Mathematical Biosciences 278 (2016) 1–4

Contents lists available at ScienceDirect

Mathematical Biosciences

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

The probability of speciation on an interaction network with unequal substitution rates

Peter Olofsson[a,b,*], Kevin Livingstone[c], Joshua Humphreys[c], Douglas Steinman[a]

[a] Department of Mathematics, Trinity University, United States
[b] School of Engineering, Jönköping University, Sweden
[c] Department of Biology, Trinity University, United States

ARTICLE INFO

Article history:
Received 5 March 2016
Revised 13 April 2016
Accepted 24 April 2016
Available online 10 May 2016

Keywords:
Bateson-Dobzhansky-Muller interactions
Interaction networks
Reproductive incompatibility
Speciation

ABSTRACT

Speciation is characterized by the development of reproductive isolating barriers between diverging groups. A seminal paper of a mathematical model of speciation was published by Orr (1995), extended by Livingstone et al. (2012) to incorporate interaction networks. Here, we further develop the model to take into account the possibility of different substitution rates for network nodes of different connectivity. Mathematically, this amounts to sampling nodes from an undirected graph where the inclusion probability for a given node depends on its degree (number of connecting edges). We establish formulas for the rate of speciation and identify a crucial parameter that is a measure of the deviation from simple random sampling.

1. Introduction

The Bateson–Dobzhansky–Muller (BDM) model describes how fixation of mutations in allopatric populations could produce inviability or sterility in hybrid offspring, without the mutations lowering fitness within either population. Briefly, the BDM model starts with an ancestral (diploid) population of genotype aabb; in one population, the A allele arises and becomes fixed, while in the other population, B arises and is fixed. The resulting hybrid from the AAbb aaBB cross would have genotype AaBb, and as A and B have never been "tested" together, they could interact epistatically to cause a genetic incompatibility. The accumulation of such Bateson–Dobzhansky–Muller incompatibilities (BDMIs) can cause permanent isolation, and hence speciation.

In [7], Orr introduces a mathematical model for the rise of BDMIs. In this model, two diverging lineages fix alleles at K loci between them, and each new allele has a probability p of being incompatible with an allele (derived or ancestral) from the other lineage at one of the $K-1$ loci at which substitutions have occurred. It is demonstrated how the expected number of incompatibilities increases as a function of K^2, a phenomenon referred to as "snowballing," (see [6]) and how the probability of speciation also depends on K through K^2. Here and subsequently, speciation is simply defined as the occurrence of at least one BDMI.

In [5], we elaborated upon Orr's model by recognizing that most nodes in a real biological network are connected to only a small number of other nodes, while very few nodes act as central hubs with a large number of interactions (see [2]). We modeled an interaction network as an undirected graph where nodes are loci and edges are existing interactions, each edge leading to a BDMI with probability p. The crucial parameters of the network turn out to be the number of nodes, N, and the number of edges, N_E, which are used to define the density $\alpha = N_E/\binom{N}{2}$ (Orr's model would thus correspond to a complete graph with $N_E = \binom{N}{2}$ and $\alpha = 1$).

The rise of K substitutions is modeled by randomly sampling K nodes from the network, resulting in a subgraph where each edge (existing interaction) may lead to a BDMI, with probability p. Thus, the more edges, the more likely a BDMI and also note that the number of edges is a random variable with range $\{0, 1, \ldots \binom{K}{2}\}$, whereas in Orr's model we always get $\binom{K}{2}$ edges. It is demonstrated that, to a first order of approximation, the formula for the probability of speciation contains α as a parameter but is otherwise similar to Orr's. Effectively, the parameter p (probability of a single BDMI) is replaced by the product αp.

The sampling in [5] is done randomly, that is, every node is equally likely to be included in the sample of size K. Biologically, this translates into all alleles being equally likely to become substituted; however, as is evident from the literature on the subject, this assumption is questionable. For example, in [3], it is observed that the connectivity of well-conserved proteins in the protein–protein interaction (PPI) network for the yeast *S. cerevisiae* is

* Corresponding author.
E-mail address: polofsso@trinity.edu (P. Olofsson).

http://dx.doi.org/10.1016/j.mbs.2016.04.010
0025-5564/

negatively correlated with their rate of evolution, and [4] reports similar findings for yeast, and also for two other species. Indeed, in [4] it is pointed out that there is a "consistent reduction in evolutionary rate for essential proteins in all three species: essential genes in the protein interaction network evolved at 70% the rate of nonessential genes (yeast: 70.5%; worm: 71.4%; fly: 70.1%)". To allow for this possibility, we now extend our model from random sampling to sampling where a node has an inclusion probability that may depend on its degree (number of interactions). Although the biological motivation is the aforementioned negative correlation, our model is not restricted to such assumptions but allows for any kind of dependence of sampling probabilities on connectivity.

2. Mathematical model

We consider a network with N nodes such that there are N_i nodes of degree i for $i = 1, 2, \ldots$, where we note that the theoretical upper limit for i is $N-1$. We define the *degree distribution* as the sequence $(q(1), q(2), \ldots)$ where

$$q(i) = \frac{N_i}{N} \tag{1}$$

so that the *mean degree* is

$$\mu = \sum_i iq(i) \tag{2}$$

which can be thought of as the expected degree of a randomly chosen node. Let N_E denote the number of edges in the network and define the *density* of the network as

$$\alpha = \frac{N_E}{\binom{N}{2}} \tag{3}$$

As every edge connects to 2 nodes, we get the relation

$$2N_E = \sum_i iN_i \tag{4}$$

and combining this with (1) and (2) yields

$$2N_E = N\mu \tag{5}$$

and as $\binom{N}{2} \approx N^2/2$, we also note that

$$\mu \approx N\alpha \tag{6}$$

relations we will make use of later.

Now sample K nodes in a way such that a given node of degree i has a probability $f(i)$ of being chosen when picking a single node. The case of random sampling, as in [5], corresponds to a uniform distribution: $f(i) = 1/N$ for all i. Let D be the degree of a node chosen according to the $f(i)$ and note that, as there are $Nq(i)$ nodes of degree i, we get

$$P(D = i) = Nq(i)f(i) \tag{7}$$

and hence expected value

$$E[D] = \sum_i iNq(i)f(i) \tag{8}$$

As this is the mean when choosing according to the $f(i)$, and μ is the mean when choosing randomly, we define the ratio

$$r = \frac{E[D]}{\mu} \tag{9}$$

which can be thought of as a measure of deviation from random sampling. Note that random sampling gives $E[D] = \mu$ so that $r = 1$. If nodes of higher degrees are less likely to be chosen (less prone to substitution), we typically get $r < 1$.

3. Results

Let X be the number of edges that we get when sampling K nodes according to the $f(i)$. The key observation used to compute both speciation probability and expected time until speciation is the following proposition, proved in the Appendix:

Proposition 1.

$$E[X] \approx \frac{K^2}{2} r^2 \alpha$$

Note: With random sampling, as in [5], the result is that

$$E[X] = \binom{K}{2}\alpha \tag{10}$$

$$\approx \frac{K^2}{2}\alpha \tag{11}$$

which is the case when $r = 1$. In [7], both r and α equal 1 and $X \equiv \binom{K}{2}$.

It is interesting to notice that, at least to a first-order approximation, the dependence of $E[X]$ on the network is light, in the sense that the network itself enters only via the density α and the sampling only via the parameter r. Thus, there are many different network topologies and many different sampling schemes (substitution mechanisms) that can lead to the same speciation rate.

Other than the usual snowballing – quadratic rather than linear dependence of $E[X]$ on K – we now also notice a quadratic dependence on r, which if $r < 1$ amounts to a "snowballing in reverse," indicating that the speciation rate might be sensitive already to small reductions in r.

A formula for the probability of speciation (at least one incompatibility) after K substitutions now follows from Proposition 1:

Proposition 2. *Let p denote the probability of an incompatibility and let S denote the event of speciation after K substitutions. Then*

$$P(S) \approx 1 - (1-p)^{K^2 r^2 \alpha/2}$$

To investigate the effect of the sampling scheme on the speciation rate, we compare $P(S)$ for a few different shapes of the $f(i)$. In general, let

$$f(i) = cr(i) \tag{12}$$

where r is a function determining the general shape of f and c is a normalizing constant. By (7)

$$c = \frac{1}{N\sum_i q(i)r(i)} \tag{13}$$

which by (8) gives the expected degree as

$$E[D] = \frac{\sum_i iq(i)r(i)}{\sum_i q(i)r(i)} \tag{14}$$

whence we can compute the speciation probability by (9) and Proposition 2.

As an example, we will again use the PPI network for *S. cerevisiae* and investigate 3 general shapes of a decreasing f: linear, polynomial, and exponential. As the yeast network appears to be well described by a power law (see [9]), we take

$$q(i) = ai^{-b}, \quad i = 1, \ldots, N-1 \tag{15}$$

where the parameter b gives the exact shape and a is a normalizing constant. Different values of b in the range 1.5–2.5 have been reported (see [1]) and as we are mainly concerned with differences

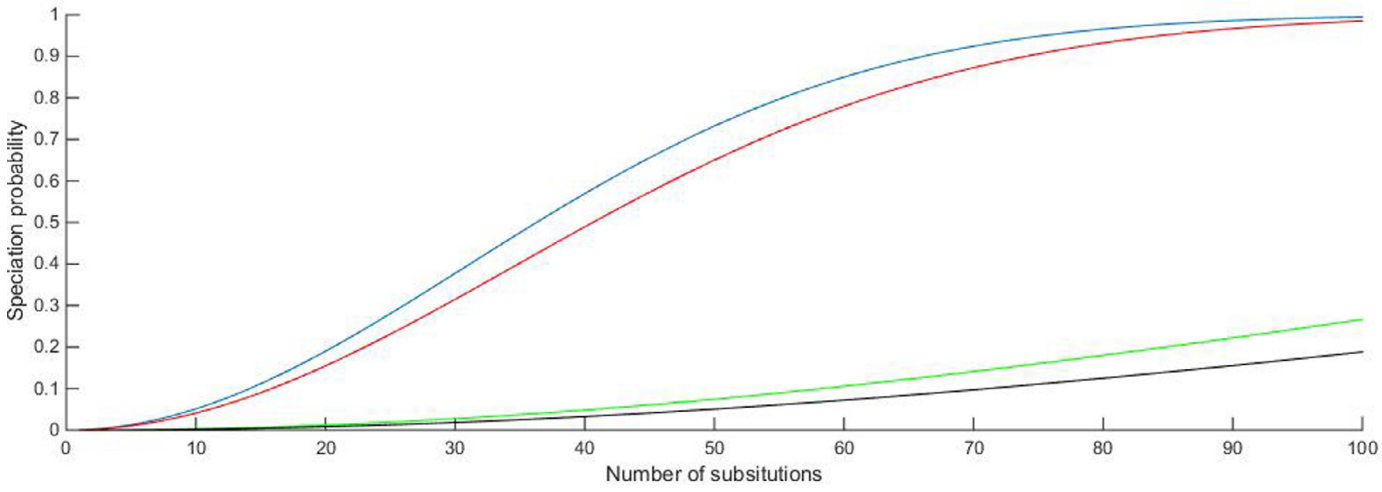

Fig. 1. The probability of speciation after K substitutions under different sampling distributions. From top to bottom: uniform, linear, polynomial, exponential.

in the sampling distribution, we simply took $b = 2$ which gives $a \approx 0.61$ (as N is so large – about 6000, see [8] – the sum of the $q(i)$ is very close to the well-known infinite series $\sum_{i=1}^{\infty} i^{-2} = \pi^2/6$). We use the following 3 shape functions:

$$r(i) = \begin{cases} N - i \text{ (linear)} \\ 1/i \text{ (polynomial)} \\ e^{-i} \text{ (exponential)} \end{cases}$$

For comparison, we also include the case of uniform sampling (corresponding to $r(i) \equiv 1$). See Fig. 1. It is clear from these figures that assuming uniform sampling may overestimate the rate of speciation, possibly by a lot. Our current model which allows for degree dependence corrects this overestimation.

We conclude with a result about the time until speciation, also proved in the Appendix.

Proposition 3. *Let T be the time until speciation, that is, the first mutation that leads to an incompatibility. Then*

$$E[T] \approx \frac{1}{r}\sqrt{\frac{\pi}{2p\alpha}}$$

Again, note the appearance of α and r as the only dependence on the network topology. To compare with previous models, in [5], $r = 1$, and in [7], both r and α equal 1. As discussed in [5], this formula can be used to estimate p based on observed values of T. If random sampling is assumed, $r = 1$, so if there is indeed a negative correlation between connectivity and rate of evolution as discussed earlier, p will be underestimated as the true r is less than 1.

Acknowledgments

We gratefully acknowledge financial support from the Howard Hughes Medical Institute through a grant to Trinity University, the NSF through Grant UBM-0926701, and the NIH through Grant 2R15GM093957-02. PO wishes to express his gratitude to the Department of Mathematics, Physics, and Chemical Engineering at Jönköping University, Sweden, where this work was finalized during the author's academic leave.

Appendix A. Proofs

A1. Proof of Proposition 1

Let N_{ij} denote the number of node pairs where one node is of degree i, and the other of degree j. Also, let $P(i, j)$ be the probability that a specific node pair of degree (i, j) is chosen. Then

$$E[X] = \sum_{i,j} N_{ij} P(i, j)$$

where the summation is over all possible degrees. For N_{ij}, note that there are $Niq(i)$ edges emanating from nodes of degree i, and each such edge connects to a node of degree j with a probability that is proportional to $jq(j)$, thus equal to $jq(j)/\mu$ (the so-called "size-biased" distribution). Thus

$$N_{ij} \approx \frac{Niq(i)jq(j)}{2\mu}$$

where the "2" appears due to the double counting of node pairs.

To get an expression for $P(i, j)$, we make the simplifying assumption that sampling is done *with* replacement rather than without; an assumption that is clearly not correct but as long as N is large and K is small in comparison with N, this simplification is justifiable. Also, by the same argument, we consider the events that the nodes of degree i and j are chosen to be independent. The probability to fail to get the node of degree i in K trials is then $(1 - f(i))^K$ and we get

$$P(i, j) \approx (1 - (1 - f(i))^K)(1 - (1 - f(j))^K)$$

and hence

$$E[X] \approx \sum_{i,j} \frac{Niq(i)jq(j)}{2\mu}(1 - (1 - f(i)^K)(1 - (1 - f(j))^K$$

$$= \frac{N}{2\mu}\left(\sum_i iq(i)(1 - (1 - f(i))^K)\right)^2$$

Next, a Taylor expansion yields

$$1 - (1 - f(i))^K \approx 1 - e^{-Kf(i)} \approx Kf(i)$$

and recalling that $\mu \approx N\alpha$, we get

$$E[X] \approx \frac{K^2}{2N\mu}\left(\sum_i Niq(i)f(i)\right)^2 = \frac{K^2}{2}\frac{E[D]^2}{\mu^2}\alpha$$

and the proof is complete.

A2. Proof of Proposition 2

Let S be the event of speciation. With X as above, the conditional probability of speciation equals

$$P(S|X) = 1 - (1-p)^X$$

and as

$$P(S) = E[P(S|X)]$$

Proposition 7 together with a first-order Taylor approximation of the type

$$E[g(X) \approx g(E[X])$$

yields the result.

A3. Proof of Proposition 3

Proof. By a well-known formula for expected values of non-negative integer valued random variables,

$$E[T] = \sum_{K\geq 0} P(T > K)$$

where

$$P(T > K) \approx (1-p)^{E[X_K]}$$

Now, letting

$$c = \frac{1}{2}r^2\alpha$$

we have $E[X_K] \approx cK^2$ and get

$$\begin{aligned} E[T] &= \sum_{K\geq 0} P(T > K) \\ &\approx \sum_{K\geq 0} (1-p)^{cK^2} \\ &\approx \sum_{K\geq 0} e^{-pcK^2} \\ &\approx \int_0^\infty e^{-pcx^2}dx \\ &= \sqrt{\frac{\pi}{4pc}} \end{aligned}$$

and inserting the expression for c yields the result. □

References

[1] F. Chung, L. Lu, G. Dewey, D.J. Galas, Duplication models for biological networks, J. Comput. Biol. 10 (5) (2003) 677.
[2] M. Costanzo, The genetic landscape of a cell, Science 327 (5964) (2010) 425.
[3] H.B. Fraser, A.E. Hirsh, L. M.Steinmetz, C. Scharfe, M.W. Feldman, Evolutionary rate in the protein interaction network, Science 296 (2008) 750.
[4] M.W. Hahn, A.D. Kern, Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks, Mol. Biol. Evol. 22 (4) (2005) 803.
[5] K. Livingstone, P. Olofsson, G. Cochran, A. Dagilis, K. MacPherson, K. Seitz, A stochastic model for the development of Bateson-Dobzhansky-Muller incompatibilities incorporating protein interaction networks, Math. Biosci. 238 (2012) 49.
[6] D.R. Matute, I.A. Butler, D.A. Turissini, J.A. Coyne, Test of the snowball theory for the rate of evolution of hybrid incompatibilities, Science 329 (5998) (2010) 1518.
[7] H.A. Orr, The genetics of speciation: the evolution of hybrid incompatibilities, Genetics 139 (1) (1995) 1805.
[8] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, M. Tyers, BioGRID: a general repository for interaction datasets, Nucl. Acids Res. 34 (2006). (Database issue): D535.
[9] A. Wagner, The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, Mol. Biol. Evol. 18 (7) (2001) 1283.