplication times simulated under different mechanisms show distinct patterns. Given the present time T and the number of gene copies n_T, the overall duplication times for nonfunctionalization tend to be larger than those for neofunctionalization and subfunctionalization, and duplication times for neofunctionalization appear to be smaller than subfunctionalization. The curves of duplication times for nonfunctionalization, neofunctionalization, and subfunctionalization are well separated (Figure 2.1a), even though three mechanisms have the same duplication rate and the same starting deletion rate. These results indicate that duplication times can be used to distinguish different mechanisms of gene retention, and to make inference about the underlying mechanism that generated the observed duplication times given the assumptions of the duplication models and their relationship to the underlying biology. These results are consistent with the caveat that the time-dependent process uses a tree-dependent clock rather than the more biological situation of a duplication-event specific process. The extension to the age-dependent birth-death model will be discussed below. The joint probability density function in Equation 2.2 can be used to obtain the maximum likelihood estimates (MLE) of parameters in the time-dependent model, when duplication times are given as input data. To visualize the divergence of the probability density functions of three mechanisms, we plotted the density curves of the first duplication time for nonfunctionalization, neofunctionalization, and subfunctionalization (Figure 2.1b) with the values of parameters in Table 2.1. Since each mechanism has a unique density curve, this result indicates that it is possible to distinguish three mechanisms using the time-dependent birth-death model. Moreover, we employed the Akaike Information Criterion (AIC) [56] to evaluate the relative quality of the time-dependent models for nonfunctionalization, neofunctionalization, and

subfunctionalization. The data sets simulated from the time-dependent model were used as input data to calculate AIC for nonfunctionalization, neofunctionalization, and subfunctionalization. For each simulated data set, the mechanism with the lowest AIC score was selected and compared with the true mechanism from which the data sets were generated. We reported the percentage of the simulated data sets successfully identifying the true mechanism (Figure 2.1c). The overall average of the percentages of samples recovering the true mechanism is about 80% (Figure 2.1c). In addition, subfunctionalization appears to be more difficult than neofunctionalization to distinguish from nonfunctionalization in this modeling framework (Figure 2.1c).

To examine the performance of maximum likelihood estimation, we use the simulated duplication times as data to estimate model parameters. The sample size (the number of duplication times) ranges from 20 to 100. The maximum likelihood estimates of parameters were obtained using Metropolis algorithm. The standard errors of the maximum likelihood estimates are displayed in Figure 2.2. For nonfunctionalization, the standard errors of the estimates of μ and λ decrease as the number of duplication times increases from 20 to 100. Similarly, the standard errors of the estimates of parameters for subfunctionalization and neofunctionalization decrease as the number of duplications grows. However, the estimation of parameter α for neofunctionalization does not improve well with the increased number of gene copies (Figure 2.2), because duplication times in the simulated data are highly correlated and the auto-correlation between two adjacent duplication times increases as the number of duplication times increases. As a result, when the number of highly correlated duplication times reaches a certain number, adding even more duplication times does not contribute more information for accurately estimating model parameters, especially for neofunctionalization where the loss rate quickly

declines to a very low level. Nevertheless, these results suggest that maximum likelihood methods can accurately estimate parameters in the time-dependent birth-death model when the sample size is large.

2.3.2 *Simulation for the age-dependent birth-death model*

The simulation for the age-dependent model was conducted with the same parameterization and simulation procedure used for the time-dependent model. We generated duplication times from the age-dependent models for nonfunctionalization, neofunctionalization, and subfunctionalization. The mean duplication times given the current time T and gene copy number n_T for the age-dependent models (Figure 2.3a) appear to be less diversified among nonfunctionalization, neofunctionalization, and subfunctionalization than those for the time-dependent models (Figure 2.1a). In addition, the density curve for subfunctionalization becomes closer to the nonfunctionalization curve under the age-dependent model (Figure 2.3b), compared to the curves for the time-dependent model (Figure 2.1b). This is consistent with our expectation, because the age of a gene copy is less than the absolute time t. Since the loss rate of subfunctionalization is concavely declining, the beginning portion of the loss rate of subfunctionalization is similar to the constant rate of nonfunctionalization. In Figure 2.3b, the density curve for neofunctionalization is well separated from the density curves for nonfunctionalization and subfunctionalization. However, the density curves for nonfunctionalization and subfunctionalization are almost identical, indicating that for the age-dependent model, it could be very difficult to distinguish subfunctionalization and nonfunctionalization (Figure 2.3b). The overall percentage of samples identifying the true mechanism increases as the number of gene copies grows (Figure 2.3c). The percentages of nonfunctionalization and neofunctionalization are significantly higher than the overall

percentage. Although the performance of subfunctionalization is below average, the percentage of samples successfully identifying the true subfunctionalization increases to 60% when the number of gene copies reaches 100. Moreover, the standard errors of the estimates of parameters in the age-dependent model appear to decrease as the number of gene copies grows, suggesting that maximum likelihood methods can accurately estimate parameters in the age-dependent model, when the sample size is large (Figure 2.4).

2.4 DISCUSSION

2.4.1 *Summary of the gene family evolution model*

We have derived the probability density function for the age-dependent birth-death model, in which the loss rate is a function of the ages of gene copies. In addition, the conditional density function and a joint density function of duplication times with age-dependent loss rate have been developed in above age-dependent model, given the current time T and the number of gene copies at the time T. The conditional density function is used to efficiently simulate duplication times, and the simulation results suggest that maximum likelihood methods can accurately estimate model parameters in both time-dependent and age-dependent models. In addition, the relative quality of various birth-death models was assessed with AIC. Both time-dependent and age-dependent models can distinguish the three mechanisms (nonfunctionalization, neofunctionalization, and subfunctionalization) with high probabilities when the sample size is large. These results indicate that the probabilistic models derived from the birth-death process with a time-dependent and age-dependent loss rates are useful for understanding the duplication and loss mechanisms of gene families that evolve over time in a single population with caveats discussed.

2.4.2 Limitations and future study

As duplication times are often not observable, it is desirable to generalize the current model to DNA sequence data. We are currently working along this line to build a generalized model that includes two stochastic processes. The birth and death process is used to derive the probability distribution of a gene family tree, while the mutation process is used to derive the probability distribution of DNA sequence data given the gene family tree. With this generalized model, we can estimate model parameters (duplication and loss rates) from DNA sequence data.

One of the limits of the current model is that it considers gene family evolution in a single population. This model cannot be applied as currently implemented to understand the evolutionary process of gene families from multiple species. To overcome this limit, the current model will be extended in the context of species trees, in which duplication process occurs along the lineages of species trees. This generalization will certainly involve intensive computation, but such a model is quite useful for understanding gene family evolution in the context of the evolution of species. Another limitation of the current age-dependent model is that the likelihood is conditioned on observed extant duplicate copies and does not consider the full generative process including duplicates that were lost before the present. Future work will examine this in the context of Approximate Bayesian Computation [57]. Further, the current model exists in the classes of interspecific models that treat all observations from a single individual from a species as fixed relative to observations from single individuals from other species. Recently, a correction for the effects of population dynamics has been introduced and can be considered in modeling efforts [9]. Missing data and genome assembly error are also not specifically addressed in the modeling framework and their impact on inference also needs to be addressed.

The gene loss models and their interpretations (the relationship between the best fit curve shape and the underlying biology) make assumptions about the relationship between the accumulation of synonymous changes and of non-synonymous changes whereas there is information in the evolution of dN/dS vs. dS that can be taken advantage of in alternative formulations of the likelihood (see [18]). Lastly, the models can be used to make predictions about functional evolution in the absence of actual functional data. While such data does not currently exist in large scale, the future may bring data on the expression levels of protein duplicates compared to an ancestral state as well as binding and enzyme specificities (and enzyme kinetics), all of which can be integrated into a phylogenetic framework. However, even with future comparative proteomic data, one still needs models that treat signals associated with selective pressures (like the models presented here), as neutral changes in expression and functional properties would not lead to changes in retention profiles (the gene loss hazard/survival model) and meaningful lineage-specific biology (see [58] for a discussion of the interplay between molecular phenotypes and biological function in an evolutionary context).

The model as currently developed also assumes that all duplicates in a gene family evolve under the same process. A future opportunity is in examination of large gene family databases like Ensembl [59], HOGENOM [60], or TAED [61], a mixture model of duplicate processes can be applied across all gene families and duplication events to enable a posteriori probabilistic identification of duplication retention mechanisms for individual gene duplication events. The work presented in this manuscript, with a birth-death model in a phylogenetic context, brings this scale of modeling one step closer.

2.5 CONCLUSIONS

We develop a generalized birth-death model for probabilistic inference of the evolutionary mechanism for duplicate gene retention based upon the average rate of loss over time of the duplicate. The time-dependent birth-death model assumes a molecular clock that starts ticking for all duplicates at the root. The time-dependent model is then extended to the age-dependent model, which allows the gene loss rate dependent of duplication events. Simulation results indicate that the mechanisms of gene retentions (nonfunctionalization, neofunctionalization, and subfunctionalization) produce distinct likelihood functions, which can be used with comparative genomic data to quantitatively distinguish those mechanisms.

2.6 REFERENCES

1. Ohta, T., *Simulating evolution by gene duplication.* Genetics, 1987. **115**(1): p. 207-13.
2. Fortna, A., et al., *Lineage-specific gene duplication and loss in human and great ape evolution.* PLoS Biol, 2004. **2**(7): p. E207.
3. Nei, M. and A.P. Rooney, *Concerted and birth-and-death evolution of multigene families.* Annu Rev Genet, 2005. **39**: p. 121-52.
4. Lynch, M., et al., *The probability of preservation of a newly arisen gene duplicate.* Genetics, 2001. **159**(4): p. 1789-804.
5. Hurles, M., *Gene duplication: the genomic trade in spare parts.* PLoS Biol, 2004. **2**(7): p. E206.
6. Ohta, T., *Role of gene duplication in evolution.* Genome, 1989. **31**(1): p. 304-10.
7. Zhang, J.Z., *Evolution by gene duplication: an update.* Trends in Ecology & Evolution, 2003. **18**(6): p. 292-298.

8. Lynch, M., *Genomics. Gene duplication and evolution.* Science, 2002. **297**(5583): p. 945-7.

9. Teufel, A.I., J. Masel, and D.A. Liberles, *What Fraction of Duplicates Observed in Recently Sequenced Genomes Is Segregating and Destined to Fail to Fix?* Genome Biol Evol, 2015.

10. Hahn, M.W., et al., *Estimating the tempo and mode of gene family evolution from comparative genomic data.* Genome Res, 2005. **15**(8): p. 1153-60.

11. Lynch, M. and J.S. Conery, *The evolutionary fate and consequences of duplicate genes.* Science, 2000. **290**(5494): p. 1151-5.

12. Hughes, A.L. and R. Friedman, *Gene duplication and the properties of biological networks.* J Mol Evol, 2005. **61**(6): p. 758-64.

13. Liberles, D.A. and K. Dittmar, *Characterizing gene family evolution.* Biol Proced Online, 2008. **10**: p. 66-73.

14. Innan, H. and F. Kondrashov, *The evolution of gene duplications: classifying and distinguishing between models.* Nat Rev Genet, 2010. **11**(2): p. 97-108.

15. Konrad, A., et al., *Toward a general model for the evolutionary dynamics of gene duplicates.* Genome Biol Evol, 2011. **3**: p. 1197-209.

16. Ohno, S., *Evolution by gene duplication*. 1970, New York: Springer.

17. Pollock, D.D., G. Thiltgen, and R.A. Goldstein, *Amino acid coevolution induces an evolutionary Stokes shift.* Proc Natl Acad Sci U S A, 2012. **109**(21): p. E1352-9.

18. Hughes, T. and D.A. Liberles, *The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation.* J Mol Evol, 2007. **65**(5): p. 574-88.

19. Teufel, A.I., et al., *On Mechanistic Modeling of Gene Content Evolution: Birth-Death Models and Mechanisms of Gene Birth and Gene Retention.* Computation, 2014. **2(3),**.

20. Force, A., et al., *Preservation of duplicate genes by complementary, degenerative mutations.* Genetics, 1999. **151**(4): p. 1531-45.

21. Rastogi, S. and D.A. Liberles, *Subfunctionalization of duplicated genes as a transition state to neofunctionalization.* BMC Evol Biol, 2005. **5**: p. 28.

22. Khan, A.A., et al., *Phylogenetic analysis of kindlins suggests subfunctionalization of an ancestral unduplicated kindlin into three paralogs in vertebrates.* Evol Bioinform Online, 2011. **7**: p. 7-19.

23. Akerborg, O., et al., *Simultaneous Bayesian gene tree reconstruction and reconciliation analysis.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(14): p. 5714-5719.

24. Basten, C.J. and T. Ohta, *Simulation study of a multigene family, with special reference to the evolution of compensatory advantageous mutations.* Genetics, 1992. **132**(1): p. 247-52.

25. Hahn, M.W., J.P. Demuth, and S.G. Han, *Accelerated rate of gene gain and loss in primates.* Genetics, 2007. **177**(3): p. 1941-9.

26. Ohta, T., *An Extension of a Model for the Evolution of Multigene Families by Unequal Crossing over.* Genetics, 1979. **91**(3): p. 591-607.

27. Thornton, J.W. and R. DeSalle, *Gene family evolution and homology: genomics meets phylogenetics.* Annu Rev Genomics Hum Genet, 2000. **1**: p. 41-73.

28. Yanai, I., C.J. Camacho, and C. DeLisi, *Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification.* Phys Rev Lett, 2000. **85**(12): p. 2641-4.

29. Karev, G.P., et al., *Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models.* BMC Evol Biol, 2004. **4**: p. 32.

30. Bailey, N., *The elements of stochastic processes*. 1964, New York.

31. Huynen, M.A. and E. van Nimwegen, *The frequency distribution of gene family sizes in complete genomes.* Mol Biol Evol, 1998. **15**(5): p. 583-9.

32. Csuros, M. and I. Miklos, *Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model.* Mol Biol Evol, 2009. **26**(9): p. 2087-95.

33. Szollosi, G.J., et al., *The inference of gene trees with species trees.* Syst Biol, 2015. **64**(1): p. e42-62.

34. Thompson, *The likelihood approach.* Human evolutionary trees, 1975.

35. Nee, S., R.M. May, and P.H. Harvey, *The reconstructed evolutionary process.* Philos Trans R Soc Lond B Biol Sci, 1994. **344**(1309): p. 305-11.

36. Kendall, D.G., *On the Generalized Birth-and-Death Process.* Annals of Mathematical Statistics, 1948. **19**(1): p. 1-15.

37. Rannala, B. and Z. Yang, *Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference.* J Mol Evol, 1996. **43**(3): p. 304-11.

38. Aldous, D. and L. Popovic, *A critical branching process model for biodiversity.* Advances in Applied Probability, 2005. **37**(4): p. 1094-1115.

39. Gernhard, T., *The conditioned reconstructed process.* Journal of Theoretical Biology, 2008. **253**(4): p. 769-778.

40. Gernhard, T., *New Analytic Results for Speciation Times in Neutral Models.* Bulletin of Mathematical Biology, 2008. **70**(4): p. 1082-1097.

41. Stadler, T., *Sampling-through-time in birth-death trees.* J Theor Biol, 2010. **267**(3): p. 396-404.

42. Rabosky, D.L., *Likelihood methods for detecting temporal shifts in diversification rates.* Evolution, 2006. **60**(6): p. 1152-64.

43. Morlon, H., T.L. Parsons, and J.B. Plotkin, *Reconciling molecular phylogenies with the fossil record.* Proc Natl Acad Sci U S A, 2011. **108**(39): p. 16327-32.

44. Hohna, S., *Fast simulation of reconstructed phylogenies under global time-dependent birth-death processes.* Bioinformatics, 2013. **29**(11): p. 1367-74.

45. Hallinan, N., *The generalized time variable reconstructed birth-death process.* J Theor Biol, 2012. **300**: p. 265-76.

46. Hohna, S., *The time-dependent reconstructed evolutionary process with a key-role for mass-extinction events.* J Theor Biol, 2015. **380**: p. 321-31.

47. Arvestad, L., et al., *Bayesian gene/species tree reconciliation and orthology analysis using MCMC.* Bioinformatics, 2003. **19 Suppl 1**: p. i7-15.

48. Arvestad, L., J. Lagergren, and B. Sennblad, *The gene evolution model and computing its associated probabilities.* J. ACM, 2009. **56**(2): p. 1-44.

49. Rasmussen, M.D. and M. Kellis, *A Bayesian approach for fast and accurate gene tree reconstruction.* Mol Biol Evol, 2011. **28**(1): p. 273-90.

50. Sjostrand, J., et al., *DLRS: gene tree evolution in light of a species tree.* Bioinformatics, 2012. **28**(22): p. 2994-5.

51. Boussau, B., et al., *Genome-scale coestimation of species and gene trees.* Genome Res, 2013. **23**(2): p. 323-30.

52. Liu, L., et al., *A Bayesian model for gene family evolution.* BMC Bioinformatics, 2011. **12**: p. 426.

53. Cotton, J.A. and R.D. Page, *Rates and patterns of gene duplication and loss in the human genome.* Proc Biol Sci, 2005. **272**(1560): p. 277-83.

54. Feller, W., *An introduction to probability theory and its applications*. 1954, New York: Wiley.

55. Zhang, P., W. Min, and W.H. Li, *Different age distribution patterns of human, nematode, and Arabidopsis duplicate genes.* Gene, 2004. **342**(2): p. 263-8.

56. Akaike, H., *Information theory and an extension of the maximum likelihood principle.* in Petrov, B.N.; Csáki, F., 2nd International Symposium on Information Theory, 1973. **September 2-8**(Budapest: Akadémiai Kiadó): p. p. 267-281.

57. Janzen, T., S. Höhna, and R.S. Etienne, *Approximate Bayesian Computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nLTT.* Methods in Ecology and Evolution, 2015. **6**(5).

58. Graur, D., et al., *On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE.* Genome Biol Evol, 2013. **5**(3): p. 578-90.

59. Flicek, P., et al., *Ensembl 2012.* Nucleic Acids Res, 2012. **40**(Database issue): p. D84-90.

60. Penel, S., et al., *Databases of homologous gene families for comparative genomics.* BMC Bioinformatics, 2009. **10 Suppl 6**: p. S3.

61. Roth, C., et al., *The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics.* Nucleic Acids Res, 2005. **33**(Database issue): p. D495-7.

2.7 APPENDICES

In the following sections, the equation numbers without the prefix, A, correspond to those in the main article.

Appendix. 1 *Calculation of mean loss rate under age-dependent model*

With the probability density function of age t', which is

$$f(t') = \frac{e^{-\int_{t-t'}^{t}(\lambda+\mu_x)dx}}{\int_0^t\left(e^{-\int_{t-t'}^{t}(\lambda+\mu_x)dx}\right)dt'} \qquad \text{(A.6)}$$

We can calculate the mean loss rate at time t using $\phi_t = E\left(\mu_{t_i'}\right) = \int_0^t \mu_{t'} f(t')dt'$ under neofunctionalization and subfunctionalization accordingly.

First let $Nu = e^{-\int_{t-t'}^{t}(\lambda+\mu_x)dx}$ and $De = \int_0^t \left(e^{-\int_{t-t'}^{t}(\lambda+\mu_x)dx}\right)dt'$.

1. Neo functionalization

Loss rate: $\mu_t = \alpha e^{-t\alpha}$

Then

$$Nu = e^{-\int_{t-t'}^{t}(\lambda+\alpha e^{-x\alpha})dx} = e^{-\lambda t' - e^{-\alpha(t-t')} + e^{-\alpha t}}$$

$$De = \int_0^t \left(e^{-\int_{t-t'}^{t}(\lambda+\alpha e^{-x\alpha})dx}\right)dt' = \int_0^t e^{-\lambda t' - e^{-\alpha(t-t')} + e^{-\alpha t}}dt' = e^{e^{-\alpha t}}\int_0^t e^{-\lambda t'} e^{e^{-\alpha(t'-t)}}dt'$$

Thus $\Phi_t = E(\mu_{t'}) = \frac{e^{e^{-\alpha t}}\int_0^t \alpha e^{-\alpha t'} e^{-\lambda t' - e^{-\alpha(t-t')}}dt'}{e^{e^{-\alpha t}}\int_0^t e^{-\lambda t'} e^{e^{\alpha(t'-t)}}dt'} = \frac{\alpha\int_0^t e^{t'(\lambda+\alpha)} e^{-e^{-\alpha(t-t')}}}{\int_0^t e^{-\lambda t'} e^{e^{-\alpha(t'-t)}}dt'}$

2. Subfunctionalization

Loss rate: $\mu_t = \frac{\alpha e^{\gamma - t}}{1+e^{\gamma - t}}$

Then

$$Nu = e^{-\int_{t-t'}^{t}\left(\lambda+\frac{\alpha e^{\gamma-x}}{1+e^{\gamma-x}}\right)dx} = e^{-\lambda t'-e^{-\alpha(t-t')}+e^{-\alpha t}} = e^{-\lambda t'-\alpha[-\ln(e^{\gamma-x}+1)]_{t-t'}^{t}}$$

$$De = \int_0^t (Nu)dt' = \int_0^t e^{-\lambda t'+\alpha\left[\ln\left(e^{\gamma-t}+1\right)-\ln\left(e^{\gamma-(t-t')}+1\right)\right]}dt'$$

$$= e^{\alpha\ln\left(e^{(\gamma-t)}+1\right)}\int_0^t e^{-\lambda t'-\alpha\ln\left(e^{\gamma-(t-t')}+1\right)}dt'$$

Thus $\Phi_t = E(\mu_{t'}) = \dfrac{\int_0^t \frac{\alpha e^{\gamma-t'}}{1+e^{\gamma-t'}} e^{-\lambda t'-\alpha[-\ln(e^{\gamma-x}+1)]_{t-t'}^{t}} dt'}{e^{\alpha\ln\left(e^{(\gamma-t)}+1\right)}\int_0^t e^{-\lambda t'-\alpha\ln\left(e^{\gamma-(t-t')}+1\right)}dt'}$

$$= \frac{\int_0^t \frac{\alpha e^{\gamma-t'}}{1+e^{\gamma-t'}} e^{-\lambda t'-\alpha\ln(e^{\gamma-(t-t')}+1)}dt'}{\int_0^t e^{-\lambda t'-\alpha\ln\left(e^{\gamma-(t-t')}+1\right)}dt'}$$

Although the integrals in these equations derived above still require a Monte Carlo integration, these simplified equations of the mean loss rate can increase the accuracy of the overall calculation.

Table 2.1: The values of parameters used in simulating duplication times under nonfunctionalization, neofunctionalization, and subfunctionalization.

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

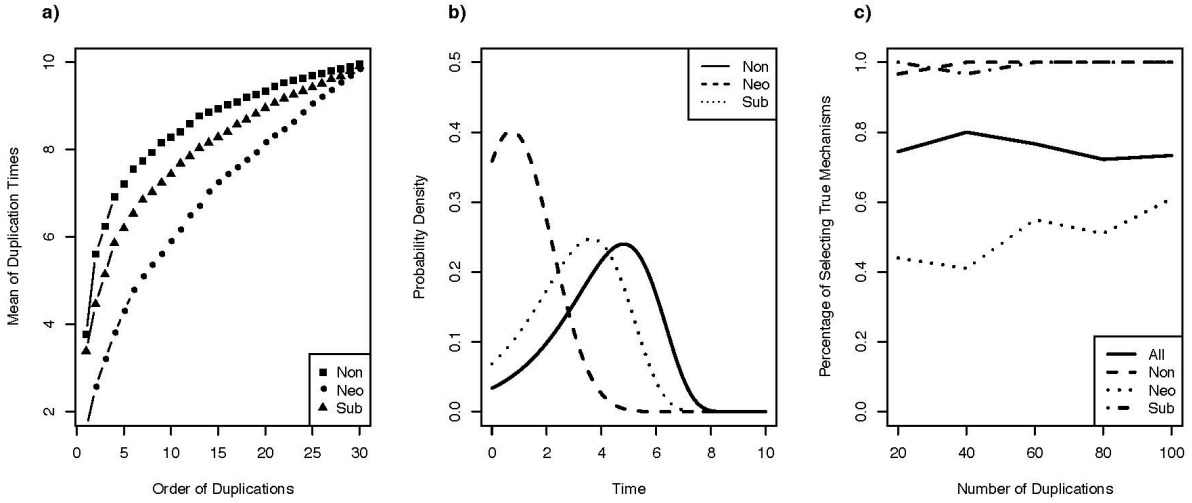

Figure 2.1: Simulation results of the time-dependent model: a) the means of duplication times simulated with 100 replicates for nonfunctionalization, neofunctionalization, and subfuncitonalization are shown; b) the probability density curves of duplication times for nonfunctionalization, neofunctionalization, and subfunctionalization under the model are shown; c) the percentage of samples identifying the true mechanism with AIC.

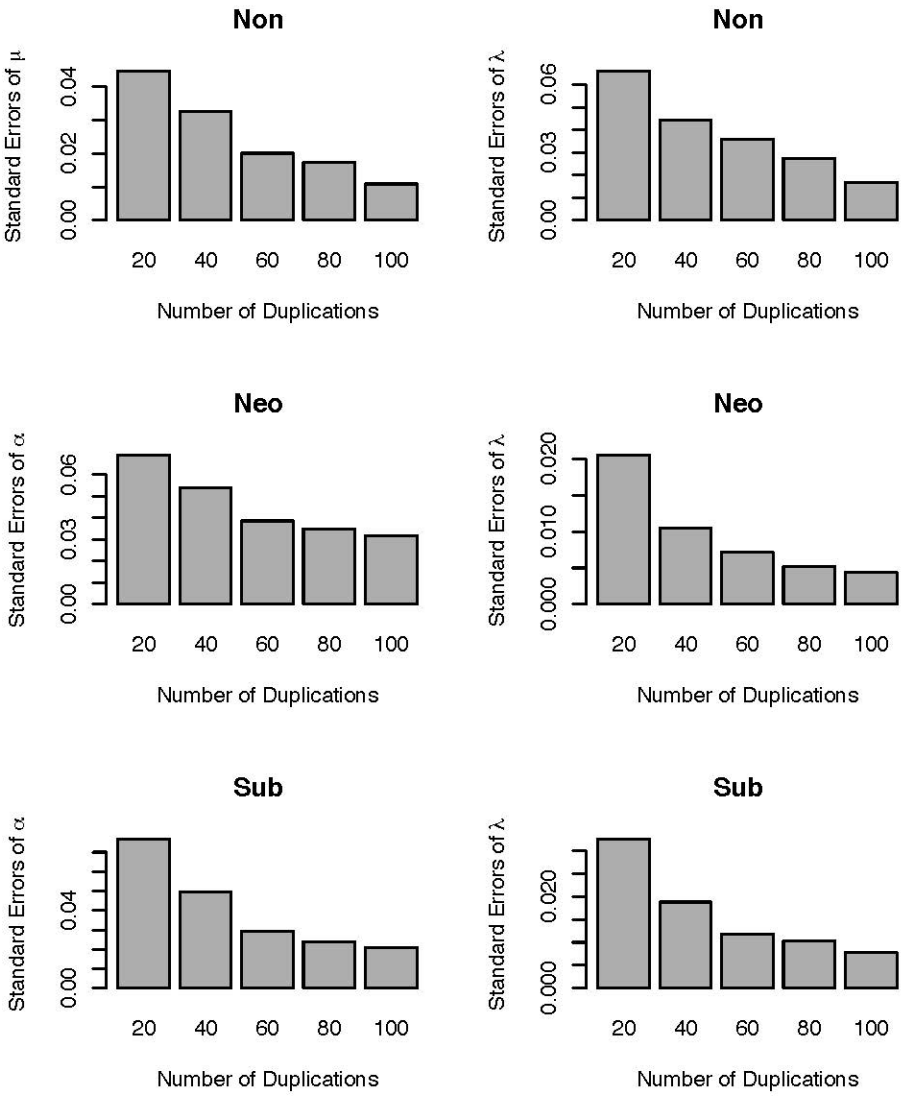

Figure 2.2: The standard errors of the maximum likelihood estimates of parameters in the age-dependent models.

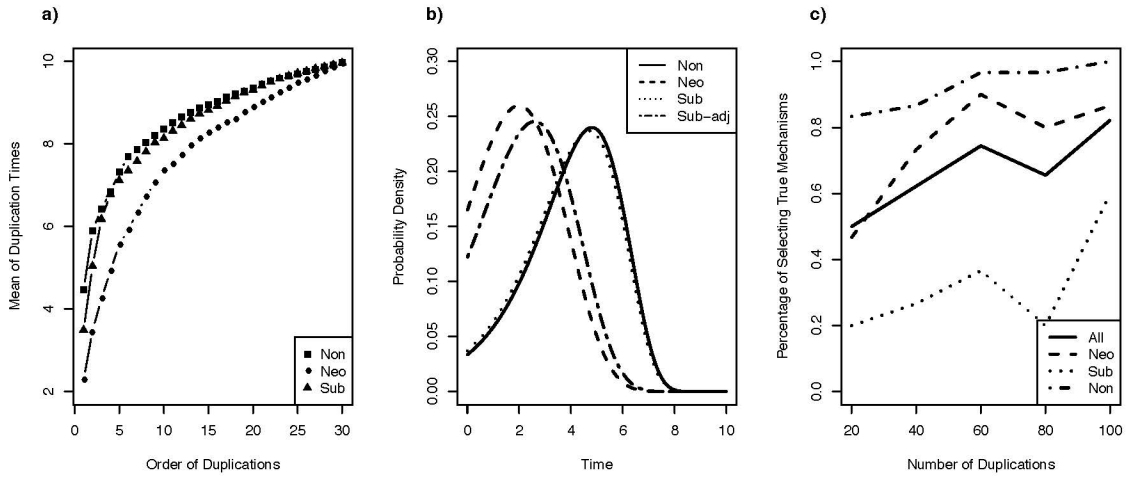

Figure 2.3: Simulation results of the age-dependent model: a) the means of duplication times simulated with 30 replicates for nonfunctionalization, neofunctionalization, and subfuncitonalization are shown; b) the probability density curves of duplication times for nonfunctionalization, neofunctionalization, and subfunctionalization under the model are shown; c) the percentage of samples identifying the true mechanism with AIC

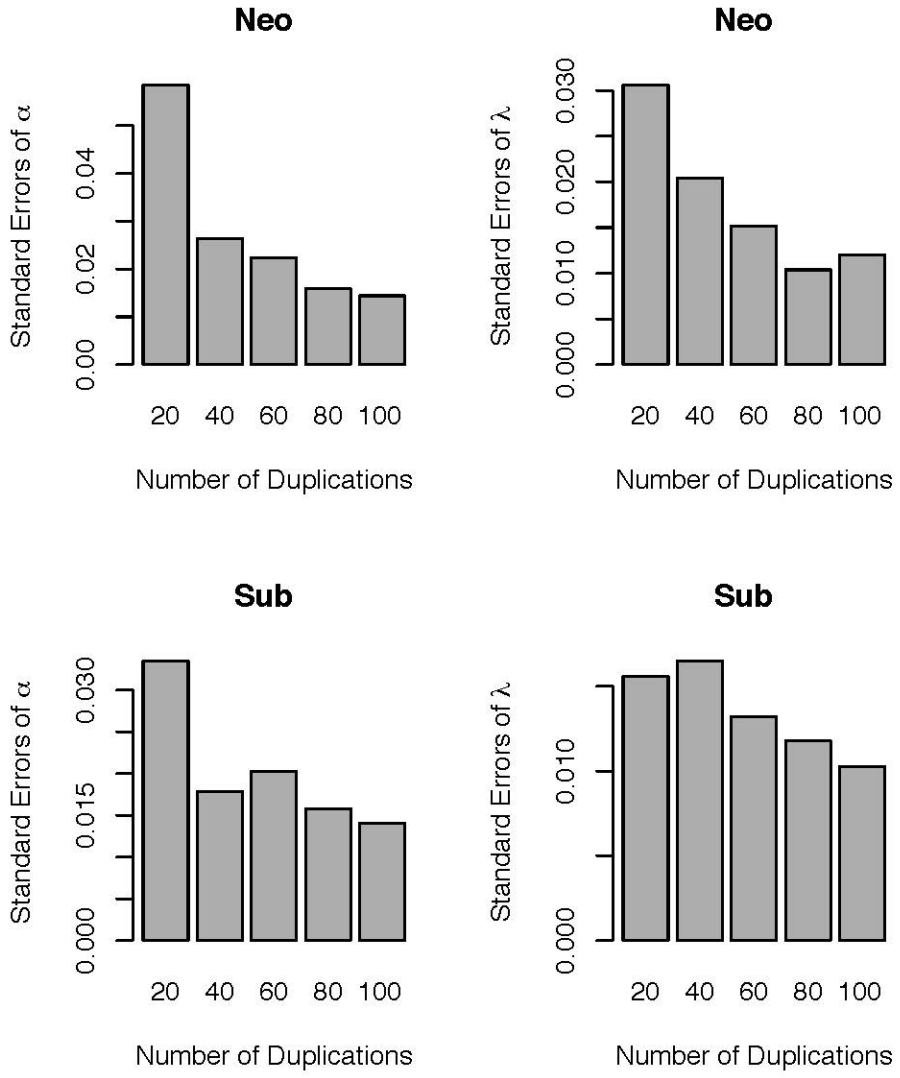

Figure 2.4: The standard errors of maximum likelihood estimates of parameters in the age-dependent models for neofunctionalization and subfunctionalization.

CHAPTER 3

A GENERALIZED BIRTH AND DEATH PROCESS FOR MODELING THE GENE FAMILY EVOLUTION WITHIN SPECIES TREE [1]

[1] Zhao, J. and Liu, L. To be submitted.

ABSTRACT

Gene families have been examined through extensive literature due to its importance in molecular evolution. Gene duplication and speciation are considered as the major forces to form gene families. In this paper, a non-homogeneous birth and death process is adopted to model gene family evolution within a species tree, in which the loss rate is a function of the gene age. In addition, specific models related to different gene retention mechanisms by incorporating distinct loss rate functions have been investigated through simulation study. Results show that different gene retention mechanisms could be distinguished by the proposed model.

3.1 INTRODUCTION

Gene family evolution is an important component in molecular evolution since it helps reveal the relationships among genes. With the acquired knowledge of these relationships, we can further explore the underlying driving forces of creation of new genes. Genes in a gene family could be paralogs due to gene duplication or orthologs because of divergence of species[1]. Duplication of genes is viewed as the dominant mechanism in developing evolutionary novelties[2-5]. Gene duplicates could be either fixed or lost in a population, and for those being fixed, it could either be silenced or be preserved with novel function from the the original copy[6].

Three different mechanisms regarding the fates of individual duplicate gene have been proposed, which are nonfunctionalization, neofunctionalization and subfunctionalization. In nonfunctionalization process, one of the gene duplicates may become a pseudogene by degenerative mutation[7, 8]. Although only a small fraction of the duplicate genes is preserved in a population, the extra gene copy may evolve a beneficial function with the original one keeping the initial function. This process is referred to as neofunctionalization, wherein the hazard function is modeled by a convexly decreasing curve [7, 9-11]. Subfunctionalization depicts a

mechanism in which each of the duplicate genes is damaged partially and both of them should be retained to perform the original functions[12-14]. The hazard function of this process , along with the non-functionalizing mutations, has been modeled by a concavely decreasing curve[7, 9-11]. Other than the three gene retention mechanisms relating to individual duplicate genes, there is an additional mechanism linked with the multiple interacting genes generated by large scale gene duplication events. This mechanism, named dosage balance, improves the preservation of duplicated interaction networks(see [7, 9, 10] for a detailed discussion).

The gene retention mechanisms discussed previously can potentially affect the gene counts in a gene family and the relationship between the gene tree and species tree. Currently, extensive computational approaches have been employed for understanding the impacts of underlying evolutionary forces on gene duplication and loss process, of which the reconciliation of gene tree with species tree and the probabilistic framework are two major strategies[1]. Goodman et al. [15, 16] firstly introduced the concept of reconciled tree in fitting gene tree into species tree using parsimony based approach which is still the most popular method in reconciliation for its computational efficiency[17]. However, there are two primary limitations of this reconciliation method when applied in modeling gene duplication and loss[1]. On one hand, the inference about gene duplication and loss would be biased if the gene tree is not estimated correctly. On the other hand, the tree reconciliation does not provide a clear interpretation about the underlying evolutionary forces since there is no parameter estimation. A second strategy is the probabilistic approach[7, 18], in which the birth and death process has been applied due to its similar underlying mechanism with gene duplication and loss. Arvestad et al. [19] introduced a gene evolution model based on a birth-death process and developed a tool to perform orthology analysis as well as gene tree/species tree reconciliation. CAFÉ [20] is a statistical analyzing tool

for the evolution of the size of gene families based on an equal rate birth and death process. Akerborg et al. [18] has developed a a probabilistic model for gene evolution by integrating gene evolution, sequence evolution, and a substitution rates model. Konrad et al.[7] have established a maximum likelihood framework based on a modified Weibull hazard function under different duplicate gene loss/retention mechanisms. It is more realistic and powerful to apply the probabilistic models over parsimony based approaches in gene family evolution. In addition, the models are possible to be compared with model assumptions[19]. However, weaknesses also exist in the probabilistic models, which are computationally expensive and highly parameterized.

In this paper, we establish a probabilistic framework to model gene family evolution in the context of species tree, in which three evolutionary mechanisms of gene retention are incorporated. This study extends the age-dependent birth and death model for gene family in one population (Zhao et al. [11]) into multiple populations by developing the joint distribution of the gene counts in the internal nodes and duplication times in all branches of the species tree. A simulation study is then performed to examine the accuracy of parameter estimation through the Bayesian methods.

3.2 MODELING GENE FAMILY EVOLUTION WITHIN A SPECIES

3.2.1 *Definitions*

Let S denote a rooted binary ultrametric K-taxon species tree with nodes $V_i(i = 1, \dots, 2K-1)$. Especially, $V_i(i = 1, \dots, K-1)$ represent the internal nodes and $V_i(i = K, \dots, 2K-1)$ denote the external nodes. We assume that the topology of the species tree ψ is given. The internal branches of the species tree represent ancestral populations, while the external branches represent contemporary species.

Let T_i be the divergence time at internal node V_i $(i = 1, \dots, K-1)$, where T_1 is the divergence time at the root of the species tree. Let T_C be the present time. We use $T_0 = 0$ to be the starting time of the evolutionary process of a gene family. Let $N_0 = 1$ be the number of gene copies at time T_0 (i.e., the process starts at a single gene copy). Let N_i $(i = 1, \dots, K-1)$ be the number of gene copies at internal node i of the species tree. Let $N_i (i = K, \dots, 2K-1)$ be the number of genes in each comtemporary species at the present time T_C (i.e., the tips of the species tree). Furthermore, we use N_i' to denote the number of gene copies at the ancestral node of V_i ($i \neq 1$, i.e., excluding the root node). Let $\boldsymbol{\tau_i} = (\boldsymbol{t_{i,N_i'+1}}, \boldsymbol{t_{i,N_i'+2}}, \dots, \boldsymbol{t_{iN_i}})$ represent the duplication times occurred in branch $\boldsymbol{i}$.

We assume that each gene may duplicate or die at any time point. As discussed in the birth and death model for the fates of gene retention [11], the duplication rate λ is constant through time and the loss rate μ_t is a function of time t. Specifically, the loss rate for nonfunctionalization is a constant, $\mu_t = \mu$, as defined in Zhao et al. [11]. Two newly defined loss rate functions of neofunctionalization and subfunctionalization are introduced to allow more flexibility than those defined by Zhao et al. [11]. A transformed exponential function, $\mu_t = \alpha e^{-t\beta}$ with $\alpha \in (0,1)$ be the starting value and $\beta > 0$ be the decreasing rate of the curve, is adopted here to model loss rate of neofunctionalization. For subfunctionalization, a transformed generalized logistic function, $\mu_t = \frac{\alpha e^{-\beta(t-\gamma)}}{1+e^{-\beta(t-\gamma)}}$ with $\alpha \in (0,1)$ be the starting value, $\beta > 0$ be the decreasing rate and $\gamma > 0$ be the time at inflection point of logistic curve, is used to model the loss rate. Moreover, we assume that when an ancestral population splits into two descendant populations, each of two descendant populations has the same number of gene copies as the ancestral population does at the divergence time.

Zhao et al. [11] have developed an age-dependent birth and death model for understanding the evolutionary process of a gene family in a single population, in which the loss rate is dependent of the age of a gene and thus substituted by the mean loss rate. In this paper, we will extend the age-dependent model to multiple populations in the context of species trees, in which the mean loss rates corresponding to neofunctionalization and subfunctionalization are calculated in Appendix 1.

3.2.2 *The probability distribution of a gene family tree within a species tree*

Given the species tree S (topology), and the number of gene copies $N_i (i = 1, \dots, 2K-1)$ at both internal and external nodes, the duplication processes in the internal and external branches are independent of one another. Let f_i be the conditional density function of the duplication times in branch i of the species tree S, conditional on the numbers of gene copies at the two ends of the branch. The conditional density function of all duplication times $\tau = \{t_{ij}, i = 1, \dots, 2K-1, j = N_i' + 1, \dots, N_i\}$, given the species tree S and the number of gene copies at the internal nodes and external nodes, is the product of density functions for individual branches, i.e.,

$$f(\tau|\psi, T_1, \dots, T_{K-1}, \theta, N_1, \dots, N_{2K-1}) = \prod_{i=1}^{2K-1} f_i \qquad (3.1)$$

In Equation 3.1, θ denotes the parameters of the age-dependent model, including the constant birth rate λ and the parameters in the death rate μ_t under three gene retention mechanisms: $\theta = (\mu, \lambda)$ for nonfunctionalization, $\theta = (\alpha, \beta, \lambda)$ for neofunctionalization, and $\theta = (\alpha, \beta, \gamma, \lambda)$ for subfunctionalization.

The density f_i can be derived from the age-dependent model. Consider an arbitrary branch (branch i) of the species tree. The time points of the two ends of branch i are denoted as $t_{iN_i'}$ and t_{iN_i}. Suppose that N_i' lineages enter the population i from its ancestral population, i.e., there are N_i' lineages at time $t_{iN_i'}$ in population i. Let N_i be the number of lineages in the

population i at time t_{iN_i}. It indicates that $N_i - N_i'$ duplication events have occurred in population i. We use $\tau_i = \{t_{ij}, j = N_i' + 1, \dots, N_i\}$ to denote the duplication times in population i. According to the age-dependent model [11], the joint probability density function of duplication times τ_i in branch i, conditional on the numbers of gene copies at the two ends of the branch, is given by

$$f_i = f(\tau_i | N_i', N_i, T_C) = \frac{\prod_{j=N_i'+1}^{N_i} (j-1)\lambda P_{t_{ij}T_C} \left(1 - \eta_{t_{i,j-1},t_{i,j}}\right)^{l-1}}{\binom{N_i-1}{N_i'-1} \left(1 - \eta_{0,T_C}\right)^{N_i'} \eta^{N_i - N_i'}} \quad (3.2)$$

In Equation 3.2, P_{t_i,t_j} and η_{t_i,t_j} are defined the same as those in Chapter 2 (Equations 2.1 and 2.2).

We then derive the joint mass function for the numbers of gene copies $N_i (i = 1, \dots, K-1)$ at the internal nodes of the species tree. Since the reconstructed process is a pure birth process, the number N_i of gene copies at an internal node *i* must be less than or equal to the number N_j of gene copies at its descendant node *j*, i.e., $N_i \leq N_j$. We use Ω to denote all possible values of the numbers of gene copies at the internal nodes that satisfy the constrain $N_i \leq N_j$. Note that $N_i \geq 1$ for any $i \in (1, \dots, K-1)$, and the number of gene copies at the tips of the species tree are given as data.

Moreover, the number of gene copies at an internal node should be less than the numbers of gene copies at the tips of its descendent lineages. By Equation (16) in Hallinan [21], the probability of number of lineages at a time point only depends on the change in the number of lineages from an earlier time point to a later time point. Thus, given the number of gene copies at the tips of the species tree, the probability mass function of the number of genes N_1 at the root of the species tree is given by

$$P(N_1|N_0 = 1, N_1^m) = \binom{N_1^m - N_0}{N_1 - N_0}\left(\frac{\eta_{T_0,T_1} - \eta_{T_0,T_C}}{1-\eta_{T_0,T_C}}\right)^{N_2 - N_1}\left(\frac{1-\eta_{T_0,T_1}}{1-\eta_{T_0,T_C}}\right)^{N_1^m - N_1} \quad (3.3)$$

In Equation 3.3, $N_1^m = min\{N_k, \dots, N_{2k-1}\}$. In general, for any $N_i, i \in (1, \dots, K-1)$, we have

$$P(N_i|N_i', N_i^m) = \binom{N_i^m - N_i'}{N_i - N_i'}\left(\frac{\eta_{T_{i-1},T_i} - \eta_{T_{i-1},T_C}}{1-\eta_{T_{i-1},T_C}}\right)^{N_i - N_i'}\left(\frac{1-\eta_{T_{i-1},T_i}}{1-\eta_{T_{i-1},T_C}}\right)^{N_i^m - N_i} \quad (3.4)$$

In Equation 3.4, N_i^m is the minimum of the numbers of gene copies at the tips of the descendant lineages of internal node *i*.

The joint mass function of the number $(N_1, \dots, N_{K-1})$ of gene copies at internal nodes, conditional on the numbers $(N_K, \dots, N_{2K-1})$ of gene copies at the tips of the species tree, is given by

$$P(N_1, \dots, N_{K-1}|\theta, N_0 = 1, N_K, \dots, N_{2K-1}) = \prod_{i=1}^{K-1} P(N_i|N_i', N_i^m) \quad (3.5)$$

Thus, the joint probability function of gene duplication times τ and the number of gene copies at the internal nodes, given the species tree S, the number of gene copies at the external nodes and parameters θ, is equal to the product of Equations 3.1 and 3.5, i.e.,

$$f(\tau, N_1, \dots, N_{K-1}|\psi, T_1, \dots, T_{K-1}, \theta, N_K, \dots, N_{2K-1}) =$$

$$P(N_1, \dots, N_{K-1}|\psi, T_1, \dots, T_{K-1}, \theta, N_0 = 1, N_K, \dots, N_{2K-1}) f(\tau|\psi, T_1, \dots, T_{K-1}, \theta, N_1, \dots, N_{2K-1}) \quad (3.6)$$

Here the birth rate λ is constant on the entire tree. However, it may vary across branches of the species tree.

In real data analysis, the input data of the previously proposed model is the gene family tree, of which the duplication times of genes and numbers of gene copies at the internal nodes can be estimated from sequence data through the substitution model. And the gene family tree is

viewed as the "True tree" which does not take into account the estimation error of the substitution model. The model parameters, including the divergence times of the species tree and other parameters θ, are estimated by the Bayesian method which is further discussed in the following section.

3.2.3 *Bayesian estimation of model parameters*

Estimation of the model parameters is based on the joint posterior probability distribution $f(\theta, T_1, \dots, T_{K-1}|\tau, N_1, \dots, N_{2K-1})$ of θ and $T_1, \dots, T_{K-1}$, i.e.,

$$f(\theta, T_1, \dots, T_{K-1}|\tau, N_1, \dots, N_{2K-1})$$

$$= \frac{f(\tau, N_1, \dots, N_{K-1}|\psi, T_1, \dots, T_{K-1}, \theta, N_K, \dots, N_{2K-1}) f(\theta) f(T_1, \dots, T_{K-1})}{\iint_{(\theta, T_1, \dots, T_{K-1})} f(\tau, N_1, \dots, N_{K-1}|\psi, T_1, \dots, T_{K-1}, \theta, N_K, \dots, N_{2K-1}) f(\theta) f(T_1, \dots, T_{K-1}) d\theta dT_1 \dots dT_{K-1}} \quad (3.7)$$

In Equation 3.7, the numerator is composed of the likelihood function of gene family tree $f(\tau, N_1, \dots, N_{K-1}|\psi, T_1, \dots, T_{K-1}, \theta, N_K, \dots, N_{2K-1})$ and the prior distribution of model parameters $\{\theta, T_1, \dots, T_{K-1}\}$. Specifically, the parameter set, θ, have different components according to the three loss rate functions. The details of the prior distribution for each parameter in θ can be found in Appendix 2. In addition, we assume that the prior distribution of each divergence time $T_i, i = 1, \dots, K-1$ is Exponential(1).

Since the integral in the denominator of Equation (3.7) is analytically intractable, the Metropolis-Hastings algorithm[22] is adopted here to approximate the posterior distribution $f(\theta, T_1, \dots, T_{K-1}|\tau, N_1, \dots, N_{2K-1})$. The algorithm starts with a set of starting values of parameters θ and $T_1, \dots, T_{K-1}$. At each iteration, each of the parameters in θ and $T_1, \dots, T_{K-1}$, taking T_1 as an example, is updated. The new value T_1' is accepted with a probability defined by the Hastings ratio

$$H = \min\left\{\frac{f(\tau, N_1, \dots, N_{K-1}|\psi, T_1', \dots, T_{K-1}, \theta, N_K, \dots, N_{2K-1}) f(\theta) f(T_1', \dots, T_{K-1})}{f(\tau, N_1, \dots, N_{K-1}|\psi, T_1, \dots, T_{K-1}, \theta, N_K, \dots, N_{2K-1}) f(\theta) f(T_1, \dots, T_{K-1})}, 1\right\}.$$

The Metropolis-Hastings algorithm converges to the posterior distribution of model parameters after the burn-in period. And the convergence of the algorithm is evaluated through monitoring the log-likelihood values of a single chain.

3.3 SIMULATION STUDY

The simulation study by Zhao et al. [11] indicates that the age-dependent birth and death model for a single population can distinguish nonfunctionalization, subfunctionalization, and neofunctionalization when gene duplication times are given. In this section, we will demonstrate the differences among the likelihood functions of the multispecies age-dependent models for nonfunctionalization, subfunctionalization, and neofunctionalization. We will describe the steps of simulating duplication times from a multiple taxa species tree under the age-dependent model. The Bayesian estimates of model parameters θ and the divergence times are obtained by maximizing the posterior probability distribution when the topology of the species tree S is given. Finally, the results will be examined through model selection procedure to identify the underlying gene retention mechanism.

3.3.1 *Procedure of generating duplication times in a species tree*

The evolutionary process of a gene family is simulated forward in time along a species tree S. All gene copies in a gene family across species are originated from a common ancestor gene. The simulation of duplication times of these gene copies proceeds as follows:

1. The numbers $(N_1, \dots, N_{K-1})$ of gene copies at internal nodes are generated consecutively through Equation 3.4.
2. Then the duplication times of gene copies in any population can be generated using the conditional density function $f\left(t_j \middle| t_{j-1}, N_i\right), j \in (N_i' + 1, \dots, N_i)$ (see Equation 3 in Zhao et

al. [11]) of duplication time $t_{ij}, j \in (N_i' + 1, \dots, N_i)$, given its previous duplication time $t_{i,j-1}, j \in (N_i' + 1, \dots, N_i)$ and the number of gene copies N_i.

3.3.2 *Simple case study*

We consider a two-taxon species tree where the gene family evolves. Specifically, the starting time is $T_0 = 0$, the root of the species tree, which is equal to the first divergent time, is at $T_1 = 7$, and the present time is $T_C = 10$. Figure 3.1 describes the process of a gene family within two-taxon species tree in graph. Each branch in this species tree represents one population. The circle means a gene duplication event and the square denotes a speciation event. Population 1 represents the common ancestral population of species 1 and 2. At time T_1, a speciation event happens and all gene copies from population 1 are inherited by the resulting two populations. The two species have N_2 and , N_3 gene copies at present time respectively. The parameterization of the duplication and loss rate under three gene retention mechanisms is shown in Table 3.1, in which all the duplication rates are set to be equal and the loss rates are same at the time when gene duplication occurs.

The topology of species tree is assumed to be known. For each mechanism, the simulation procedure repeats 30 times. We change the numbers of gene copies at current time N_2 and N_3 such that the number of simulated data, i.e. the duplication times, varies accordingly. Particularly, N_2 ranges from 20 to 100 by 20 and N_3 ranges from 30 to 110 by 20. By increasing the number of gene copies, the accuracy of parameter estimation is expected to be improved.

3.3.3 *Performances of related probability distribution functions*

Since the generation procedure of duplication times of genes within a species tree depends on the probability mass function of the gene copy number at speciation nodes and the conditional density function of the duplication times in a single population, it is very important to check the

performance of these two functions under three mechanisms before implementing the simulation. Therefore, we first visualize these two functions using the simple case we defined in previous section. The number of genes in the current two species are given as $N_2 = 55$ and $N_3 = 60$.

In Figure 3.2, we compared the cumulative density function curves and probability mass function curves of the number of gene copies at first speciation. Specifically, Figure 3.3(a) shows that it tends to have more gene copies at the speciation time under neofunctionalization than under subfunctionalization and nonfunctionalization and curves in Figure 3.3(b) are consistent with those in Figure 3.3(a). Thus we can conclude that the mass functions of gene copies can be used to generate samples that are distinguishable under three mechanisms.

In addition, with Equation 3.3, we are able to calculate the expected number of gene copies at the speciation node, of which the numbers are: 8 under nonfunctionalization, 10 under subfunctionalization and 16 under neofunctionalization. These results also meet our expectation that the birth and death process of a gene family tends to generate more copies under neofunctionalization with a convexly decreasing curve of loss rate than under the other two processes. The calculation of the expectation is shown in Appendix 4.

To examine the divergence of the probability density functions of three mechanisms, the conditional density curves of the first duplication time is shown in Figure 3.3. It is obvious that each of the three mechanisms demonstrate a particular pattern and neofunctionalization tends to have the earliest duplication time, followed by subfunctionalization and nonfunctionalization. Thus the simulated duplication times could tell the difference among those mechanisms.

3.3.4 *Visualization of simulated data*

By visualizing the distribution of gene copy numbers through time, one can infer what effects of these different gene retention mechanisms have on the evolutionary process of gene family. For

the studies based on the reconstructed evolutionary process, two approaches have been explored in these visualizations. First of all, the average number of lineages through time in the process is plotted[23]. Secondly, the divergence times of the lineages for all the simulated trees given a certain model are plotted in one graph[24-26]. Each of these approaches has strengths and weaknesses. Plotting the average number provides a simple illustration of the process, however, with no deviation from fit information. Incorporating all the trees in one plot does not give a clear view of the process. Thus in this paper, we draw a side by side box plot for the gene copy numbers under different mechanisms through the increased gene copy number at present (Figure 3.4). In this plot, it is intuitive for one to compare the means of the number of genes at the speciation time under different model. It tends to have more genes under neofunctionalization than under the other two models, which is consistent to our expectation. In addition, the dispersions of the number of genes for each mechanism with increased current gene copy number are displayed, from which one can clearly distinguish neofunctionalization from the other two mechanisms, whereas subfunctionalization and nonfunctionalization have a higher degree of overlap.

3.3.5 *Estimation results*

To access the performance of Bayesian estimation, we use the simulated duplication times as input data to estimate model parameters. For nonfunctionalization, the root-mean-square error (RMSE) of the estimate of μ and λ decrease as the number of current gene copies N_2 and N_3 increase (Figure 3.1). Additionally, the root-mean-square errors (RMSE) of the estimates of parameters for neofunctionalization and subfunctionalization decrease as N_2 and N_3 increase. However, the decline pattern in neofunctionalization and subfunctionalization are not so obvious as in nonfunctionalization. This is mainly caused by two facts. On one hand, there are more

parameters in neofunctionalization and subfunctionalization, of which some are correlated in the loss rate function. On the other hand, the duplication times are highly correlated, and the auto-correlation between two adjacent duplication times increases as the number of the duplication time increases. In other words, when the number of duplication times reaches a certain number, it appears that adding more duplication times does not contribute more information for accurately estimating model parameters.

3.4 DISCUSSION

3.4.1 *Summary of the gene family evolution model within species tree*

In this paper, we have extended the age-dependent birth and death model in one population [11] into the multiple populations scenario, that is, a species tree. In this model, the joint probability distribution of gene duplication times τ and the numbers of gene copies at the internal nodes, given the number of gene copies at present and the species tree, is developed such that the model parameters could be estimated through likelihood based methods. Noticing that the loss rate functions of neofunctionalization and subfunctionalization are further polished to represent more flexible curves by incorporating one additional parameter. The results from the simulation study show that the proposed model is able to differentiate the three gene retention mechanisms and the Bayesian estimates of model parameters demonstrate an increasing accuracy as the sample size increases.

3.4.2 *Future study*

The input data in the current model is the duplication times in a gene family tree which are assumed to be the truth. However, this kind of data is not observable in reality and need to be estimated from the sequence data, through which more estimation errors would be introduced. Thus it is necessary to build a hierarchical model that includes a mutation process modeling the

sequence given a gene family tree and a birth and death process describing the gene family evolution in a species tree.

It is worth noticing that the neofunctionalization and subfunctionalization mechanisms for duplicated genes would place a selection pressure and change the mutation rates correspondingly. This would invalidate the existing substitution models whose mutation rates do not reflect the effects of neofunctionalization and subfunctionalization. Thus a mutation process that integrates different gene retention mechanisms is in need.

3.5 REFERENCES

1. Demuth, J.P. and M.W. Hahn, *The life and death of gene families.* Bioessays, 2009. **31**(1): p. 29-39.
2. Hurles, M., *Gene duplication: the genomic trade in spare parts.* PLoS Biol, 2004. **2**(7): p. E206.
3. Lynch, M., et al., *The probability of preservation of a newly arisen gene duplicate.* Genetics, 2001. **159**(4): p. 1789-804.
4. Ohta, T., *Role of gene duplication in evolution.* Genome, 1989. **31**(1): p. 304-10.
5. Zhang, J.Z., *Evolution by gene duplication: an update.* Trends in Ecology & Evolution, 2003. **18**(6): p. 292-298.
6. Walsh, J.B., *How often do duplicated genes evolve new functions?* Genetics, 1995. **139**(1): p. 421-8.
7. Konrad, A., et al., *Toward a general model for the evolutionary dynamics of gene duplicates.* Genome Biol Evol, 2011. **3**: p. 1197-209.
8. Lynch, M. and J.S. Conery, *The evolutionary fate and consequences of duplicate genes.* Science, 2000. **290**(5494): p. 1151-5.

9. Hughes, T. and D.A. Liberles, *The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation.* J Mol Evol, 2007. **65**(5): p. 574-88.

10. Teufel, A.I., et al., *On Mechanistic Modeling of Gene Content Evolution: Birth-Death Models and Mechanisms of Gene Birth and Gene Retention.* Computation, 2014. **2(3),**.

11. Zhao, J., et al., *A generalized birth and death process for modeling the fates of gene duplication.* BMC Evol Biol, 2015. **15**: p. 275.

12. Force, A., et al., *Preservation of duplicate genes by complementary, degenerative mutations.* Genetics, 1999. **151**(4): p. 1531-45.

13. Khan, A.A., et al., *Phylogenetic analysis of kindlins suggests subfunctionalization of an ancestral unduplicated kindlin into three paralogs in vertebrates.* Evol Bioinform Online, 2011. **7**: p. 7-19.

14. Rastogi, S. and D.A. Liberles, *Subfunctionalization of duplicated genes as a transition state to neofunctionalization.* BMC Evol Biol, 2005. **5**: p. 28.

15. Goodman, M., et al., *Fitting the Gene Lineage into Its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences.* Systematic Zoology, 1979. **28**(2): p. 132-163.

16. Page, R.D.M. and M.A. Charleston, *From gene to organismal phylogeny: Reconciled trees and the gene tree species tree problem.* Molecular Phylogenetics and Evolution, 1997. **7**(2): p. 231-240.

17. Nakhleh, L., *Computational approaches to species phylogeny inference and gene tree reconciliation.* Trends Ecol Evol, 2013. **28**(12): p. 719-28.

18. Akerborg, O., et al., *Simultaneous Bayesian gene tree reconstruction and reconciliation analysis.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(14): p. 5714-5719.

19. Arvestad, L., et al., *Bayesian gene/species tree reconciliation and orthology analysis using MCMC.* Bioinformatics, 2003. **19 Suppl 1**: p. i7-15.

20. De Bie, T., et al., *CAFE: a computational tool for the study of gene family evolution.* Bioinformatics, 2006. **22**(10): p. 1269-71.

21. Hallinan, N., *The generalized time variable reconstructed birth-death process.* J Theor Biol, 2012. **300**: p. 265-76.

22. HASTINGS, W.K., *Monte Carlo sampling methods using Markov chains and their applications.* Biometrika, 1970. **57**(1): p. 97-109.

23. Nee, S., R.M. May, and P.H. Harvey, *The reconstructed evolutionary process.* Philos Trans R Soc Lond B Biol Sci, 1994. **344**(1309): p. 305-11.

24. Rabosky, D.L., *Likelihood methods for detecting temporal shifts in diversification rates.* Evolution, 2006. **60**(6): p. 1152-64.

25. Stadler, T., *Mammalian phylogeny reveals recent diversification rate shifts.* Proc Natl Acad Sci U S A, 2011. **108**(15): p. 6187-92.

26. Crisp, M.D. and L.G. Cook, *Explosive radiation or cryptic mass extinction? Interpreting signatures in molecular phylogenies.* Evolution, 2009. **63**(9): p. 2257-65.

3.6 APPENDICES

In the following sections, the equation numbers without the prefix, A, correspond to those in the main article.

Appendix. 1 *Calculation of mean loss rates of the age-dependent model in the context of multiple populations.*

According to probability density function of age t' derived by Zhao et al. [11] (Equation 6):

$$f(t') = \frac{e^{-\int_{t-t'}^{t}(\lambda+\mu_x)dx}}{\int_0^t \left(e^{-\int_{t-t'}^{t}(\lambda+\mu_x)dx}\right)dt'}$$

For neofunctionalization: $\mu_t = \alpha e^{-t\beta}$ and mean loss rate is simplified as

$$\phi_t^{neo} = E\left(\mu_{t'}\right) = \int_0^t \mu_{t'} f(t')dt'$$

$$= \int_0^t \mu_{t'} \frac{e^{-\int_{t-t'}^{t}(\lambda+\mu_x)dx}}{\int_0^t \left(e^{-\int_{t-t'}^{t}(\lambda+\mu_x)dx}\right)dt'} dt'$$

$$= \frac{\int_0^t \alpha e^{-t'\beta} e^{-\int_{t-t'}^{t}(\lambda+\alpha e^{-x\beta})dx} dt'}{\int_0^t \left(e^{-\int_{t-t'}^{t}(\lambda+\alpha e^{-x\beta})dx}\right)dt'}$$

$$= \frac{\int_0^t \alpha e^{-t'\beta} e^{-\lambda t' + \frac{\alpha}{\beta}\left(e^{-t\beta} - e^{-(t-t')\beta}\right)} dt'}{\int_0^t e^{-\lambda t' + \frac{\alpha}{\beta}\left(e^{-t\beta} - e^{-(t-t')\beta}\right)} dt'}$$

For subfunctionalization: $\mu_t = \frac{\alpha e^{-\beta(t-\gamma)}}{1+e^{-\beta(t-\gamma)}}$ and mean loss rate is simplified as

$$\phi_t^{sub} = E\left(\mu_{t'}\right) = \int_0^t \mu_{t'} f(t')dt' = \int_0^t \mu_{t'} \frac{e^{-\int_{t-t'}^{t}(\lambda+\mu_x)dx}}{\int_0^t \left(e^{-\int_{t-t'}^{t}(\lambda+\mu_x)dx}\right)dt'} dt'$$

$$= \frac{\int_0^t \frac{\alpha e^{-\beta(t'-\gamma)}}{1+e^{-\beta(t'-\gamma)}} e^{-\int_{t-t'}^{t}\left(\lambda+\frac{\alpha e^{-\beta(x-\gamma)}}{1+e^{-\beta(x-\gamma)}}\right)dx} dt'}{\int_0^t \left(e^{-\int_{t-t'}^{t}\left(\lambda+\frac{\alpha e^{-\beta(x-\gamma)}}{1+e^{-\beta(x-\gamma)}}\right)dx}\right)dt'}$$

$$= \frac{\int_0^t \frac{\alpha e^{-\beta(t'-\gamma)}}{1+e^{-\beta(t'-\gamma)}} e^{-\lambda t'+\frac{\alpha}{\beta}\ln\left(\frac{e^{-\beta t}+e^{-\beta\gamma}}{e^{-\beta(t-t')}+e^{-\beta\gamma}}\right)} dt'}{\int_0^t e^{-\lambda t'+\frac{\alpha}{\beta}\ln\left(\frac{e^{-\beta t}+e^{-\beta\gamma}}{e^{-\beta(t-t')}+e^{-\beta\gamma}}\right)} dt'}$$

The simplified expressions of ϕ_t^{neo} and ϕ_t^{sub} have successfully solved one fold integral which can definitely improve the accuracy and speed of the mean loss rate calculation.

Appendix. 2 *The prior distributions of each component in* θ.

For nonfunctionalization, the duplication rate and loss rate are constant, so $\theta = (\mu, \lambda)$. Since both μ and λ range in $(0,1)$, we assume that the prior distribution $f(\lambda)$ of the duplication rate and $f(\mu)$ of the loss rate are uniform (0,1).

For neofunctionalization, the duplication rate is constant and loss rate is a function of the age of gene, so $\theta = (\alpha, \beta, \lambda)$. And we know that $\alpha \in (0,1), \beta > 0 \; and \; \lambda \in (0,1)$, we assume the prior distributions of α and λ are uniform(0,1) and the prior of β is exponential (1).

For subfunctionalization, duplication rate is constant and loss rate is a function of gene age, thus $\theta = (\alpha, \beta, \gamma, \lambda)$. Both α and λ lie in $(0,1)$ and $\beta > 0$ and $\gamma > 0$, so the prior distributions of α and λ are uniform(0,1) and priors of β and γ are exponential(1).

Appendix. 4 *Calculation of expectation of gene copies at speciation using Equation (7).*

$$P(N_i|N_{i-1}, N_i(i = K+1, \dots, 2K))$$

$$= \binom{N_S^m - N_{i-1}}{N_i - N_{i-1}} \left(\frac{\eta_{T_{i-1},T_i} - \eta_{T_{i-1},T_C}}{1 - \eta_{T_{i-1},T_C}}\right)^{N_i - N_{i-1}} \left(\frac{1 - \eta_{T_{i-1},T_i}}{1 - \eta_{T_{i-1},T_C}}\right)^{N_S^m - N_i} \quad (A.7)$$

$$E\left(N_i|N_{i-1}, N_i(i = K+1, \dots, 2K)\right)$$

$$= \sum_{l=N_{i-1}+1}^{N_S^m} N_l \binom{N_S^m - N_{i-1}}{N_l - N_{i-1}} \left(\frac{\eta_{T_{i-1},T_i} - \eta_{T_{i-1},T_C}}{1 - \eta_{T_{i-1},T_C}}\right)^{N_l - N_{i-1}} \left(\frac{1 - \eta_{T_{i-1},T_i}}{1 - \eta_{T_{i-1},T_C}}\right)^{N_S^m - N_l}$$

Appendix. 5 *The estimation error of parameters*

Table A.1 Estimation errors of parameters for nonfunctionalization

	$N_2 = 20$	$N_2 = 40$	$N_2 = 60$	$N_2 = 80$	$N_2 = 100$
μ (0.8)	0.162	0.125	0.094	0.090	0.086
λ (0.2)	0.047	0.035	0.029	0.022	0.021

Table A.2 Estimation errors of parameters for neofunctionalization

	$N_2 = 20$	$N_2 = 40$	$N_2 = 60$	$N_2 = 80$	$N_2 = 100$
α (0.8)	0.193	0.167	0.145	0.113	0.095
β (0.3)	0.062	0.051	0.049	0.040	0.034
λ (0.2)	0.053	0.041	0.033	0.031	0.029

Table A.3 Estimation errors of parameters for subfunctionalization

	$N_2 = 20$	$N_2 = 40$	$N_2 = 60$	$N_2 = 80$	$N_2 = 100$
α (0.8)	0.203	0.201	0.194	0.173	0.152
β (0.5)	0.170	0.129	0.113	0.095	0.067
γ (7)	1.538	1.437	1.177	0.995	0.834
λ (0.2)	0.051	0.040	0.039	0.037	0.033

Table 3.1: The values of parameters in defining the duplication and loss rates used in generating duplication times under nonfunctionalization, neofunctionalization, and subfunctionalization.

	λ	μ	α	β	γ
Nonfunctionalization	0.2	0.8			
Neofunctionalization	0.2		0.8	0.3	
Subfunctionalization	0.2		0.8	0.5	7

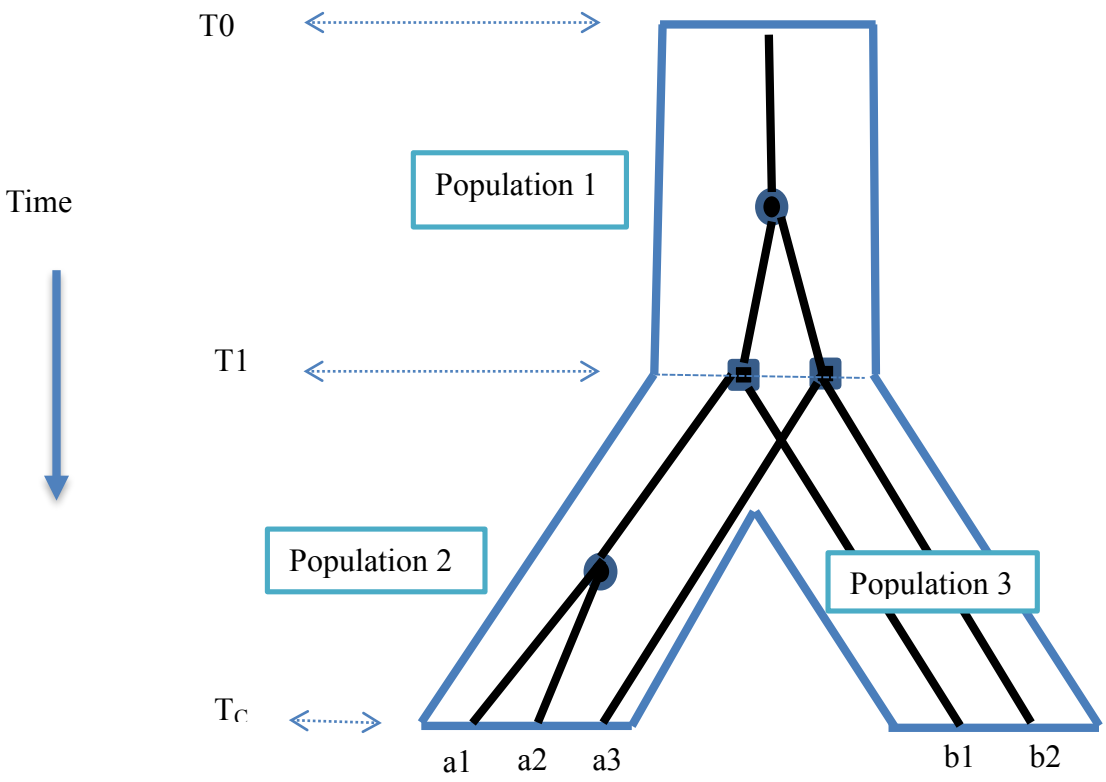

Figure 3.1. The evolutionary process of a gene family tree within a species tree is shown.

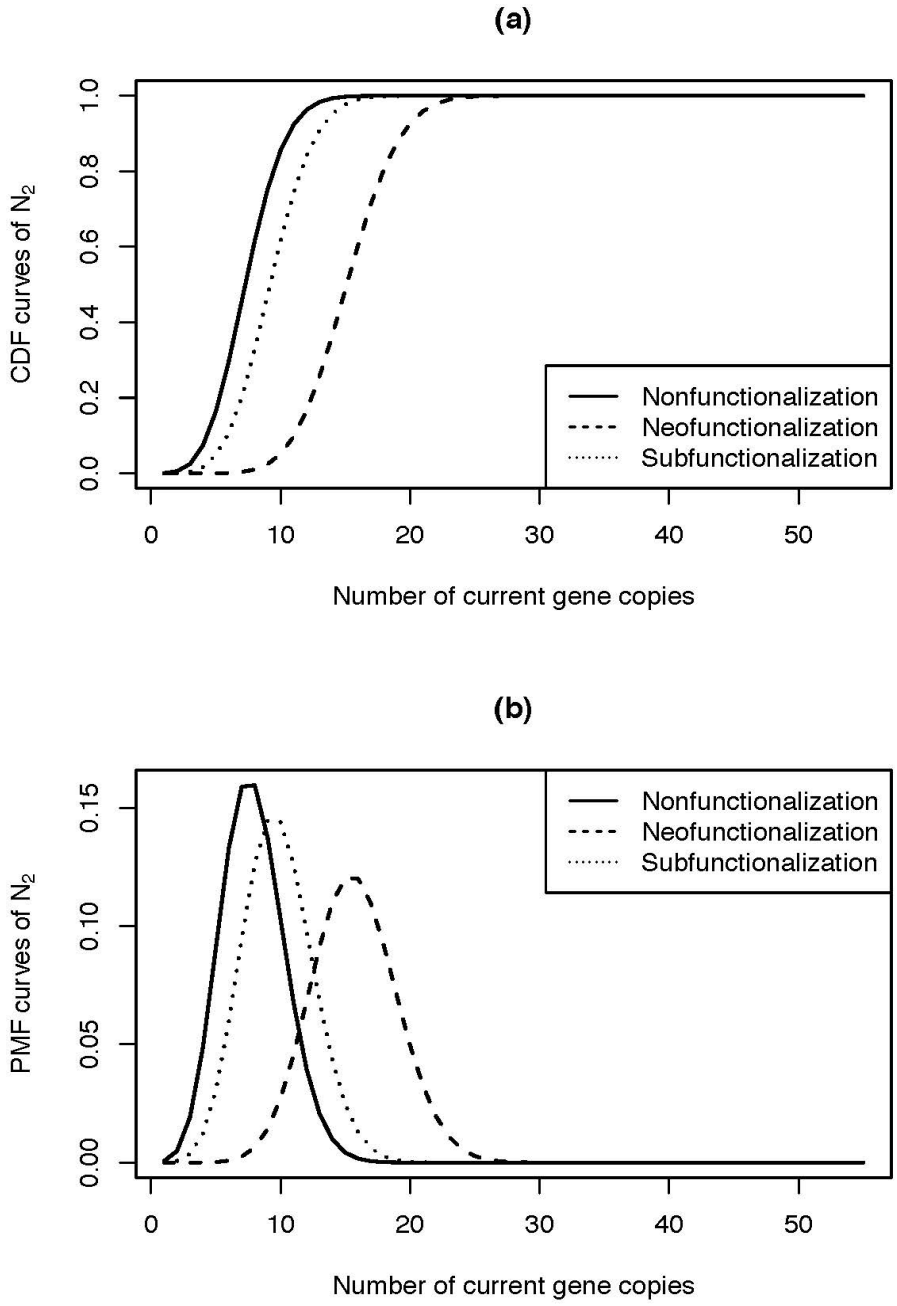

Figure 3.2 Performances of the probability mass function of the gene copy number at divergence time: (a) the cumulative density curves for nonfunctionalization, neofunctionalization, and subfuncitonalization are shown; (b) the probability mass curves for nonfunctionalization, neofunctionalization, and subfuncitonalization are shown.

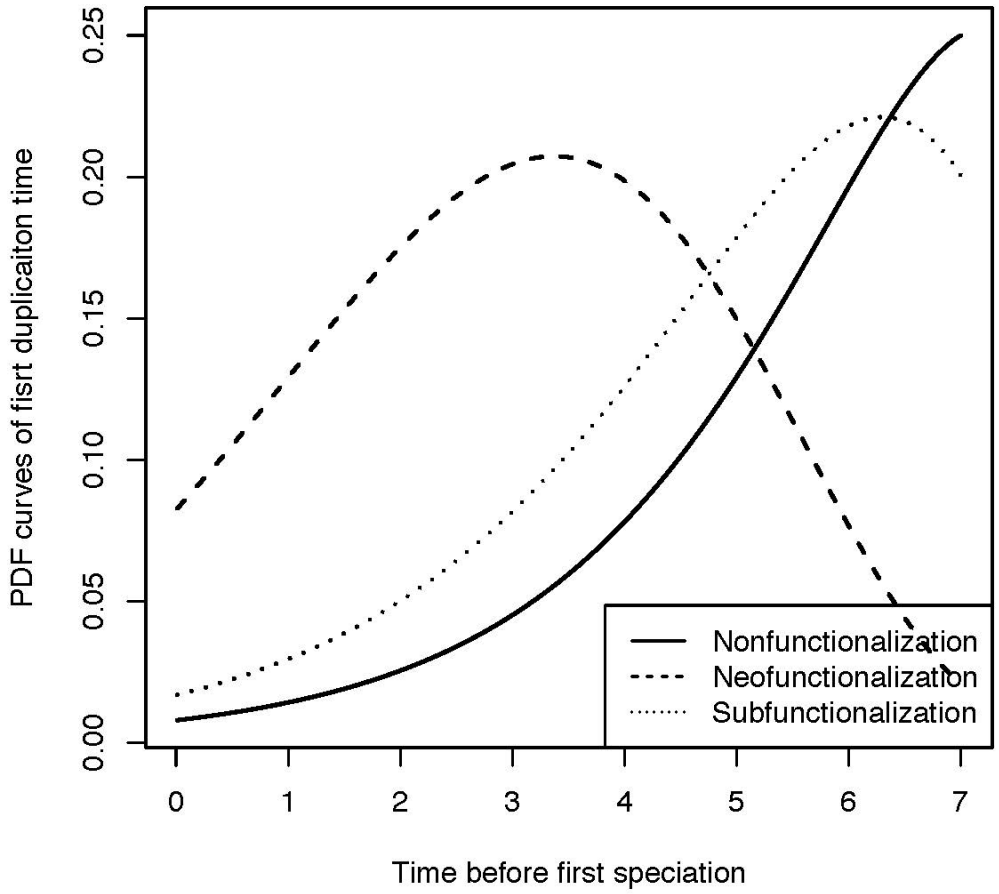

Figure 3.3 The probability density curves of the first duplication times for nonfunctionalization, neofunctionalization, and subfunctionalization under the model are shown

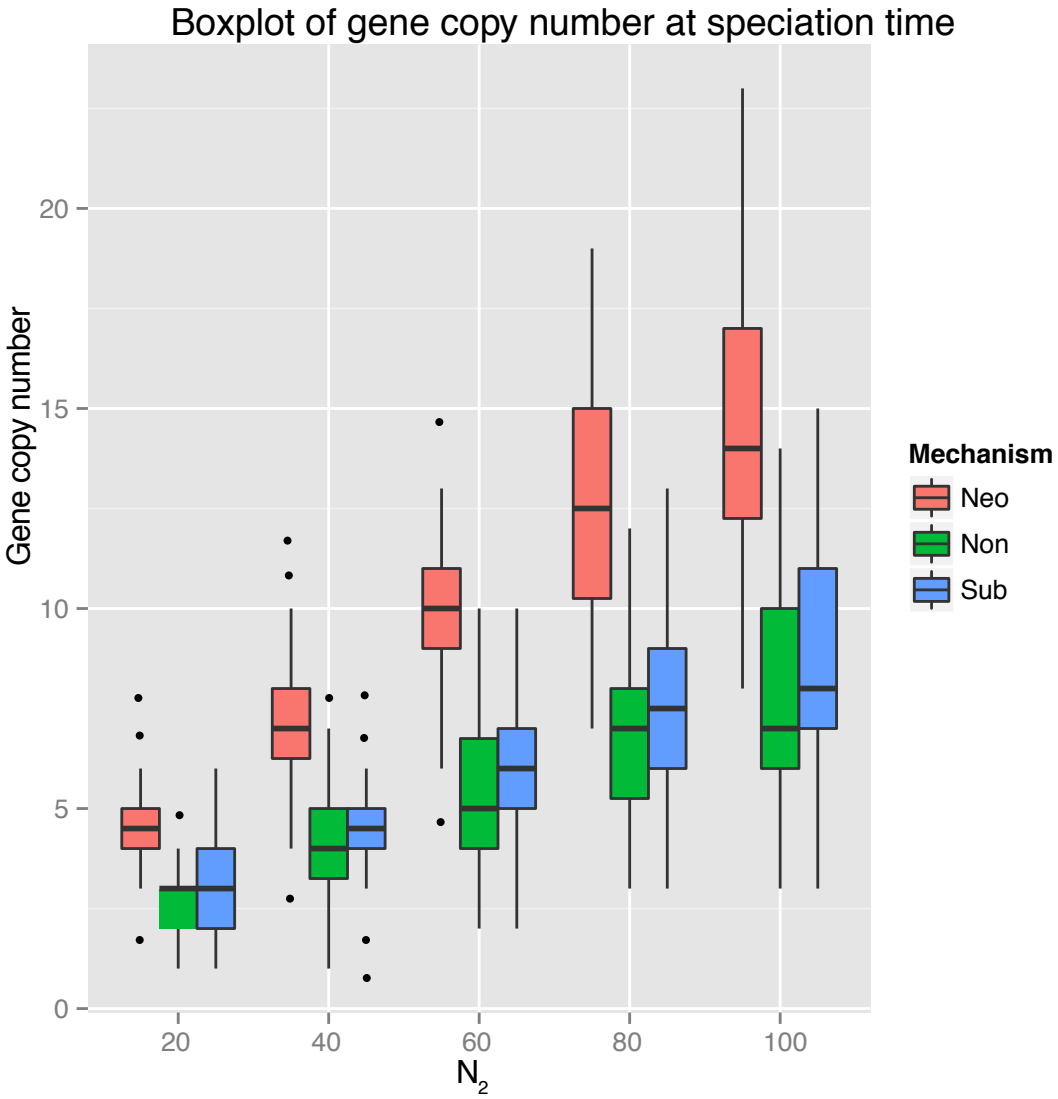

Figure 3.4 The boxplots of simulated gene copy numbers at the divergence time with increased current gene copy numbers in nonfunctionalization, neofunctionalization and subfunctionalization.

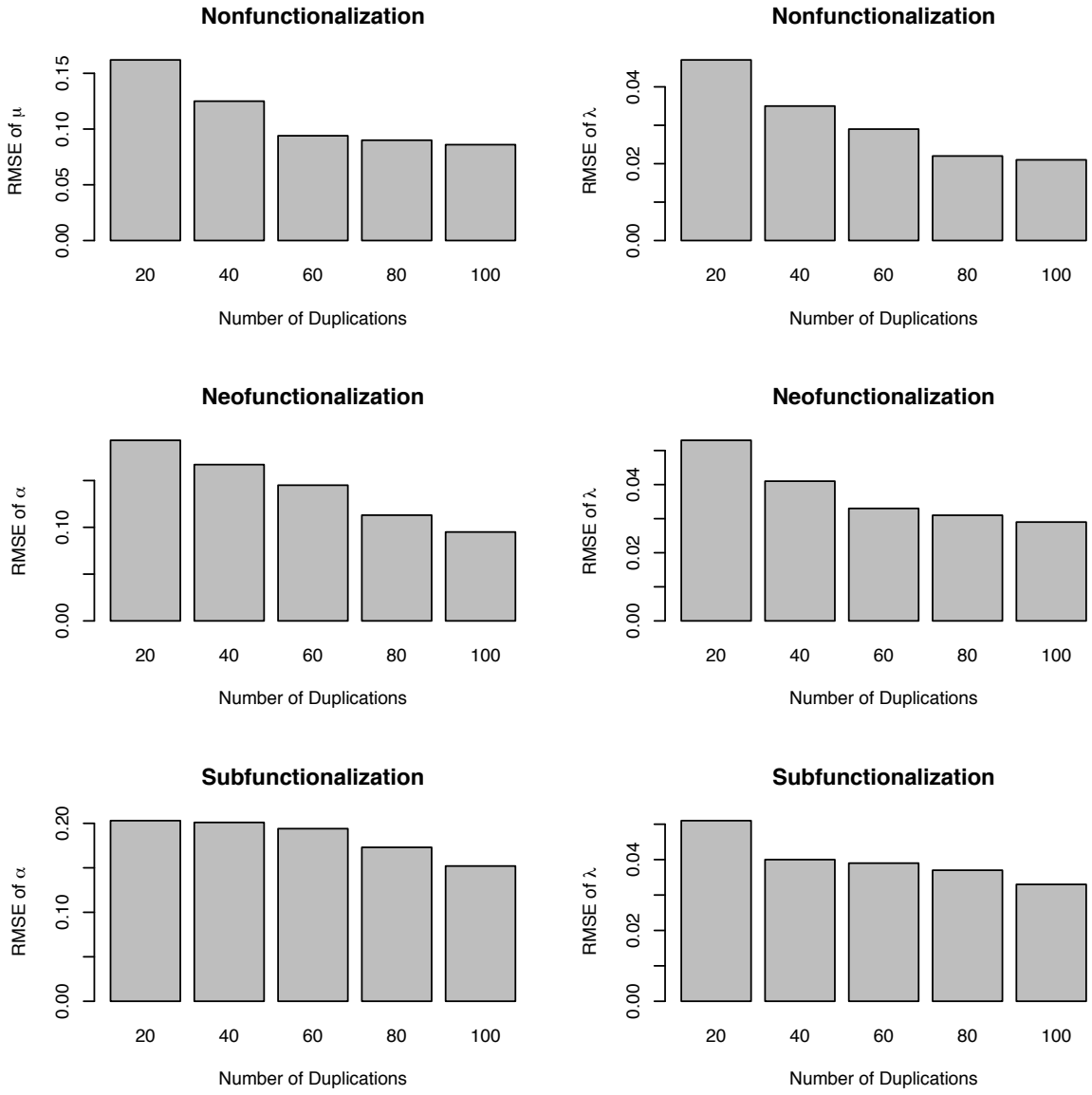

Figure 3.5 The root of mean squared errors (RMSE) of the Bayesian estimates under three gene retention mechanisms in the simulation study.

CHAPTER 4

A BAYESIAN HIERARCHICAL MODEL FOR GENE FAMILY EVOLUTION

In this chapter, a conceptual Bayesian hierarchical model will be developed to explain the relationship between the species tree, gene family tree, and the DNA sequences (data).

4.1 THE PROBABILITY DENSITY FUNCTIONS

We assume that gene family trees are generated from a duplication/loss process (i.e., a non-homogeneous birth and death process) occurring along the lineages of the species tree. In addition, we assume that DNA sequences evolve on the gene family trees, following a substitution model. Thus the Bayesian hierarchical model includes the following probability density functions:

4.1.1 *The probability density function of gene family trees given the species tree*

As gene family trees are conditionally independent given the species tree, we have

$$f(\boldsymbol{G}|S,\theta) = \prod_{i=1}^{K} f(G_i|S,\theta) \qquad (4.1)$$

In Equation 4.1, $\boldsymbol{G}$ is a vector of multiple gene family trees, G_i denotes the i^{th} gene family tree, K is the number of gene family trees, S represent the topology and branch lengths of the species tree, and θ denotes the parameters related to the age-dependent loss rate and constant duplication rate in the age-dependent model discussed in Chapter 3 and 4.

The probability density function of a gene family tree within a species tree $f(G_i|S,\theta)$ could be derived from Equation 3.6. In Chapter 3, the gene family tree is estimated from sequence data and considered as the input data to estimate parameters of the model, in which the numbers of gene copies at the internal nodes are fixed. While the gene family tree is treated as a random variable in current Bayesian hierarchical model, wherein the numbers of gene copies at the internal nodes are

varied in the space Ω defined in Chapter 3. For this reason, $f(G_i|S,\theta)$ is calculated by summing over all possible values of the numbers of gene copies at the internal nodes in Equation 3.6 as shown below:

$$f(G_i|S,\theta) = f(\tau,|\psi,T_1,\dots,T_{K-1},\theta,N_K,\dots,N_{2K-1})$$

$$= \sum_{(N_1,\dots,N_{K-1})\in\Omega} f(\tau,N_1,\dots,N_{K-1}|\psi,T_1,\dots,T_{K-1},\theta,N_K,\dots,N_{2K-1}) \quad (4.2)$$

In Equation 4.2, all the notations are directly inherited from Chapter 3.

4.1.2 *The probability density function of DNA sequences given the gene family tree*

The substitution processes of DNA sequences occur along the branches of a gene family tree. The substitution of nucleotide is modeled as a continuous time Markov process conditional on a given gene tree topology and a vector of branch lengths. It is assumed that different nucleotide sites within a sequence, and different lineages, experience independent substitutions, and no recombination occurs. In this way the probabilities of nucleotides observed at different sites can be calculated independently.

The following abbreviations are used here: $\boldsymbol{D}$ is sequence data (from multiple gene families), D_i is sequence data from one gene family, $\boldsymbol{G}$ denotes a vector of gene family trees, and ζ denotes the parameters in the substitution model. The probability density function $f(D_i|G_i,\zeta)$ of DNA sequences given a gene family tree is defined by the sequence evolutionary model. This model is conditionally independent of the species tree since it only depends on the topology and branch lengths of the gene family tree.

$$f(\boldsymbol{D}|\boldsymbol{G},\zeta) = \prod_{i=1}^{K} f(D_i|G_i,\zeta) \quad (4.3)$$

It is worth noticing that the probability density function $f(D_i|G_i,\zeta)$ is not a simple application of the existing substitution model, such as HKY and F81. This is caused by the fact

that the selection pressure on sequence evolution is decreased when neofunctionalization or subfunctionalization happens. In addition, the mutation rate of single nucleotide, which is proportional to the selection pressure, will decrease accordingly. Therefore, a new substitution model is desired to be developed in order to incorporate the changes of mutation rate in the context of neofunctionalization and subfunctionalization.

4.2 PRIOR DISTRIBUTIONS OF MODEL PARAMETERS

The parameters in the hierarchical model include the duplication rate λ, parameters $\alpha, \beta \; and \; \gamma$ in the function of the loss rate, gene family trees, and the species tree. We assume that the topology of the species is fixed, while the branch lengths of the species tree will be estimated from data. The branch lengths τ in the gene family trees and the species tree are in the mutation units, i.e., $\tau = \eta t$, where η is the mutation rate. We assume that the birth rate may vary across populations on the species tree (we use λ_i to denote the birth rate for population *i*). In addition, we assume the loss rate parameter follows the same function on the entire tree. The parameters of a gene family tree include not only the topologies, but also the node times (duplication times) of the tree. The node times of gene family trees are distributed on the branches of the species tree, which can be used to identify duplication events occurring in each population of the species tree. Moreover, the estimates of the parameters $\alpha, \beta \; and \; \gamma$ in the loss function will help identify the evolutionary fates of gene duplicates. All the parameters in the hierarchical model have important biological implications.

In addition to the two probability density functions described in last section, we need to specify a prior distribution for each parameter in the hierarchical model. The choice of prior distributions depends on the nature of the data at hand. Different users may choose different priors for the parameters. By default, we assume that the birth rate λ and the starting value α of loss rate

follow uniform (0, c) distribution. The constant c will be sufficiently large in order to cover practically possible realization of λ and α. We choose exponential distributions as the priors of the decreasing rate $\beta$and the time at inflection point of logistic curve γ in loss rate function. It has been suggested that the branch length of a gene family tree follows an exponential distribution. Thus the prior of the branch length of a gene family tree is an exponential distribution with mean v (by default, v = 10). Similarly, we assume that the branch length of the species tree has an exponential distribution with mean 10 by default. We can adjust the values of the hyper parameters c and v.

4.3 POSTERIOR DISTRIBUTIONS

Estimation of parameters in the hierarchical model is based on the posterior distribution, which is the combination of the probability density function of data and the prior distribution of model parameters,

$$f(S,\alpha,\beta,\gamma,\lambda|\boldsymbol{D}) = \frac{f(\boldsymbol{D}|\boldsymbol{G},\zeta)\times f(\boldsymbol{G}|S,\alpha,\beta,\gamma,\lambda)\times f(\theta)}{\int_{\theta} f(\boldsymbol{D}|\boldsymbol{G},\zeta)\times f(\boldsymbol{G}|S,\alpha,\beta,\gamma,\lambda)\times f(\theta)d\theta} \qquad (4.4)$$

Here θ denotes all the parameters in the hierarchical model, i.e. $\theta = \{\alpha,\beta,\gamma,\lambda,\zeta,S\}$.

Due to the intractable integral on the denominator of the posterior distribution, we apply Markov Chain Monte Carlo (MCMC) algorithm to approximate the posterior distribution. We will examine the performance on convergence of this algorithm by generating samples with increased sample sizes. If the estimations from these samples converge to the true parameters as sample size increases, then we can conclude that the Bayesian inferences on model parameters can be achieved using samples generated from the MCMC algorithm.

4.4 BAYESIAN INFERENCE OF GENE FAMILY EVOLUTION

In this section, we will discuss the applications of the Bayesian model in addressing real problems regarding gene family evolution.

4.4.1 *Testing the homogeneous birth and death rate model*

As we discussed above, it is of significant interest in the studies of gene family evolution to test whether the duplication and loss rates are constant over time. The branches with significantly high birth rates may have important biological implications. For instance, it may indicate that whole genome duplication may have occurred on those branches of the species tree. The hypothesis of a homogeneous birth rate (or death rate) can be tested under the Bayesian model. In addition, the heterogeneous death rate may imply the existence of a gene retention mechanism, such as neofunctionalization and subfunctionalization. In this study, the hypothesis of a homogeneous death rate is of more interest. Thus the null and alternative hypotheses are

H_0: homogeneous death rate vs H_1: heterogeneous death rates

We evaluate the homogeneous death rate model and the heterogeneous death rate model under the Bayesian framework and then calculate the Bayes factor for the two models, $BF = f(X|H_1) / f(X|H_0)$, where $f(X \mid H_0)$ is the marginal likelihood under the null hypothesis and $f(X \mid H_1)$ is the marginal likelihood under the alternative hypothesis. The evidence for supporting the null hypothesis (H_0) against the alternative hypothesis (H_1) is based on the Bayes Factor. In general, $Ln(BF) > 10$ is interpreted as strong evidence for supporting the alternative hypothesis.

4.4.2 *Classifying gene families based on their evolutionary fates*

Classification of gene family is the act of grouping genes or proteins into families, which leads to a better understanding of the evolutionary forces and functions of genes[1-3]. During last 20 years, lots of sequence-based methods have been established for classification of gene family. These

methods have been divided into three categories according to Frech and Chen [3], which are phylogenetic inference, classification based on sequence signatures, and pairwise comparisons of full protein sequences. In this study, we are focusing on the classification of gene families based on the Bayesian hierarchical model developed in this dissertation.

As described in Chapter 3 and Chapter 4, a specific parameterization of the loss rate function corresponds to a unique evolutionary fate of gene duplicates (nonfunctionalization, neofunctionalization, or subfunctionalization). The estimation of the parameters in the loss rate function can provide information regarding the retention mechanisms of the gene families in the data. Gene family data can be obtained from the adaptive evolution database (TAED) [4] or a whole genome gene family database (Hobacgen) [5]. Then the families with the same parameterization of the loss rate function are grouped. To achieve this goal, additional parameters $\{\nu_i, i = 1, \dots, K\}$ are added to the probabilistic model we described in section 4.3, where ν_i indicates the category that gene family i belongs to. There are three categories, i.e., $\nu_i = 1, 2, 3$. Each category represents a particular underlying mechanism that the gene family has undergone, i.e., 1:neofunctionalization, 2: nonfunctionalization, and 3: subfunctionalization. The parameters $\{\nu_i, i = 1, \dots, K\}$ can be estimated under the Bayesian framework with a uniform prior for ν_i, such that $P(\nu_i = 1) = P(\nu_i = 2) = P(\nu_i = 3) = 1/3$. The uniform prior indicates that gene family i has an equal probability to fall into any of the three categories. The assignment of gene family i is determined by the maximum posterior probability of ν_i.

4.5 DISCUSSION

A theoretical Bayesian framework has been established to illustrate the relationship among DNA sequence data, gene family tree and species tree in this chapter. The probability density function of the gene family tree within a species tree is obtained by extending the result in Chapter 3. The

substitution model of sequence data is suggested to be revised to include the effects of neofunctionalization and subfunctionalization. Furthermore, the priors of model parameters are discussed and a posterior distribution is provided based on the probability density functions and the priors. However, without the exact expression of the substitution model, it is impossible to implement the simulation study and real data analysis. Nevertheless, the current Bayesian hierarchical model still bring a distinct perspective in modeling the sequence data, gene family tree and species tree simultaneously in the context of different gene retention mechanisms. Finally, two possible empirical applications of the Bayesian hierarchical model have been suggested and the implementation processes are given to facilitate future work.

4.6 REFERENCES

1. Wu, C.H., et al., *Protein family classification and functional annotation.* Comput Biol Chem, 2003. **27**(1): p. 37-47.
2. Demuth, J.P., et al., *The evolution of mammalian gene families.* PLoS One, 2006. **1**: p. e85.
3. Frech, C. and N. Chen, *Genome-wide comparative gene family classification.* PLoS One, 2010. **5**(10): p. e13409.
4. Roth, C., et al., *The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics.* Nucleic Acids Res, 2005. **33**(Database issue): p. D495-7.
5. Perriere, G., L. Duret, and M. Gouy, *HOBACGEN: database system for comparative genomics in bacteria.* Genome Res, 2000. **10**(3): p. 379-85.

CHAPTER 5

OVERALL CONCLUSIONS

5.1 SUMMARY

In this dissertation, we first explored the process of gene family evolution in a single population, in which the loss rate of duplicated gene is associated with different gene retention mechanisms. Based on the theory of nonhomogeneous birth and death process, we have derived the probability density function of the gene family tree given a species tree with an age-dependent loss rate. For each of the gene retention mechanisms, we have performed a simulation study, wherein the duplication times of genes in a single population are generated in a forward direction. These generated duplication times are then used as the input data to estimate model parameters through maximum likelihood estimation. It is shown that the established model is able to accurately estimate parameters and distinguish gene retention mechanisms.

We have also extended the age-dependent birth and death model in one population into multiple populations, we have once again derived the joint probability density function of duplication times and number of gene copies at the internal nodes. Unlike the one population scenario, for the multiple populations, we have considered the distribution of the number of gene copies at internal nodes of species tree. Once again the simulation study is performed and estimation results from Bayesian method are examined to see the validity of the model.

Finally, a Bayesian hierarchical model is discussed in conceptual level, in which the mutation process of DNA sequences and the birth and death process of genes are combined. The probability density function of the gene family tree within a species tree is derived based on the

age-dependent model. It is also suggested that the substitution model of sequence data should be developed to include the effects of neofunctionalization and subfunctionalization. The priors of model parameters are discussed and a posterior distribution is provided. The proposed Bayesian hierarchical model leads to a novel viewpoint in modeling the sequence data, gene family tree and species tree simultaneously. Furthermore, two applications of the Bayesian hierarchical model have been suggested in solving real problems which can benefit future study.

5.2 LIMITATIONS AND FUTURE STUDY

Although the age-dependent birth and death process was generalized to sequence data in conceptual level, there is still lack of a proper substitution model allowing for effects of different gene retention mechanisms. Specifically, the mutation rate of nucleotide in a sequence would decrease due to the reduced selection pressure when neofunctionalization and subfunctionalization occur. In addition, the current age-dependent model is conditional on observed duplicate copies which does not account for the full productive process including duplicates that were lost before the present. Thus we plan to examine this in the context of Approximate Bayesian Computation. Furthermore, missing data and genome assembly error are not specifically addressed in the modeling framework and their impact on inference also needs to be addressed. Lastly, the models can be used to make predictions about functional evolution in the absence of actual functional data. While such data does not available in large scale, the future may bring expression data of protein duplicates that can be integrated into a phylogenetic framework. However, even with comparative proteomic data in the future, one still needs models that account for signals related to selective pressures (like the models presented here), since neutral changes in expression and functional properties would not lead to changes in gene retention profiles and meaningful lineage-specific

biology (see [3] for a discussion of the interaction between molecular phenotypes and biological function in an evolutionary view).

The model developed currently made an assumption that all duplicates in a gene family evolve under the same process. It is possible to examine large gene family databases like Ensembl [4], HOGENOM [5], or TAED [6] and establish a mixture model of duplicate processes applied across all gene families to enable a probabilistic identification of gene retention mechanisms for individual gene duplication events. The work presented in this dissertation with a birth and death model in a phylogenetic context, brings this scale of modeling one step closer.

5.3 REFERENCES

1. Teufel, A.I., J. Masel, and D.A. Liberles, *What Fraction of Duplicates Observed in Recently Sequenced Genomes Is Segregating and Destined to Fail to Fix?* Genome Biol Evol, 2015.
2. Hughes, T. and D.A. Liberles, *The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation.* J Mol Evol, 2007. **65**(5): p. 574-88.
3. Graur, D., et al., *On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE.* Genome Biol Evol, 2013. **5**(3): p. 578-90.
4. Flicek, P., et al., *Ensembl 2012.* Nucleic Acids Res, 2012. **40**(Database issue): p. D84-90.
5. Penel, S., et al., *Databases of homologous gene families for comparative genomics.* BMC Bioinformatics, 2009. 10 Suppl 6: p. S3.
6. Roth, C., et al., *The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics.* Nucleic Acids Res, 2005. **33**(Database issue): p. D495-7.