Bull. Aust. Math. Soc. **108** (2023), 513–515
doi:10.1017/S0004972723000801

MECHANISTIC MARKOV MODELS FOR THE EVOLUTION OF GENE FAMILIES

JIAHAO DIAO

(Received 2 July 2023; first published online 10 August 2023)

2020 *Mathematics subject classification*: primary 60J20; secondary 92D10.

Keywords and phrases: gene family evolution, level-dependent quasi-birth-death process, Markov chains, subfunctionalisation model, neofunctionalisation model, hidden Markov model.

Gene duplication has been identified as one of the key processes for driving functional change in genomes. Along with the duplication processes, point mutations may occur in coding regions or regulatory regions of a gene. Point mutations can disrupt the function of a coding region or regulatory region. Modelling these evolutionary processes can help us to understand how genomes can maintain or modify functions through the evolution of time.

This thesis considers two Markov models for the evolutionary processes of gene family duplication, referred to as the Detailed Binary Matrix (DBM) model and the Level-Dependent Quasi-Birth-Death (LD-QBD) model. We also develop an LD-QBD model for neofunctionalisation, in which we consider how genes can become associated to perform a joint function.

Chapter 1 introduces some related biological background and the main mathematical and statistical techniques we apply to model evolutionary processes. We review the relevant literature and discuss current gaps in knowledge.

In Chapter 2, we apply an application of a DBM model to the evolutionary process of gene duplication. We describe how a binary matrix can be used to record relevant information about a gene family, including the number of genes, the number of genes permitted to be lost and the number of functions in the gene family. We derive expressions for the rates of transition between states and the probabilities of corresponding events. This approach allows us to model how a gene can obtain a new function (neofunctionalisation) and how genes can specialise to perform only a subset of the original functions (subfunctionalisation). The DBM model has a large state space which means that, while it can be used for simulation, it is not suitable for mathematical analysis using the theory of Markov chains. Therefore, we next develop

Thesis submitted to the University of Tasmania in October 2022; degree approved on 27 March 2023; supervisors Barbara Holland and Malgorzata O'Reilly.

an LD-QBD model as an alternative way to model the same dynamics but in a more numerically efficient way. Also we illustrate how to approximate the DBM model applying the LD-QBD model. We construct several numerical examples to compare the qualitative behaviour of the DBM model and the LD-QBD model.

Subfunctionalisation occurs when different copies of genes maintain the ability to perform different functions. In Chapter 3, using the DBM model from Chapter 2, we explore how gene duplication followed by subfunctionalisation effects the shape and balance of gene trees. We describe how to generate a gene tree under the DBM model. Moreover, we analyse the conditions under which the process has a stationary tree size based on different rates of gene duplication, gene loss and function loss in a gene. We find conditions under which gene trees are more balanced compared to trees generated under the constant rate birth-and-death model.

Chapter 4 presents a joint paper [2] that applies matrix analytic methods to the study of both species trees, gene trees and their reconciliation. I contributed to the sections related to gene trees and species tree/gene tree reconciliation. In these sections, we extended the state space of the LD-QBD model from Chapter 2, so that it is possible to track tree balance. Based on a given species tree, we develop algorithms to compute the maximum likelihood reconciliation, which results in the most likely embedding of a gene tree in a species tree.

In Chapter 5, we propose a different view of neofunctionalisation compared to Chapter 2. Rather than modelling it as acquisition of new regulatory regions, we consider the case where multiple genes may become associated to perform some joint function. We develop two models of neofunctionalisation. In the first model, when two genes become associated at some point on a species tree, their subsequent rates of gain and loss are dependent; whereas in parts of the species tree unaffected by the neofunctionalisation, the genes are gained and lost independently. When two genes are associated and they are both present, the species can perform some new function. In the second more advanced model, we develop an LD-QBD model which considers the association among more than two genes. We assume that a new beneficial function is obtained when all the genes are present.

In Chapter 6, we discuss further work which has appeared in [1]. The key next step will be to obtain some biological datasets and fit our models to the data. For example, we could measure the tree balance for empirical gene trees and compare the results to predictions from the models in Chapter 3.

Overall, we have developed several new models to describe gene family evolution. We show how mathematical techniques from the area of matrix analytic methods and stochastic modelling can help us better understand evolutionary processes.

References

[1] J. Diao, M. M. O'Reilly and B. Holland, 'A subfunctionalisation model of gene family evolution predicts balanced tree shapes', *Mol. Phylogenet. Evol.* **176** (2022), Article no. 107566.

[2] J. Diao, T. L. Stark, D. A. Liberles, M. M. O'Reilly and B. R. Holland, 'Level-dependent QBD models for the evolution of a family of gene duplicates', *Stoch. Models* **36**(2), 285–311.

https://doi.org/10.1017/S0004972723000801 Published online by Cambridge University Press

JIAHAO DIAO, School of Mathematics and Statistics,
University of Melbourne, Parkville, Victoria 3010, Australia
e-mail: jiahao.diao@unimelb.edu.au

https://doi.org/10.1017/S0004972723000801 Published online by Cambridge University Press