J. Math. Biol. (1986) 24: 353-360

Journal of
Mathematical
Biology

A model of weak selection in the infinite alleles framework

E. D. Rothman and N. C. Weber*

Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

Abstract. Ewens (1972) proposed a model in the infinite allele framework for populations with neutrality of all alleles at a particular locus. This paper proposes a generalisation of Ewens' result for situations where there is a form of weak selection. The models considered here are continuous time, discrete state space Markov processes.

Key words: Ewens sampling formula — Infinite allele models — Mutation — Weak selection — Markov process

1. Introduction

The advent of a variety of biochemical techniques to distinguish between alleles at a locus focused attention on the neutralist versus selectionist hypotheses as explanations of genetic variability. Models of these hypotheses have also received considerable attention in recent years. An extensive bibliography and description of these models may be found in Ewens (1979). Of particular concern to us here are the so-called infinite alleles models. These models require each mutant allele to be a new (unique) allele.

Among the collection of models proposed as models for neutrality a model of Ewens (1972) provides a prototype. Ewens sampling distribution is based on four assumptions:

(i) neutrality of all alleles at a locus;
(ii) a fixed population size that is large compared with the sample size;
(iii) a stationary stochastic process of mutation and drift; and,
(iv) a potentially infinite number of alleles, where only unique alleles result from mutation.

Though other collections of assumptions have been studied, results similar to Ewens sampling formula emerge. Notable in this regard are the models of

* *Permanent address*: Department of Mathematical Statistics, University of Sydney, N.S.W. 2006, Australia

Karlin and McGregor (1966), Watterson (1974, 1976), Kingman (1977a, b) and Rothman and Templeton (1980). Indeed, a special case of our model provides yet another approach to Ewens formula.

Models of symmetric selection have also been studied elsewhere. Gillespie (1977) shows that a model of selection in a random environment yields Ewens sampling formula too. Thus any test of neutrality may have little power for certain forms of selection. On the other hand Watterson (1977) describes an alternative form of selection which does produce an alternative model.

Our approach is via a continuous time, discrete state space Markov process. Each class of alleles is assumed to evolve according to a birth and death process. Alleles in a given class are assumed to be selectively neutral, so they have the same birth and death rates. But alleles in different classes have different birth and death rates. The classes are then coupled by assuming that mutants arising from births in one class may belong to another class. Though the number of classes is assumed fixed the number of possible alleles within any class may be infinite.

Our purpose in introducing this model is two-fold. On one hand, by providing a framework for a future study of hypothesis tests, in the spirit of Neyman and Pearson we hope to shed some light on the neutralist versus selectionist debate. The other motivating factor involves an ongoing study of mutation rates in man. Since indirect estimates are based on neutrality, among other assumptions, an investigation of this sort will allow us to describe the impact of departures of certain assumptions on our estimators. In particular, data collected in Neel and Rothman (1981) indicates a higher frequency of rare variant alleles in tribal populations than in civilized populations. To investigate this difference without accepting a higher mutation rate in tribal populations requires such a framework.

2. The basic model

A simple linear birth and death process with immigration was first proposed as a model for allele behaviour by Karlin and McGregor (1966). Our model extends their process to several classes with possibly different birth and death rates. Each class in isolation is a Karlin and McGregor model but in our model the immigration process is a transfer of mutant alleles from one class to another.

Consider the following modification of the classical infinite allele birth-death process. Alleles are divided into C classes with birth and death rates in the jth class being b_j and d_j respectively. Let $\omega_j = b_j/d_j$. If a birth occurs in class j then with probability q the birth is a mutation to a new allele which we assume is a member of one of the existing C classes. The conditional probability that a mutant, whose parent is in class j, belongs to class i is u_{ji}, $\sum_{i=1}^{C} u_{ji} = 1$.

Let $G_j(t, i)$ denote the number of alleles with i copies in class j at time t, let $M_j(t) = \sum_{i=1}^{\infty} iG_j(t, i)$ be the total number of individuals in class j at time t and let $K_j(t) = \sum_{i=1}^{\infty} G_j(t, i)$ denote the total number of different alleles present in class j at time t. The moment generating function (m.g.f.) for $\{G_j(t, i)\}$ is

$$H(\boldsymbol{\theta}, t) = E\left[\exp \sum_{j=1}^{C} \sum_{i=1}^{\infty} \theta_{ij} G_j(t, i)\right]$$

and $H(\boldsymbol{\theta}, t)$ satisfies the differential equation

$$\begin{aligned}\frac{\partial H}{\partial t} &= \sum_{j=1}^{C} \sum_{i=1}^{\infty} [b_j pi \exp(\theta_{i+1,j} - \theta_{ij}) - 1] \frac{\partial H}{\partial \theta_{ij}} \\ &\quad + \sum_{j=1}^{C} \sum_{i=2}^{\infty} d_j i[\exp(\theta_{i-1,j} - \theta_{ij}) - 1] \frac{\partial H}{\partial \theta_{ij}} \\ &\quad + \sum_{j=1}^{C} d_j(e^{-\theta_{1j}} - 1) \frac{\partial H}{\partial \theta_{1j}} + q \sum_{j=1}^{C} (e^{\theta_{1j}} - 1) \sum_{r=1}^{C} b_r u_{rj}. \\ &\quad \cdot E\left[M_r(t) \exp\left(\sum_{j=1}^{C} \sum_{i=1}^{\infty} \theta_{ij} G_j(t, i) \right) \right], \end{aligned} \tag{1}$$

where $p = 1 - q$ and $\partial H / \partial \theta_{ij} = E[G_j(t, i) \exp \sum_{j=1}^{C} \sum_{i=1}^{\infty} \theta_{ji} G_j(t, i)]$.

Setting $\theta_{ij} = i\theta_j$ in (1) we obtain the following differential equation for the moment generating function of $\{M_j(t)\}$,

$$\begin{aligned}\frac{\partial L}{\partial t} &= \sum_{j=1}^{C} \frac{\partial L}{\partial \theta_j} [b_j p(e^{\theta_j} - 1) + d_j(e^{-\theta_j} - 1)] + q \sum_{j=1}^{C} (e^{\theta_j} - 1) \sum_{r=1}^{C} b_r u_{rj} \frac{\partial L}{\partial \theta_r} \\ &= \sum_{j=1}^{C} \frac{\partial L}{\partial \theta_j} \left[(e^{\theta_j} - 1)\{b_j p + b_j q u_{jj} - d_j\, e^{-\theta_j}\} + q b_j \sum_{r \neq j} (e^{\theta_r} - 1) u_{jr} \right], \end{aligned} \tag{2}$$

where $L(\boldsymbol{\theta}, t) = E[\exp \sum_{j=1}^{C} \theta_j M_j(t)]$.

Restricting attention to the marginal m.g.f. of $M_j(t)$, we get

$$\begin{aligned}\frac{\partial}{\partial t} E\, e^{\theta M_j(t)} &= \left\{ \frac{\partial}{\partial \theta} E\, e^{\theta M_j(t)} \right\} (e^{\theta} - 1)[b_j p + b_j q u_{jj} - d_j\, e^{-\theta}] \\ &\quad + (e^{\theta} - 1) q \sum_{r \neq j} b_r u_{rj} E M_r(t)\, e^{\theta M_j(t)}. \end{aligned} \tag{3}$$

Thus from (2) and (3) we see that a necessary condition for the $M_j(t)$ to be independent random variables is

$$\sum_{j=1}^{C} (e^{\theta_j} - 1) \sum_{r \neq j} b_r u_{rj} \{ E M_r(t) - E(M_r(t)\, e^{\theta_r M_r(t)}) / (E\, e^{\theta_r M_r(t)}) \} = 0$$

which is trivially satisfied if $u_{jj} = 1$, $j = 1, \ldots, C$, but is not true in general.

If there is only one class of alleles, so that $C = 1$, then the model reduces to the classical linear birth and death process discussed, for example, in Bailey (1964). This process becomes extinct with probability one if $\omega_1 \leq 1$, and if $\omega_1 > 1$ then the population becomes extinct with probability $\omega_1^{-M_0}$, or explodes with probability $1 - \omega_1^{-M_0}$ where M_0 is the initial population size. Thus even if $\omega_1 = 1$, the system does not settle down to a stable equilibrium distribution.

By extending the linear birth and death model to incorporate several classes with different birth and death rates we hope to establish a model that, at least under certain conditions on the ω_i and u_{ij} parameters, will yield a suitable stable equilibrium population.

If we differentiate (2) with respect to θ_j and set $\theta_1 = \cdots = \theta_C = 0$, we get

$$\frac{d}{dt} E M_j(t) = E M_j(t)\, d_j(\omega_j p - 1) + q \sum_{r=1}^{C} E M_r(t) b_r u_{rj}. \tag{4}$$

Therefore a necessary condition for the existence of stable expected class sizes is
(A) there exist non-negative constants $M_1, M_2, \ldots, M_C$ such that

$$M_j\, d_j(1-\omega_j p) = q \sum_{r=1}^{C} M_r b_r u_{rj}, \qquad j=1,2,\ldots,C.$$

If there is a positive solution to this system of equations then $\omega_j(p+qu_{jj}) \leqslant 1$ with equality only if $u_{rj}=0$ for all $r \neq j$. Denote by (B) the condition

(B) $$\omega_j(p+qu_{jj})<1, \qquad j=1,2,\ldots,C.$$

Under conditions (A) and (B) there is also a consistent, stable expected number of alleles with i copies in class j at equilibrium, viz.,

$$M_j(1-\omega_j p)(\omega_j p)^{i-1}/i, \qquad i \geqslant 1, \qquad j=1,2,\ldots,C.$$

A discussion of the conditions under which (A) is satisfied is given in Sect. 4. One special case of interest is $u_{ij}=C^{-1}$, the case where a mutant is equally likely to be a member of any of the C classes. If $u_{ij}=C^{-1}$ for all i, j then condition (A) holds if

$$q \sum_{j=1}^{C} \omega_j/(1-\omega_j p) = C \tag{5}$$

whence M_j is proportional to $[d_j(1-\omega_j p)]^{-1}$. Condition (5) is trivially satisfied if $\omega_j=1, j=1,\ldots,C$. If the ω_j are not all equal then condition (5) indicates the degree of weak selection allowed in the model.

3. The modified process

Even if (A) and (B) are satisfied the simple model put forward in Sect. 2 needs to be modified in order to obtain a non-trivial equilibrium distribution for the $G_j(t, i)$ process. Suppose that (A) and (B) hold and $u_{jj}<1$ for at least one $j=1,2,\ldots,C$. One way of modifying the above model to yield an interesting equilibrium behaviour is to adjust the rate at which mutant alleles are produced in such a way that the probability of a mutant offspring being produced in class r and joining class j in a small time interval $(t, t+dt)$ is $(M_r b_r q u_{rj})\, dt$.

We will use lowercase to denote the modified process. From (1) the modified process has m.g.f.

$$h(\boldsymbol{\theta}, t) = E \exp\left(\sum_{j=1}^{C} \sum_{i=1}^{\infty} \theta_{ij} g_j(t, i) \right)$$

which satisfies

$$\begin{aligned}\frac{\partial h}{\partial t} &= \sum_{j=1}^{C} \sum_{i=1}^{\infty} b_j p i[\exp(\theta_{i+1,j}-\theta_{ij})-1]\frac{\partial h}{\partial \theta_{ij}} \\ &\quad + \sum_{j=1}^{C} \sum_{i=2}^{\infty} d_j i[\exp(\theta_{i-1,j}-\theta_{ij})-1]\frac{\partial h}{\partial \theta_{ij}} + \sum_{j=1}^{C} d_j(e^{-\theta_{1j}}-1)\frac{\partial h}{\partial \theta_{1j}} \\ &\quad + q \sum_{j=1}^{C} (e^{\theta_{1j}}-1) \sum_{r=1}^{C} b_r M_r u_{rj} h(\boldsymbol{\theta}, t).\end{aligned}$$

Setting $\partial h/\partial t = 0$ we find that this differential equation is solved by the joint m.g.f.

$$h(\boldsymbol{\theta}, t) = \prod_{j=1}^{C} \prod_{i=1}^{\infty} \exp[\lambda_{ij}(e^{\theta_{ij}} - 1)]$$

where $\lambda_{ij} = M_j(1-\omega_j p)(\omega_j p)^{i-1}/i$. That is, suppressing the t, at equilibrium the numbers of alleles in class j with i copies, $g_j(i)$, are independent Poisson random variables with mean λ_{ij}, $j = 1, \ldots, C$; $i = 1, 2, \ldots$. Also at equilibrium the total number of alleles represented in class j, $k_j = \sum_{i=1}^{\infty} g_j(i)$, has a Poisson distribution with mean $-M_j[(1-\omega_j p)/\omega_j p]\log(1-\omega_j p)$.

At equilibrium the total number of individuals in the jth class, m_j, has m.g.f.

$$\prod_{i=1}^{\infty} \exp[\lambda_{ij}(e^{i\theta_j} - 1)] = [(1-\omega_j p\, e^{\theta_j})/(1-\omega_j p)]^{-\phi_j}$$

where $\phi_j = M_j(1-\omega_j p)/(\omega_j p)$, and so m_j has a negative binomial distribution,

$$P(m_j = m) = (1-\omega_j p)^{\phi_j}\binom{\phi_j + m - 1}{m}(\omega_j p)^m.$$

It is interesting at this point to go back and consider the linear birth and death process model proposed in Sect. 2 of Karlin and McGregor (1966). Suppose we have C *independent* linear birth and death processes operating with the jth class having birth and death rates b_j and d_j respectively and with new alleles entering the jth class as a Poisson process at constant rate $q\sum_{r=1}^{C} M_r b_r u_{rj}$, for some constants $M_1, M_2, \ldots, M_C$. Further suppose the new alleles evolve with the same birth and death rates as the other members of the class they enter. From Karlin and McGregor Theorem 2.1 we have that in class j the $g_j(t, i)$, $i = 1, 2, \ldots$, have independent Poisson *transient* distributions, similar to the equilibrium distributions in the model above, and since the classes are independent it is a simple matter to sum across classes to get the transient distributions of $\sum_{j=1}^{C} g_j(t, i)$ and $m_j(t)$, the total number of individuals in class j at time t. The Karlin and McGregor results follow without any conditions on $M_1, \ldots, M_C$. The role of conditions (A) and (B), restricting the values of $M_1, \ldots, M_C$ that can be considered, is to ensure that the proposed "modified process" is a closed system with class sizes that remain finite.

Next we will investigate some features of the proposed modified process. First, the conditional distribution of $g_j(1), g_j(2), \ldots$ given that the total number of alleles represented in the jth class is $k_j = k$, is multinomial; the conditional probability that an allele selected at random from the jth class is represented by i copies being

$$-(\omega_j p)^i/[i\log(1-\omega_j p)], \qquad i = 1, 2, \ldots.$$

More generally, the probability that an allele selected at random from the population is represented by i copies is

$$E\left[\sum_{j=1}^{C} g_j(i)\Big/\sum_{j=1}^{C} k_j\right] = \sum_{j=1}^{C} (\omega_j p)^i/\{i|\log(1-\omega_j p)|\}\cdot E\left[k_j\Big/\sum_{j=1}^{C} k_j\right]$$

$$= \sum_{j=1}^{C} a_j(\omega_j p)^i/\{i|\log(1-\omega_j p)|\}$$

where

$$a_j = \frac{M_j(1-\omega_j p)[\log(1-\omega_j p)]/(\omega_j p)}{\sum_{j=1}^{C} M_j(1-\omega_j p)[\log(1-\omega_j p)]/(\omega_j p)},$$

as the k_j are independent Poisson random variables.

Another aspect of this model which is of interest is the conditional distribution of the number of alleles represented by $1, 2, 3, \ldots$ copies, i.e. $\sum_{j=1}^{C} g_j(1)$, $\sum_{j=1}^{C} g_j(2), \ldots$, given the total number of individuals in the population, $\sum_{j=1}^{C} m_j$. First $\sum_{j=1}^{C} g_j(i)$ has a Poisson distribution with mean $p^i \sum_{j=1}^{C} (\phi_j \omega_j^i)/i$. Also $\sum_{j=1}^{C} m_j$ has m.g.f.

$$\prod_{j=1}^{C} [1-\omega_j p\, e^{\theta})/(1-\omega_j p)]^{-\phi_j},$$

so

$$P\left(\sum_{j=1}^{C} m_j = n\right) = \prod_{j=1}^{C} (1-\omega_j p)^{\phi_j} p^n A_n,$$

where

$$A_n = \left[\sum_{\substack{x_1+\cdots+x_C=n \\ x_i \geq 0}} \left\{\prod_{j=1}^{C} \binom{\phi_j + x_j - 1}{x_j} \omega_j^{x_j}\right\}\right].$$

Thus

$$\begin{aligned} P&\left(\sum_{j=1}^{C} g_j(1) = a_1, \sum_{j=1}^{C} g_j(2) = a_2, \ldots \middle| \sum_{j=1}^{C} m_j = n\right) \\ &= \left[\prod_{i=1}^{\infty} \left\{\exp\left(-p^i \sum_{j=1}^{C} (\phi_j \omega_j^i) \middle/ i\right)\right.\right. \\ &\quad \left.\left. \cdot \left[p^i \sum_{j=1}^{C} (\phi_j \omega_j^i) \middle/ i\right]^{a_i} \middle/ a_i!\right\}\right] \middle/ P\left(\sum_{j=1}^{C} m_j = n\right) \\ &= A_n^{-1} \prod_{i=1}^{n} \left[\left(\sum_{j=1}^{C} \phi_j \omega_j^i\right) \middle/ i\right]^{a_i} \middle/ a_i!, \end{aligned} \tag{6}$$

where $n = \sum_{i=1}^{\infty} i a_i$.

From the discussion in Sect. 4, if condition (A) is satisfied then the model with no selection acting is precisely the model with $\omega_j = 1$, $j = 1, \ldots, C$. Thus if there is no selection (6) reduces to

$$P\left(\sum_{j=1}^{C} g_j(1) = a_1, \sum_{j=1}^{C} g_j(2) = a_2, \ldots \middle| \sum_{j=1}^{C} m_j = n\right) = \binom{\theta+n-1}{n}^{-1} \prod_{i=1}^{n} (\theta/i)^{a_i} \middle/ a_i!,$$

where $\theta = (q/p) \sum_{j=1}^{C} M_j$, which is Ewens' sampling formula. Thus the modified process yields an equilibrium behaviour which provides a generalisation of Ewens' result to certain cases of weak selection.

Further,

$$E\left(\prod_{i=1}^{n} \left(\sum_{j=1}^{C} g_j(i)\right)^{[n_i]} \middle| \sum_{j=1}^{C} m_j = n\right) = A_n^{-1} \prod_{i=1}^{n} \left[\left(\sum_{j=1}^{C} \phi_j \omega_j^i\right) \middle/ i\right]^{n_i} A_{n-N},$$

where $N=\sum_{i=1}^{n} in_i$ and $x^{[n_i]}=x(x-1)\cdots(x-n_i+1)$. In particular,

$$E\left(\sum_{j=1}^{C} g_j(i)\,\middle|\,\sum_{j=1}^{C} m_j=n\right)=A_n^{-1}\left(\sum_{j=1}^{C}\phi_j\omega_j^i\Big/ i\right)A_{n-i}.$$

4. Comments on condition (A)

Condition (A) imposes constraints on the ω_i and u_{ij} parameters and the ω_i in turn reflect the relative selective advantages of the various classes of alleles. Summing the equations in (A) we get

$$\sum_{j=1}^{C} M_j d_j(1-\omega_j)=0 \tag{7}$$

and so if $\omega_1=\omega_2=\cdots=\omega_C$ and some M_j is positive then (7) implies that $\omega_j=1$ for all j. Also if ω_j are not all equal then at least one ω_j is greater than 1 and at least one ω_j is less than 1.

In general we can write the system of equations given in (A) in the form $\mathbf{AM}=\mathbf{0}$, where $\mathbf{M}'=(M_1, M_2, \ldots, M_C)$ and $\mathbf{A}$ is a $C\times C$ matrix. A non-zero solution to these equations exists if $\det \mathbf{A}=0$. Let $\mathbf{C}=\mathbf{U}'-\mathbf{I}$, where $\mathbf{U}=(u_{ij})$ is a matrix of probabilities. With the convention $(1-\omega_j)(\omega_j/(1-\omega_j))=1$ if $\omega_j=1$, we can write

$$\det \mathbf{A}=\left[\prod_{j=1}^{C} d_j(1-\omega_j)\right](-1)^C \cdot\left\{1+\sum_{k=1}^{C-1}(-q)^k\sum_{i_1,\ldots,i_k}\left(\frac{\omega_{i_1}}{1-\omega_{i_1}}\right)\cdots\left(\frac{\omega_{i_k}}{1-\omega_{i_k}}\right)\det \mathbf{C}_{i_1\ldots i_k}\right\}, \tag{8}$$

where $\mathbf{C}_{i_1\ldots i_k}$ is the $k\times k$ matrix formed by the elements in both the $i_1,\ldots,i_k$ rows and columns of $\mathbf{C}$ and $\sum_{i_1,\ldots,i_k}$ denotes summation over all distinct subsets of size k drawn from $\{1, 2, \ldots, C\}$. The dependence of $\det \mathbf{A}$ on ω_j and u_{ij} is made clear in (8).

If $C=2$ then $\det \mathbf{A}=0$ implies either $\omega_1=\omega_2=1$ or

$$q\left[u_{12}\left(\frac{\omega_1}{\omega_1-1}\right)+u_{21}\left(\frac{\omega_2}{\omega_2-1}\right)\right]=1.$$

In both cases, if (B) holds, positive solutions for M_1 and M_2 exist. Conditions ensuring the existence of positive solutions for $M_1,\ldots,M_C$ can be derived for the cases $C>2$ by considering (8) and the equations in (A).

One special case of the above model which is of some interest is the case $u_{ij}=C^{-1}$. This model represents the situation where the class of a mutant is selected at random from the C available. In this case the (i,j)th element of $\mathbf{A}$ is

$$\begin{aligned} a_{ij}&=d_j\omega_j qC^{-1}, && i\neq j\\ &=d_j(\omega_j p-1-q\omega_j C^{-1}), && i=j \end{aligned}$$

and so

$$\det \mathbf{A}=\left[\prod_{j=1}^{C} d_j(\omega_j p-1)\right]\left[1+qC^{-1}\sum_{j=1}^{C}(\omega_j/(\omega_j p-1))\right].$$

If $\omega_j p < 1, j = 1, 2, \ldots, C$ then $\det \mathbf{A} = 0$ if

$$C = q \sum_{j=1}^{C} (\omega_j/(1 - \omega_j p))^{-1}.$$

and M_j is proportional to $[d_j(1 - \omega_j p)]^{-1}$.

Acknowledgement. The authors would like to thank Professor Nagylaki and the referees for their helpful comments which led to an improved presentation of the results in this paper.

References

Bailey, N. T. J.:The elements of stochastic processes with applications to the natural sciences. New York: Wiley, 1964

Ewens, W. J.: The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **4**, 251-259 (1973)

Ewens, W. J.: Mathematical population genetics. Berlin Heidelberg New York: Springer 1979

Gillespie, J. H.: Sampling theory for alleles in a random environment. Nature (Lond.) **266**, 443-445 (1977)

Karlin, S., McGregor, J.: The number of mutant forms maintained in a population. Proc. Fifth Berk. Symp. of Math. Stat. and Probab. **4**, 403-414 (1966)

Kingman, J. F. C.: A note of multi-dimensional models of neutral mutation. Theor. Popul. Biol. **11**, 285-290 (1977a)

Kingman, J. F. C.: The population structure associated with the Ewens sampling formula. Theor. Popul. Biol. **11**, 274-283 (1977b)

Neel, J. V., Rothman, E. D.: Is there a difference among human populations in the rate with which mutation produces electrophoretic variants? Proc. Natl. Acad. Sci. USA **78**, 3108-3112 (1981)

Rothman, E. D., Templeton, A. R.: A class of models of selectively neutral alleles. Theor. Popul. Biol. **18**, 135-150 (1980)

Watterson, G. A.: The sampling theory of selectively neutral alleles. Adv. Appl. Probab. **6**, 463-488 (1974)

Watterson, G. A: The stationary distribution of the infinitely-many neutral alleles diffusion model. J. Appl. Probab. **13**, 639-651 (1976)

Watterson, G. A.: Heterosis or neutrality? Genetics **85**, 789-814 (1977)

Received April, 1985/Revised March 26, 1986