Journal of Theoretical Biology 378 (2015) 56–64

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A coarse-grained biophysical model of sequence evolution and the population size dependence of the speciation rate

Bhavin S. Khatri [a,b,*], Richard A. Goldstein [b]

[a] The Francis Crick Institute, Mill Hill Laboratory, The Ridgeway, London NW7 1AA, UK
[b] Division of Infection & Immunity, University College London, London WC1E 6BT, UK

HIGHLIGHTS

- We explore a biophysical mechanism of speciation.
- We develop a novel coarse-grained stochastic dynamics of sequence evolution.
- Analytical results show that hybrid binding energies diffuse neutrally.
- Sequence entropy and drift poise common ancestors closer to incompatible regions.
- So as population size decreases hybrid incompatibilities arise more quickly.

ARTICLE INFO

Article history:
Received 3 October 2014
Received in revised form
20 February 2015
Accepted 20 April 2015
Available online 30 April 2015

Keywords:
Dobzhansky Muller incompatibilities
Evolution
Genotype phenotype map
Sequence entropy
Free fitness

ABSTRACT

Speciation is fundamental to understanding the huge diversity of life on Earth. Although still controversial, empirical evidence suggests that the rate of speciation is larger for smaller populations. Here, we explore a biophysical model of speciation by developing a simple coarse-grained theory of transcription factor-DNA binding and how their co-evolution in two geographically isolated lineages leads to incompatibilities. To develop a tractable analytical theory, we derive a Smoluchowski equation for the dynamics of binding energy evolution that accounts for the fact that natural selection acts on phenotypes, but variation arises from mutations in sequences; the Smoluchowski equation includes selection due to both gradients in fitness and gradients in sequence entropy, which is the logarithm of the number of sequences that correspond to a particular binding energy. This simple consideration predicts that smaller populations develop incompatibilities more quickly in the weak mutation regime; this trend arises as sequence entropy poises smaller populations closer to incompatible regions of phenotype space. These results suggest a generic coarse-grained approach to evolutionary stochastic dynamics, allowing realistic modelling at the phenotypic level.

1. Introduction

Speciation underlies the diversity of life on Earth today. Yet the detailed genetic mechanisms by which distinct species arise are still largely not understood. Darwin (1859), despite the title of his magnus opus, struggled to understand how natural selection could give rise to hybrid inviability or infertility. If the hybrid inviability were due to a single locus, how could two species evolve from a common ancestor, as one of these species would have to evolve past an inviable heterozygotic state. A resolution came with the understanding that epistatic (non-linear) interactions between different loci can give rise to the so-called Dobzhansky–Muller incompatibilities (DMI) between independently evolving lineages (Dobzhansky, 1936; Muller, 1942; Bateson, 1909; Gavrilets, 2004). For example, two lineages evolving independently through geographic isolation (allopatric evolution) from a common ancestor *ab* can fix the genotypes *aB* and *Ab*, yet the hybrid genotype *AB* may be inviable. Through a similar mechanism incompatibilities can arise in polygenic systems, where the effective contribution to fitness of the many loci coding a quantitative trait fitness is epistatic. Even if the loci contribute additively to the trait, stabilising selection (usually modelled as quadratic) on a trait value induces epistasis. Populations diverge, under the action of drift, by shifting between different stable equilibria that encode the same optimal trait value, but with different allelic combinations (Wright, 1935a,b); when combined in hybrids this can lead to hybrid incompatibilities (Barton, 1989). Field data (Coyne and Orr,

* Corresponding author at: The Francis Crick Institute, Mill Hill Laboratory, The Ridgeway, London NW7 1AA, UK.
E-mail addresses: bhavin.khatri@physics.org (B.S. Khatri), r.goldstein@ucl.ac.uk (R.A. Goldstein).

http://dx.doi.org/10.1016/j.jtbi.2015.04.027

2004; Mayr, 1963) and specific introgression studies (Wu and Beckenbach, 1983; Vigneault and Zouros, 1986) suggest that the most dominant form of speciation involves the generation of hybrid incompatibilities in geographically isolated populations with no or very little gene flow.

The development of quantitative models that can predict speciation rates will allow better understanding of the different factors that maintain bio-diversity along with the processes of extinction and environmental change (Coyne and Orr, 2004; Rosenzweig, 2001). An important aspect of such models is the dependence of speciation rate on population size. Although, the question of a population size dependence of the rate of speciation has received little empirical attention and there have yet to be any definitive studies, there is indirect evidence that the rate of speciation is higher in smaller populations (Santos and Salzburger, 2012; Mayr, 1970; Glor et al., 2004), including the large species diversity of fish in the East African Great Lakes (Owen et al., 1990) compared to marine animals (Mayr, 1970, 1954; Rubinoff and Rubinoff, 1971) and birds (Fitzpatrick, 2004) which have large ranges and populations sizes, and the population size dependence observed in net diversification rates inferred from phylogenetic trees (Coyne and Orr, 2004; Nee, 2001; Barraclough and Nee, 2001). Strikingly, although cichlid fishes in Lake Malawi, whose effective population sizes are of order 100–10 000 (Oppen et al., 1997; Fiumera et al., 2000), develop reproductive isolation within 1–10 Myr after divergence (Stelkens et al., 2010), domestic chickens (*Gallus gallus*) can still hybridise with helmeted guineafowl (*Numida meleagris*) after roughly 55 Myr divergence (Cooper and Penny, 1997), potentially reflecting the large effective population size of domestic chickens estimated to range between 10^5 and 10^6 (Sawai et al., 2010).

Models of speciation that require positive selection to drive divergence are unlikely to be able to explain these trends as larger populations take less time to fix beneficial mutants and so evolve more quickly (Gavrilets, 2003). Founder event or peak shift models where reproductive isolation arises when a small population passes through a fitness valley could explain this trend, as the rate of valley crossing increases at small population sizes (Lande, 1979, 1985; Barton and Charlesworth, 1984; Barton and Rouhani, 1987). However, these models require a small fitness valley to give speciation on realistic timescales, meaning that the reproductive isolation this model seeks to explain is generally destroyed. In the strong mutation regime (mutation rate large relative to the inverse population size), polymorphisms will be common, and the larger variation found in larger populations is predicted to result in a slower average substitution rate, reducing the rate of speciation (Gavrilets, 1999; Nei et al., 1983). Polygenic models of divergence of additive traits under stabilising selection, also in the strong mutation regime, predict that smaller populations can shift between stable equilibria more quickly, leading to more rapid isolation (Barton, 1989). More recently, sequence-level simulations of protein–DNA binding similar to the model we examine here, showed in the intermediate to strong mutation regime, that hybrid fitness decayed more rapidly for smaller populations (Tulchinsky et al., 2014); however, the underlying mechanism or growth of DMIs was not explored. Despite these results in the strong mutation regime, many traits involved in speciation are found to be monogenic or oligogenic (involving only one or a few loci) (Orr, 2001) and so are expected to arise in monomorphic populations in the weak mutation regime. In this respect, Orr constructed a model that considers the combinatorics of how potential incompatibilities grow between two independent lineages in the weak mutation regime. For pair-wise interactions between loci this growth is quadratic in the number of substitutions by which they are separated (Orr, 1995; Orr and Turelli, 2001); however, the model assumes that populations diverge neutrally and so predicts no population size dependence. To summarise, although theory predicts that in the strong mutation regime we would expect a slower rate of accumulation of DMIs for larger populations, there are no theories of speciation that predict this population size effect in the very relevant weak mutation regime.

In this paper, we examine the process of how incompatibilities arise in allopatry for a biophysical model of a transcription factor binding to DNA by developing a coarse-grained model of how the transcription factor protein and DNA sequences co-evolve within a stochastic dynamics framework. Our key innovation is to develop a general equation of phenotypic evolution in the weak mutation regime, which accounts for the fact that selection acts on phenotypes, but variation in phenotype arises from mutations in sequence through the mapping of genotype to phenotype. In particular, we need to include the number of sequences corresponding to a particular phenotype, the log of which we call the "sequence entropy" in analogy to statistical mechanics entropy. This approach normally gives rise to an often intractable master equation. By considering the continuous limit, however, we can convert the master equation into a diffusion equation called the Smoluchowski equation, which includes selection, sequence entropy, and random drift. By including the effects of sequence entropy, the stochastic dynamics framework we present allows investigation of the effect of population size on evolution in the weak mutation regime, including its role in speciation dynamics. Our work differs from previous diffusion-based models of phenotypic evolution, such as Lande (1976), by considering a generic genotype–phenotype map and also in focusing on the weak mutation regime where we can ignore polymorphisms and restrict our attention to movement between monomorphic genotypes. What we find is a picture of speciation different from that of the Orr model, in that it features a latency in the development of DMIs as hybrid populations need a finite time to reach incompatible regions of phenotype space. Importantly, the model predicts a higher rate of speciation in smaller populations in the weak mutation regime, providing an explanation for the trend seen in the observations described above.

Gene expression divergence has been shown to be a major factor in driving differences between species (King and Wilson, 1975; Wolf et al., 2010; Wray, 2007; Wittkopp et al., 2008), and there is direct evidence of speciation driven by the evolution of genes related to transcription factors in *Drosophila* (Ting et al., 1998; Brideau et al., 2006). Thus the binding of transcription factors to DNA to control gene expression is arguably one of the most important co-evolving systems for organisms and crucial for their correct development, making them an ideal case study for a biophysical model of speciation. However, despite our focus on transcription factor binding, the model is in fact very generic and could form the basis for the co-evolution of a number of interacting macromolecules including protein–protein interactions, antibody–antigen interactions, or the interaction of genes expressed by nucleus and mitochondria.

We first derive a diffusion equation (Smoluchowski equation) for studying the coarse-grained stochastic evolutionary dynamics of co-evolving sequences, and then adapt this model to the case of two interacting genes represented by the binding of a transcription factor to a region of DNA. We then consider two populations evolving independently from a common ancestor, and consider the viability of reproductive crosses between these populations.

2. A Smoluchowski equation for evolutionary stochastic dynamics

Natural selection acts on phenotypes. In general, however, many genotypes code for the same phenotype (Fontana, 2002;