Journal of Theoretical Biology 341 (2014) 111–122

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Revisiting the diffusion approximation to estimate evolutionary rates of gene family diversification

Erida Gjini [a,*], Daniel T. Haydon [c,d,e], J. David Barry [e], Christina A. Cobbold [b,d]

[a] Instituto Gulbenkian de Ciência Oeiras, Portugal
[b] School of Mathematics and Statistics, College of Science and Engineering, University of Glasgow, Glasgow, United Kingdom
[c] Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
[d] The Boyd Orr Centre for Population and Ecosystem Health, University of Glasgow, Glasgow, United Kingdom
[e] Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, University of Glasgow, Glasgow, United Kingdom

HIGHLIGHTS

- We model genetic diversification of a multi-gene family by explicitly simulating the evolutionary processes of point mutation and gene conversion.
- We link the stochastic dynamics of diversification to the Wright–Fisher model in population genetics and the diffusion approximation.
- We compare simulations and the diffusion approach across many relevant parameter regimes, showing a very good match for large family size, long gene sequences and small relative conversion tract length.
- We apply the diffusion approximation to estimate rates of diversification within the antigen gene family of African trypanosomes.

ARTICLE INFO

Article history:
Received 18 December 2012
Received in revised form
21 June 2013
Accepted 2 October 2013
Available online 11 October 2013

Keywords:
Identity distribution
Wright-Fisher model
Mutation rate
Gene conversion
Trypanosome antigen archive

ABSTRACT

Genetic diversity in multigene families is shaped by multiple processes, including gene conversion and point mutation. Because multi-gene families are involved in crucial traits of organisms, quantifying the rates of their genetic diversification is important. With increasing availability of genomic data, there is a growing need for quantitative approaches that integrate the molecular evolution of gene families with their higher-scale function. In this study, we integrate a stochastic simulation framework with population genetics theory, namely the diffusion approximation, to investigate the dynamics of genetic diversification in a gene family. Duplicated genes can diverge and encode new functions as a result of point mutation, and become more similar through gene conversion. To model the evolution of pairwise identity in a multigene family, we first consider all conversion and mutation events in a discrete manner, keeping track of their details and times of occurrence; second we consider only the infinitesimal effect of these processes on pairwise identity accounting for random sampling of genes and positions. The purely stochastic approach is closer to biological reality and is based on many explicit parameters, such as conversion tract length and family size, but is more challenging analytically. The population genetics approach is an approximation accounting implicitly for point mutation and gene conversion, only in terms of per-site average probabilities. Comparison of these two approaches across a range of parameter combinations reveals that they are not entirely equivalent, but that for certain relevant regimes they do match. As an application of this modelling framework, we consider the distribution of nucleotide identity among VSG genes of African trypanosomes, representing the most prominent example of a multi-gene family mediating parasite antigenic variation and within-host immune evasion.

1. Introduction

Quantifying the contributions of different evolutionary processes to the generation of genetic diversity is important to understand the evolution, adaptation and persistence of organisms. Key functions are often encoded by multi-gene families such as the major histocompatibility complexes (MHC) in man and mouse, the *Amy* multigene family of *Drosophila melanogaster*, and variable antigen genes of parasites such as *Plasmodium falciparum* and African trypanosomes. Typically multigene families contain genes that have arisen primarily via gene duplication, a driving force in molecular evolution (Ohno, 1970; Lynch and Conery, 2000; Bailey et al., 2002).

* Corresponding author at: Instituto Gulbenkian de Ciência, Oeiras, Portugal. Tel.: +351 214407968.
E-mail address: egjini@igc.gulbenkian.pt (E. Gjini).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.10.001

Gene duplication is then followed by gene conversion and point mutation, besides other processes such as unequal crossing over, recombination, random genetic drift and selection. Gene conversion is a special type of non-reciprocal transfer of genetic material in which one segment of DNA contributes genetic information to another, making the recipient location identical to the donor, but not altering the donor sequence. This process is very important for the concerted evolution of gene families and the functions they encode across organisms (Ohta, 2010). The combined effects of gene conversion and point mutation determine the diversification of duplicated genes, with gene conversion playing a major role in accelerating the spread of beneficial mutations through all gene family members.

Theoretical treatment of concerted evolution of multigene families presents many challenges, because the pattern of polymorphism in multigene families is much more complicated than that in single-copy genes. However some major advances using population genetic approaches were accomplished as early as the 1970s and 80s (Ohta, 1976, 1983; Nagylaki, 1984; Walsh, 1983), and modern approaches using the coalescent to understand multigene family complexity and evolution are increasing (see e.g. Griffiths and Watterson, 1990; Innan, 2002). Typical population genetic analyses focus on fixation probabilities of alleles and equilibrium identity coefficients under different scenarios (Mano and Innan, 2008; Innan, 2009). The picture of evolution gets complicated when the gene copy number is not constant in time (see Tachida and Kuboyama, 1998 for a gene duplication model), when gene conversion occurs in a biased or sequence dependent manner (Walsh, 1983, 1987), when mutation is biased and when selective forces are at play.

Although many characteristics of identity coefficients have been modelled, the temporal dynamics driving gene families towards such equilibria have usually received more minor attention, and simulation of idealized single-nucleotide events, rather than explicit whole genetic events has generally been adopted, with few exceptions (Innan, 2002). Because of analytical tractability, small multigene families with two copies of genes have been modelled more frequently, and distribution of allele frequencies between genomes rather than within genomes have been investigated. Depending on whether gene homology is studied by nucleotide identity or amino-acid identity, a K-allele model (Kimura and et al., 1968) or an infinite-allele model (Kimura and Crow, 1964) have been used respectively.

With the increasing availability of genetic data comes the challenge of quantifying the rates and characteristics of mutation, gene conversion, and other evolutionary forces that shape gene families from the molecular signatures they leave on DNA sequences. The rate and tract length of gene conversion between duplicated genes are among the most difficult parameters to infer. The empirical approach usually requires mutation accumulation experiments in transgenic model systems, while polymorphism (SNP) data is usually analyzed from a more theoretical standpoint, when DNA sequence data are available. More recently maximum likelihood methods have been proposed that overcome the limitation of estimates being model-dependent (Mansai et al., 2011). Empirical approaches for estimating tract length of gene conversions rely on identification of donor and recipient genes and involve the analysis of selected markers (see Song et al., 2011 for a recent review). In contrast, evolutionary data are not very informative for the tract length, mainly because of their dependence on the overall accumulation of footprints of historical gene conversions that potentially overlap with one another.

In this study, we investigate the dynamics of within-genome diversification of a multi-gene family as a result of only two recurring processes: point mutation and gene conversion among its members. We adopt two approaches, one based on simulation of discrete events and the other based on a diffusion approximation to extract information about the magnitude of genetic diversity attainable in a family of genes, and how it depends on the rates and characteristics of these evolutionary forces and on the family size. We analyze the role of various parameters, such as gene length, family size and conversion tract length in the distribution of pairwise identity in a gene family. We link the classical Wright–Fisher model (Fisher, 1930; Wright, 1931) to the dynamics of multigene family diversification, providing an avenue for further quantitative exploration of genomic evolution.

Finally, we apply our modelling framework to the nucleotide diversity of antigen (*VSG*) genes in African trypanosomes, to examine the interplay of gene conversion and point mutation within a group of related genes, representative of a multi-gene family that has originated through duplication. The overall distribution of genetic identity in this antigen gene family has two main implications for the fitness of the parasite: first, it interferes with higher-scale processes such as mosaic gene formation, often driven by identity-related recombination (Barbet and Kamper, 1993; Marcello and Barry, 2007a) involving pseudogenes; second, it can determine antigenic cross-reactivity between parasite variants that appear sequentially in infections and are targeted by the host immune system. Applying the diffusion approximation to the empirical genetic identity distribution of this multi-gene family, we lay a new bridge between mathematical theory and parasite genetic data, and are able to extract the rates of the evolutionary processes that can shape antigen gene diversification.

2. Modelling framework

To model the evolution of pairwise genetic identity in a multigene family we first consider all conversion and mutation events as they happen, keeping track of the donors and times of their occurrence; then we consider only the infinitesimal effect of these processes on pairwise identity accounting for random sampling of genes and positions. The first approach is purely stochastic, based on many explicit parameters, such as conversion event rate and mutation rate per unit of time, as well as conversion tract length, gene length and gene number, and it serves to visualize exact trajectories of the system of genes. The second approach is an approximation of the biological stochastic process, implicitly taking into account the characteristics of point mutation and gene conversion, but depending basically on just three parameters: mutation and conversion probabilities per base pair per generation and gene length, which makes it more amenable to analytic treatment.

2.1. Stochastic simulation of genetic events

Consider a population of N genes, each of length L, subject to gene conversion between pairs of genes and random point mutation. The state of a gene is represented by an array of L integers, corresponding to the 4 nucleotide types (A–C–T–G), in line with the K-allele ($K=4$) model. At time 0 a random initial sequence of length L is generated and applied to all genes, making them identical. We model the stochastic occurrence of single genetic events in such a multigene family as a Poisson process, which we simulate using the Gillespie Algorithm (Gillespie, 1977). The rate at which a gene is converted per unit of time is denoted by γ, while the rate at which mutations occur is given by μ. The global event rates of the two processes per unit of time are γN and μN. Stochastic events (mutations and conversions) are indexed $1,2,\ldots,T\in\mathbb{Z}$, which occur at the times $t_1,t_2,\ldots,t_T\in\mathbb{R}$. The inter-event times are exponentially distributed with mean $1/(\gamma N+\mu N)$. Only one event can happen at a time. Point mutation is chosen