EVOLUTION
INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

The role of weak selection and high mutation rates in nearly neutral evolution

Journal:	*Evolution*
Manuscript ID:	draft
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Lawson, Daniel; Biomathematics and Statistics Scotland Jensen, Henrik; Imperial College London, Institute for Mathematical Sciences; Imperial College London, Mathematics
Keywords:	Fitness, Models/Simulations, Variation

The role of weak selection and high mutation rates in nearly neutral evolution

Daniel John Lawson*
daniel@bioss.ac.uk
Biomathematics and Statistics Scotland
Macaulay Institute, Craigiebuckler, Aberdeen, UK. AB15 8QH
Tel: +44 (0)1224 498200 Fax: +44 (0)1224 311556

Henrik Jeldtoft Jensen
h.jensen@imperial.ac.uk
Institute for Mathematical Sciences, Imperial College London
53 Princes Gate, South Kensington Campus
London, UK. SW7 2PG
and:
Department of Mathematics, Imperial College London
South Kensington campus, London, UK. SW7 2AZ

July 15, 2008

Abstract

Neutral dynamics occur in evolution if all types are 'effectively equal' in their reproductive success. Population dynamics with selection imply that the definition of 'effectively equal' depends on the population size and the details of mutations. Genetic data for extremely large clonal populations indicates that many genes evolve neutrally, which current models can only explain if selection on those genes is completely absent. Such models typically consider the case where mutations are rare, so that population dynamics occurs at a different timescale to evolution and there are at most two competing types. However, if the probability of a mutation in

*Corresponding author.

the population as a whole is large, then the whole distribution of types must be considered. We show that this has important consequences for the occurrence of neutral dynamics. In highly connected type spaces, neutral dynamics can occur for all population sizes despite significant selective differences, via the forming of effectively neutral networks connecting rare neutral types. Biological implications include an explanation for the high diversity of rare types that survive in large clonal populations, and a theoretical justification for the use of neutral null models.

1 Introduction

The evolution of a population is influenced by both chance events and selection. Selection acts on a population via differential reproductive success brought about by heritable differences. Chance events include mutations causing heritable differences, and the random process of population dynamics. The perceived relative importance of these various process has changed over time. Darwin (1859) believed that selection with variation was paramount, but more recently Kimura (1983) and many others (Tachida, 1991; Ohta, 2003; Nei, 2005a) have demonstrated that chance in population dynamics best describes the fixation of many mutations using the ‘nearly neutral model of molecular evolution’. Very recently the relative importance of chance has again been challenged (Hahn, 2008). The current genetic inference framework (Felstenstein, 1988) measures phylogenetic relationships in terms of the number of mutations and therefore requires neutral evolution of at least some loci. It is therefore essential to address the relevance of the neutral model as a null hypothesis.

By current methods it is estimated that 50% of loci in some bacterial genes (Charlesworth and Eyre-Walker, 2006) are shaped by adaptation. This leaves a huge proportion of the genome shaped by effectively neutral substitutions.

Could undetected selection at these loci be relevant for evolution? Mutations resulting in a small change to reproductive ability are common in both coding and non-coding regions of the genome (Ohta, 1997), arising for example via the stability of RNA folding (Aita et al., 2003), gene regulation (Ohta, 2002) and increased efficiency of shorter genomes. Under the nearly neutral theory of molecular evolution, each genetic component usually contributes independently to reproductive success. Effectively neutral dynamics are observed for selection less than some critical value which decreases inversely with increasing population size. In bacterial populations, selective differences would have to be essentially absent for neutral evolution to occur at large population sizes. Since (very) small fitness differences are to be expected in all mutations, it is important to address why neutral evolution should be appropriate at all for viruses and bacteria.

The standard theoretical approach to evolution is to assign 'fitness' to genes under given genetic and environmental conditions, which translates to a reproductive ability for the individual. In a sexually reproducing organism, genes are regularly recombined in different combinations and over evolutionary time an average fitness may be assigned to each gene by averaging over all possible genetic environments. However, in asexually reproducing organisms, recombination is rare and gene interactions are more important in determining long term reproductive success. In this case a better model is to assign a 'fitness' to a combination of genes, i.e. to the type of the individual. Using this approach we demonstrate that effectively neutral evolution may occur at relatively strong selection in large populations, when compared with the more frequently studied model of independent contribution to fitness of each gene.

We find that neutral dynamics cannot be supported in a large population when mutation rates are low, such that no mutations are expected during a generation. In this case population dynamics and evolutionary dynamics occur on different timescales so at most two types will be present at a given time. At higher mutation rates, trends or correlations in reproductive success through type space of the order of $\sqrt{Np_m}$ mutations are relevant, where N is the population size and p_m the mutation rate. When strong long distance trends in reproductive ability are present (for example, when alleles contributes independently), non-neutral dynamics are observed at smaller selection strength than expected at low mutation rate. In this case small differences in fitness per mutation lead to a large difference in fitness over the whole *population*. However, if there are no long distance trends in reproductive ability, genetically very different types may have similar reproductive success and neutral dynamics can still be observed for much larger selection strength. An 'effectively neutral' network of types is formed, in which *nearest neighbour* types need not be competitively neutral. Competitively neutral types are connected by less fit types in a way that does not affect the statistics of the evolution of the population as a whole.

If we assume strong linkage in sexually reproducing populations then our model becomes an appropriate description. In this case we explain the 'paradox of variation' (Hahn, 2008), that more variation is not observed in larger populations under the neutral model. The same *degree* of neutrality is observed regardless of population size. Thus recombination rate is likely to play a vital role in the applicability of neutral models.

Our model predicts the conditions for emergence of neutral networks without *a-priori* assuming neutral dynamics should occur. Neutral networks themselves have found application to viral evolution (van Nimwegen, 2006) and have been well studied previously (van Nimwegen et al., 1999; van Nimwegen and Crutchfield, 2000). When mutations off a neutral network are deadly, a 'holey fitness landscape' (Gavrilets, 1999; Bastolla et al., 2002) is instead formed, for which moving across non-neutral regions is impossible. Our work shows that both of these models are valid when selection is present even for very large populations.

Our results support the argument (Nei, 2005b) that phenotypes and genotypes will evolve qualitatively differently. This is because the dimensionality of a genotype space is considered to be higher than that of a phenotype space (Huynen et al., 1996), and therefore the connectively of the neutral network is also higher. Genotypes (i.e. a very high dimension space of possible mutations) may evolve neutrally even when population sizes become large. Hence genetically diverse asexual individuals in a large population may compete effectively neutrally, and therefore a large number of cryptic species would be expected. This cannot happen for simple phenotypes, or sexually reproducing individuals if fast genetic exchange results in the emergence of a 'fittest' combination of genes.

We consider a simple evolution model, to which we apply a combination of simple semi-rigorous arguments and simulation. This allows a description of the conditions required for fully neutral models to accurately represent the evolution of a population in which small selective differences may be present.

2 A conserved population nearly-neutral evolution model

A simple Moran birth/death process (Moran, 1962) is considered with clonal reproduction in a type space. The type of an individual is it's position in type space, which determines it's reproductive probability using a 'fitness landscape' model to describe how reproductive success varies with type. Three possible representative fitness landscapes are considered.

2.1 Definition of the model

A conserved number of individuals N are considered, with each individual i belonging to a given type $\mathbf{x}_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(D)})$ in an infinite ranged D-dimensional type space. Each type $\mathbf{x}_i$ has a fitness $F(\mathbf{x}_i)$ which determines that types reproductive success. A generation consists of performing N of the following timesteps:

1. Select an individual i uniformly from the population which will be killed at the end of the timestep.

2. Select an individual j of type $\mathbf{x}_j$ for reproduction with probability $p_{\mathit{off}}(\mathbf{x}_j) = F(\mathbf{x}_j)/(\sum_{k=1}^{N} n_k F(\mathbf{x}_k))$.

3. Create an offspring of individual j with initial type $\mathbf{x}_j$. With probability p_m a mutation occurs in a single type dimension, say $\mathbf{x}_i^{(\alpha)}$, with $\alpha \in (1, D)$ each chosen with probability $1/D$. The mutation involves $\mathbf{x}_i^{(\alpha)}$ changing by $+1$ or -1 with equal probability.

It is simple to see that if all fitnesses $F(\mathbf{x}_j)$ are equal then the dynamics are fully neutral. We will now define the various fitness landscapes $p_{\mathit{off}}(\mathbf{x})$.

2.2 Fitness

To capture important qualitative features of the change in fitness with type, the following three simple definitions of a fitness landscape are considered. Landscape 1: the random fitness landscape, and landscape 2: the 'top-hat' correlated fitness landscape are considered as two extremes of fitness landscapes that are globally bounded. In landscape 3: the linear fitness landscape the potential fitness difference within a population is unbounded.

Landscape 1: The random uncorrelated fitness landscape is maximally rugged, and created by the following function:

$$F(\mathbf{x}; s) = 1 - sy(\mathbf{x}), \tag{1}$$

where $y(\mathbf{x})$ is a random number generated uniformally in $(0, 1)$ for each $\mathbf{x}$. Hence the fitness $F(\mathbf{x}; s)$ is uncorrelated between types and is in the range $[1 - s, 1]$.

This fitness landscape can also be related to simple correlated (i.e. smoothly varying but random) fitness landscapes by rescaling. Consider a correlated random fitness landscape with correlation length η, such that $\langle F(\mathbf{x})F(\mathbf{x}')\rangle \propto \exp(-(\mathbf{x} - \mathbf{x}')/\eta)$. By rescaling mutation size and mutation rate (i.e. 'coarse-graining' the fitness landscape) the correlated fitness landscape can be reduced to an uncorrelated fitness landscape. Mutation at rate p_m creates a random walk in type space for a given lineage (Yi-Cheng Zhang et al., 1990) in which the mean population position $\|\mu\|(t) = \sqrt{\sum_{i=1}^{N}\sum_{d=1}^{D}(x_i^d)^2} \propto \sqrt{p_m t}$ at time t. Therefore scaling space as $x' = x/\eta$ requires scaling mutation rate as $p'_m = p_m/\eta^2$. The correlated fitness landscape in the unprimed variables is described statistically by the random fitness landscape in the primed variables.

Sufficiently large random correlated landscapes require a great amount of care to construct (Laird and Jensen, 2006) and are therefore not considered here. The random landscape is generated by using a pseudo-random number generator with seed given by the location in type space $\mathbf{x}$.

Landscape 2: The 'top-hat' correlated fitness landscape is an extreme example of a correlated landscape, given by:

$$\begin{aligned} p_{\mathit{off}}(\mathbf{x}; s) &= 1, & \text{if all } |\mathbf{x}^{(\alpha)}| < L, \\ &= 1 - s, & \text{if any } |\mathbf{x}^{(\alpha)}| \geq L. \end{aligned} \tag{2}$$

As before, the label $\alpha \in (1, D)$ refers to directions in type space. Equation 2 describes a 'top hat' function such that fitness decreases by an amount s outside a square (in D dimensions) of side $2L$. This represents type spaces with a single well defined fit area. As before the maximum fitness gradient is s.

Landscape 3: The linear fitness landscape has fitness increasing linearly in all dimensions:

$$p_{off}(\mathbf{x}; s) = 1 + s \sum_{\alpha=1}^{D} x^{(\alpha)}. \tag{3}$$

Again the maximum fitness gradient is s between neighbouring types, but the maximum fitness difference over the whole population is unbounded.

3 Theory

The nearly-neutral case with high mutation rates is difficult to approach analytically, and so we use simple semi-rigorous but informative 'mean-field' arguments which are backed by numerical simulation. The size of the 'neutral regime' is considered, i.e. the range of selection strengths for which effectively neutral dynamics are observed.

3.1 Characterisation of neutral dynamics

Since neutral evolution is itself dynamically rich, a careful characterisation is necessary. The neutrally evolving population distribution can be accurately described (Lawson and Jensen, 2007) as a 'cloud' of individuals forming a number of distinct clusters in type space moving in a correlated manner. This can be described *statistically* as a 'peak': i.e. a single entity with a given mean position and width, both of which change in time. However the distribution is not continuous as in e.g. (Schuster, 1997). A correct model of neutral dynamics is useful for calculations when selection is small but significant, which are usually expanded around the neutral case (e.g. (Traulsen et al., 2006)).

A neutrally evolving population is described as a statistical distribution with a mean position $\mu(t)$ (i.e. centre of mass) and a standard deviation $w(t)$ (referred to as a width to avoid confusion with the standard deviation *of* the width). The mean position and the width evolve as random variables with known behaviours. The mean position performs a simple random walk (Bailey, 1964) characterised by:

$$\|\mu\|(t) \propto t^{\beta} \tag{4}$$

with $\beta = 1/2$. The width fluctuates around the time-averaged width w^* given by:

$$w^* = \lim_{T_m \to \infty} \frac{1}{T_m} \int_{t=0}^{T_m} w(t)dt. \tag{5}$$

There are two possible statistically relevant effects of weak selection on the neutral population distribution. The first is a change in effective diffusion rate of $\mu(t)$ via either pinning, i.e. a reduced velocity of the mean population type, or an active selection gradient, i.e. an increase in velocity. The second effect is that selective forces alter either the population distribution size w or the magnitude of its fluctuations compared with the neutral case.

The time-averaged width w^* takes a different value to the 'equilibrium' width w^{equil}, for which the expected change of width in time is zero:

$$\langle \frac{dw}{dt} \rangle_{w=w^{\text{equil}}} = 0. \tag{6}$$

Therefore reduced fluctuations of $w(t)$ produce a contraction of $\langle w \rangle$ towards w^{equil}, and conversely for increased fluctuations. A change in w^{equil} will likewise produce a change in the time-averaged width $\langle w \rangle$. Thus $\langle w \rangle$ is an accessible measure characterising neutral dynamics and together with $\|\mu\|(t)$ characterises neutral evolution.

3.2 Two competing types

It is instructive to recap the well understood case of low mutation rates, such that only two types compete at a given time. The higher mutation rate cases will be compared to this simple case.

Consider a population of size N at a good type with $F = 1$. A single mutation occurs to a poor type i on which $F(i) = 1 - s$. For no selection ($s = 0$), the population size $n_i(t)$ of type i would do a random walk starting from 1. Type i typically becomes extinct ($n_i = 0$) but after N attempts it is expected to succeed, i.e. n_i reaches N (Fraser, 1976). A less fit type that can succeed in $O(N)$ attempts is called effectively neutral. For this reason neutral evolution occurs more slowly in larger populations when mutation rates are small.

This problem is solved under the name of 'Gambler's Ruin' (see e.g. (Ash, 1970)), when a time step is defined as waiting until the population of the unfit type $n_i(t)$ changes. The ratio of the probability of n_i increasing to the probability of decreasing is $1 - s + O(s^2)$, with s considered small. By comparison to the Gambler's Ruin problem with this ratio, a population of the poor type i of initial size $n_i(0) = 1$ in a total population of N will eventually reach population size N with probability:

$$p_{\text{poor}} = \frac{s}{(1+s)^N - 1}. \tag{7}$$

The neutral case with $s = 0$ succeeds with probability $p_0 = 1/N$. The ratio p_{poor}/p_0 is 'exponential like' in s with the characteristic scale:

$$s^* = \frac{2}{N}, \tag{8}$$

or equivalently, effectively neutral evolution is observed for $s < s^* = 2/N$. s^* is called the *critical* selection value.

When only two types ever compete, selection acts via pinning, i.e. long waiting times at high fitness types. In this case the above argument can be extended to fitness landscapes with multiple high and low fitness areas. Ref. (Aranson et al., 1997) performs such an argument mathematically in a slightly different fitness landscape to ours. By analogy to pinning in anomalous diffusion (Bouchaud and Georges, 1990; Ralf Metzler and Joseph Klafter, 2000), if there is some maximum to the time the population can spend at fit sites then a rescaling of the mutation rate will recover standard mutation-drift dynamics. In this case fitness variation is irrelevant over long times. However, if the time taken to leave fit types is unbounded then the motion of the population becomes subdiffusive in type space. This means that the average root-mean-square position $\|\mu\|(t) \propto t^{\beta}$, with $\beta < 1/2$ and the dynamics are not statistically neutral.

As p_m increases a range of types can coexist. We will address whether the existence of a population distribution around a high fitness type allows faster escape, or if the low fitness of the surrounding types prevents the establishment of a wide population distribution.

In summary, if mutations are rare ($Np_m \ll 1$) and selection is weak ($s \ll 1$), fully neutral behaviour of the peak is expected for $s < s^* \propto N^{-1}$.

3.3 Predictions for a large population in a fitness landscape

When mutation rates are high, analytical techniques become difficult and we will resort to simulation. However, some predictions can be made by making strongly simplifying assumptions, which are explained here without mathematical detail.

Landscape 1: the random uncorrelated fitness landscape represents a correlated landscape upon rescaling p_m. The critical selection $s^* \propto N^{-\alpha}$ must follow $\alpha \to 1$ as $p_m \to 0$, but for large p_m the population contains a number of types and therefore the distribution of distances between fit types will play a role. As $D \to \infty$ the probability of high fitness types being close increases and hence α may decrease. This is not simple to model theoretically and is the main target of the simulation study.

Landscape 2: the top-hat fitness landscape can be understood theoretically as described mathematically by van Nimwegen et al. (1999) for a more general case. This can be considered as a single fit type consisting of the whole fit region competing with a single less fit type consisting of the whole unfit region. The dynamics between regions are related to the two-type case with mutation rate across the fitness boundary depending on the specific population distribution. Within a single region the dynamics are neutral. Since there are effectively only two types competing regardless of D, the dynamics follow the low mutation rate case above with the upper bound in selection strength for neutral dynamics $s^* \propto N^{-1}$.

Landscape 3: the linear fitness landscape can also be understood theoretically at large mutation rates, as discussed in detail by Kessler et al. (1997) for the more general case of large selection. The width of the population distribution $w \propto (p_m N)^{1/2}$ for all selection s. Therefore the effective fitness difference of individuals *within* the population $s_{\text{diff}} \propto s(p_m N)^{1/2}$. The best and worse types compete with small effective mutation rate as in the two type case with $s^*_{\text{diff}} \propto N^{-1}$. Therefore $s^*_{\text{diff}} \propto s^*(p_m N)^{1/2}$ and by rearrangement the upper bound in selection strength for neutral dynamics $s^* \propto N^{-3/2}$. Additionally, the peak position performs a biased random walk in the usual way (e.g. (Bailey, 1964)), with deterministic drift component $v \propto p_m s_{\text{diff}}$ and variance component $D^2 \propto 1/N^2$. Deterministic drift dominates the random walk if $v > D$, or $s_{\text{diff}} > s^*_{\text{diff}} \propto (Np_m)^{-1}$ and again $s^* \propto N^{-3/2}$. This holds for arbitrary dimension D as all mutations have an equal chance of increasing or decreasing fitness.

4 Results

Since the theoretical predictions for landscapes 2 and 3 correctly describe the simulation results, we focus on the case of landscape 1, the random fitness landscape. Results for all landscapes are presented in summary form. Non-neutral dynamics are observed for $s > s^*$, where $s^* = \min(s^*_w, s^*_\mu)$ is the critical selection found by observing changes to either the average width (observed for $s > s^*_w$) or mean position (observed for $s > s^*_\mu$).

The general simulation approach is to perform ensemble averaging over a large number (100+) of runs at a range of parameters, and use statistical bootstrapping techniques (Davison and Hinkley, 1997) to provide accurate standard errors. The focus is the relationship between population size and effectively neutral dynamics. To avoid repetition, detailed results are given for the case of low dimension which is well understood theoretically under truly neutral dynamics (Lawson and Jensen, 2007). It is important to stress that the general features discussed extend to arbitrary dimension, including the genetically relevant infinite dimension limit.

4.1 Effects on the width

As discussed, the average width of the population distribution provides a strong indicator of neutral dynamics in a population. Fig. 1 (left) shows the average width of the population distribution as a function of selection s. Each population size displays a different region $s < s^*_w(N, p_m, D)$ for which the width $w(s) = w(0)$, i.e. selection is not effecting the observed average width. As selection is increased above s^*_w the average width decreases as selection suppresses fluctuations.

Considering only the effect of N on s^*_w, the neutral data can be collapsed as $s' = s/s^*_w(N)$ and $w'(s) = w(s)/w(s = 0)$ shown in Fig. 1 (right), with

s_w^* shown as a vertical line. Accurate confidence intervals for s_w^* are found by statistical bootstrapping; see Appendix. For $D = 1$ the critical selection is $s^*(N) = 8N^{-0.94}$ (see Fig. 2 for error margins). The data collapse is intended only for $s < s^*$, though in this case holds over the whole parameter region. The region $s > s^*$ corresponds to non-neutral dynamics. The critical selection observed via a change in the average width is of the general form:

$$s_w^* = aN^{-\alpha(D,p_m)}D^{\beta(p_m)}, \tag{9}$$

where a is a constant, β is an exponent for the dimension dependence and α for population size dependence. We focus on the case of constant D and p_m such that $s_w^* \propto N^{-\alpha}$ and measure α.

The data for landscape 2 (the top-hat fitness landscape) is fit similarly to the Random Landscape. For landscape 3 (the linear fitness landscape), the fitting procedure is more difficult since a general nonlinear form must be used for $w(s)$, so averages and plausible ranges of s^* are estimated by eye for this case only.

The details for all cases and dimensions are summarised in Fig. 2. Theoretical predictions for the scaling of the critical selection $s^* \propto N^{-\alpha}$ are supported. For landscape 2: the top-hat landscape theory suggested $\alpha = 1$, and for landscape 3: the linear landscape the prediction was $\alpha = 3/2$. For landscape 1: the random fitness landscape there is a clear relationship with dimension, starting in $D = 1$ at a value away from $\alpha = 1$ and decreasing (close to linearly) with dimension towards $\alpha = 0$ where it remains for $D \to \infty$. The gradient of the decrease is mutation-rate dependent.

4.2 Mean Position in type space

In the neutral case, the mean position of the population distribution in type space (i.e. the average type) is a random variable performing a random walk.

Deviations from this pattern are of interest. Using Eq. 4 a simplistic method to find the diffusion exponent β is as the regression coefficient when plotting the root-mean-squared distance from the origin $\|\mu\|$ as a function of time t on double logarithmic axes. This is $\beta = 1/2$ in the fully neutral case. When there is pinning to a particular type with high fitness then β should decrease. When there is a fitness gradient, a velocity should be induced and β should increase.

The evolution of a population in a random fitness landscape is related to the behaviour of a random walker in a random potential. Studying this numerically is notoriously difficult (Bouchaud and Georges, 1990) and the above method poorly captures the asymptotic behaviour often considered mathematically. However, in this case it is the short and medium time scales that are of relevance to biological evolution which are captured in β as measured by the above regression method.

As expected from Eq. 4, $\beta = 1/2$ is observed for neutral dynamics as seen for all selection strengths $s < s^*_\mu$ (not shown). The critical selection observed s^*_μ in the mean position has the form

$$s^*_\mu = bD^\gamma(p_m), \tag{10}$$

for some constant b and exponent γ, with no dependence on population size N for all $N \geq 500$.

4.3 Width as the important measure

Using the definition of $s^* = \min(s^*_w, s^*_\mu)$ and Equations 9 and 10, the mean position dynamics provide an important constraint on neutral dynamics if $s^*_\mu < s^*_w$. By rearrangement:

$$\alpha \log(N) + (\gamma - \beta)\log(D) < [\log(a) - \log(b)]. \tag{11}$$

Therefore the signs of $(\gamma - \beta)$ and $[\log(a) - \log(b)]$ determine which constraint holds as N and D become large. From the results it is found that both are positive using simple regression in the dimension variable D. For example, at $p_m = 0.05$ in the random landscape $(\gamma - \beta) = 4.5 \pm 0.7$ and $[\log(a) - \log(b)] = 6.9 \pm 0.8$. The signs of the combined constants (which hold for all tested mutation rates) imply all terms in Equation 11 are positive. Therefore pinning is important (i.e. observed at smaller selection than distribution width fluctuation changes) only when both N and D are small. Therefore, at low N and D a change in the mutational drift rate can be observed before the population distribution changes shape. If $N \to \infty$ or $D \to \infty$ then $s_w^* < s_\mu^*$ and $s^* = s_w^*$, i.e. we need only observe the average width.

4.4 Interpreting the results

A statistical description of neutral evolution was used to characterise the effect of selection on a population distribution. This defined a 'neutral regime' in which the population as a whole evolved effectively neutrally.

For low mutation rates $Np_m \ll 1$ all individuals are distributed over a maximum of two types that compete with each other. In this case neutral dynamics are observed for $s < s^* \propto N^{-\alpha}$ with $\alpha = 1$, as is found in classical models. This occurs regardless of the distribution of fitter types in the fitness landscape.

For high mutation rates $Np_m \geq 1$ the population forms a distribution over many types. The relevant selection parameter s measures the *maximum range* of fitnesses experienced by the population. Fitness landscapes with a single maxima (Eq. 2), or with long range trends (Eq. 3, s is redefined as s_{diff}) also have critical selection $s^* \propto N^{-\alpha}$ with $\alpha = 1$. However, fitness landscapes with large fluctuations but no long distance trend (Eq. 1) allow neutral dynamics to be observed for a larger range of population sizes N. In suitably connected fitness landscapes such as that of genotypes ($D \to \infty$) there is *no effect* of population size on the critical selection strength s^*. When selection is below s^*, taking the limit $N \to \infty$ results in a neutral model. The studied case of a random landscape is important because large populations $N \to \infty$ have distribution width $w \propto \sqrt{p_m N}$ so cover a wide range of types. Therefore short or medium sized correlations in fitness will be irrelevant - there may be no important trend in fitness for large populations. The counter-intuitive result of neutral dynamics occurring in a large population with a range of fitnesses can be understood as follows.

In the neutral regime for high mutation rates, the population will contain a large number of types with differing fitness. The fitness difference may be large enough to be measured as selectively important at the level of single mutations, but at the population distribution level the fitness of these types will be effectively ignored. Fig. 3 illustrates how distinct fit neutral types may be connected by less fit types. Neutral types do not have to be adjacent, but only within the fluctuation region of order $\sqrt{p_m N}$. It is only when selection strength $s > s^*$ (of order 1) that crossing unfit regions becomes unlikely.

Fig. 3 illustrates the importance of connectivity and hence the dimensional dependence for neutral dynamics to be observed. Percolation theory (Grimmett, 1999) may be an appropriate description, but is difficult to apply because the population forms a wide distribution. As $N \to \infty$ the proportion of truly neutral types tends to zero but the number of types within the fluctuation region increases. The fluctuation region of a good type contains intermediate fitness types (grey in Fig. 3) for which the extinction time is critically determined by the number of mutations onto the type. Therefore the threshold for ‘good enough’ types is not well defined, but depends on the distance from a fit type.

Hence this simple percolation argument should not be taken beyond an illustrative 'cartoon', although the simple linear relationship of α with dimension in Fig. 2 is indication that an analytical approach may be fruitful. Using an alternative argument, van Nimwegen and Crutchfield (2000) evaluate the time taken to cross variable sized fitness barriers, but the argument does not extend easily to the case of a random fitness landscape.

5 Discussion

5.1 Relation to other works

We have described a mechanism by which a neutral network may form for large clonal populations when fitness variation can occur on every mutation. This provides theoretical support to models representing large populations with neutral networks (van Nimwegen, 2006) and holey fitness landscapes (Gavrilets, 1999; Drossel, 2001). Our work adds to previous models by demonstrating how a neutral network can form without *a-priori* assuming that neutral dynamics should be observed.

Other authors consider the case when the number of possible types is relatively low. The deterministic 'quasi-species' model (Jain and Krug, 2007) provides a good description when taking $N \to \infty$ whilst the number of possible types remain finite. Distant fitter types are populated by rare long ranged mutations which grow until they become a dominant species and provide the source for new expansions until a fittest population is reached. Evolution occurs deterministically along a specific path towards the most fit type. Our work contrasts this by considering an infinite type space, whereby the behaviour in the infinite population limit remains stochastic.

Our model applies primarily to clonal populations but could apply to all organisms in principle. The conceptual basis for neutral dynamics in large populations is the formation of an effectively neutral network via a rugged fitness landscape produced by gene interactions. The NK model for gene interactions (Kauffman and Levin, 1987) provides one theoretical framework in which very rugged fitness landscapes can occur in sexually reproducing organisms, which has been studied (Ohta, 1997) in the nearly neutral model of molecular evolution for low mutation rates.

In many other theoretical studies a large population $N \to \infty$ is assumed. This limit is not consistent with the assumption of small mutation $Np_m \ll 1$ and therefore a distribution of types must be considered. Under weak selection, we found stochastic evolution models are more typically appropriate to genotypes (justifying the Coalescent (Donnelly and Tavare, 1995)) or perhaps complex ecological traits (relating to Ecological Neutral Theory (Hubbell, 2001)). Deterministic evolution models such as Adaptive Dynamics (Waxman and Gavrilets, 2005) are more appropriate to simple phenotypic traits or when selection is strong or directed. Our model therefore provides important links between previously unrelated models.

5.2 Biological implications

The reproductive ability of real organisms is not well described by a static fitness landscape over evolutionary time. Reproductive success is caused by a wide range of features including environmental effects and interaction of individuals. However, each of the landscapes discussed may qualitatively describe individuals in a population for a time. The linear fitness landscape describes directional selection, the top-hat fitness landscape describes stabilising selection, and the random fitness landscape is appropriate when there is fitness variation without long range trends. We would therefore expect to observe each behaviour only in a subset of type space for a given evolutionary environment.

The important contribution from our model is a theoretical justification for the widespread use of stochastic models of genetic evolution for large populations. Neutral or nearly neutral evolution can be a reasonable assumption for large clonally reproducing populations even when there is fitness variation between types. It is important to integrate the effects of large mutation rates and high population sizes into the current theoretical frameworks. Although population size cannot be inferred from genetics data alone (Stephens, 2007) our model demonstrates that the qualitative nature of dynamics need not change with population size. Hence stochastic models for *changing* population sizes are also reasonable. Our model is not directly applicable to genetics data, but does translate conceptually to problems involving mutation of DNA. Neutral dynamics may naturally occur under different selective conditions for recombining and non-recombining areas, which may be important to inference about mitochondrial DNA (William et al., 1995; Rand et al., 1994) and the Y chromosome (e.g. (Handley et al., 2006)).

The most useful model of biological evolution will differ from situation to situation, particularly depending on the speed of recombination. Our model best describes low recombination rates and therefore asexual populations. It predicts that neutral dynamics can persist for much larger selection strength and population sizes than standard models for sexual species predict. This explains the existence of cryptic asexual species. The fundamental mechanism is simply that mutation rates are large, and not all genes will be good for all types but instead interact to determine fitness. We found that a surprisingly 'large' selective advantage may be present in a population and neutral dynamics can still be observed.

[Data collapse]

The data are collapsed to a single curve for $s < s^*$ by normalising the width and selection. The width is normalised to 1 at zero selection, for which truly neutral dynamics are observed. Selection is normalised to $s' = s/N^\alpha$, where α is a dimension dependent constant to be determined.

To determine α we fit a broken-stick model to the data, i.e. a piecewise linear function with two pieces. This is defined by $\log(w) = c_1$ for $s < s^*$, and $\log(w) = c_2 + c_3 \log(s)$ for $s > s^*$. This is fit by maximum likelihood of the model parameters given the datapoints.

To determine confidence intervals for s^* and α we use a 'bootstrap' method. Since observations were taken at specific values of selection s_i there is uncertainty in the values of w for the s in between. Sample s_i are obtained by perturbing each s_i by an amount x_i, with a 'tent' distribution of mean s_i and extension $(s_{i+1} - s_{i-1})/4$ (i.e. the mean halfway point to its neighbouring points). Sample widths w_i for each value of selection s_i are obtained via a bootstrapping of the n runs (that is, the average of resampling n values with replacement). This provides a distribution of w_i and s_i. These are once again bootstrapped, so that each w_i is sampled and a maximum likelihood s^* is obtained, giving a distribution for s^*. A linear regression is obtained for $\log(s^*)$ as a function of $\log(N)$, giving $-\alpha$ as the slope. Confidence intervals are obtained via the regression variance.

acknowledgments

The work was partially funded by the Scottish Government's Rural and Environment Research and Analysis Directorate (RERAD), and the Engineering and Physical Sciences Research Council (EPSRC).

References

Aita, T., M. Ota, and Y. Husimi, 2003. An in silico exploration of the neutral network in protein sequence space. J. Theor. Biol. 221:599–613.

Aranson, I., L. Tsimring, and V. Vinokur, 1997. Evolution on a rugged landscape: Pinning and aging. Phys. Rev. Lett. 79:3298–3301.

Ash, R. B., 1970. Basic Probability Theory. John Wiley & Sons, Inc., London.

Bailey, N. T. J., 1964. The elements of Stochastic Processes. John Wiley & Sons, Inc, New York.

Bastolla, U., M. Porto, H. E. Roman, and M. Vendruscolo, 2002. Lack of Self-Averaging in Neutral Evolution of Proteins. Phys. Rev. Lett. 89:208101.

Bouchaud, J.-P. and A. Georges, 1990. Anomalous diffusion in disordered media: Statistical mechanisms, models and physical applications. Phys. Rep. 195:127–293.

Charlesworth, J. and A. Eyre-Walker, 2006. The rate of adaptive evolution in enteric bacteria. Mol. Biol. and Evol. 23:1348–1356.

Darwin, C., 1859. The Origin of Species. Penguin, Penguin Books Ltd, 27 Wrights Lane, London W8 5TZ England.

Davison, A. C. and D. Hinkley, 1997. Bootstrap Methods and their Applications. Cambridge University Press.

Donnelly, P. and S. Tavare, 1995. Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. 29:401–421.

Drossel, B., 2001. Biological evolution and statistical physics. Advances in Physics 50(2):209–295.

Felstenstein, J., 1988. Phylogenies from molecular sequences: inference and reliability. Ann. Rev. Genetics 22:521–565.

Fraser, D. A. S., 1976. Probability & Statistics: Theory and Applications. Duxbury Press, North Scituate, Massachusetts.

Gavrilets, S., 1999. A dynamical theory of speciation on holey adaptive landscapes. American Naturalist 154.

Grimmett, G., 1999. Percolation. Springer. Second Edition.

Hahn, M. W., 2008. Toward a selection theory of molecular evolution. Evolution 62:255–265.

Handley, L. J. L., L. Berset-Brő́ndli, and N. Perrin, 2006. Disentangling reasons for low y chromosome variation in the greater white-toothed shrew (crocidura russula). Genetics 173:935–942.

Hubbell, S., 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press, 41 William Street, Princeton, New Jersey 08540.

Huynen, M., P. Stadler, and W. Fontana, 1996. Smoothness within ruggedness: The role of neutrality in adaptation. Proc. Natl. Acad. Sci. 93:397–401.

Jain, K. and J. Krug, 2007. Deterministic and stochastic regimes of asexual evolution on rugged fitness landscapes. Genetics 175:1275–1288.

Kauffman, S. and S. Levin, 1987. Towards a general theory of adaptive walks on rugged landscapes. J. Theor. Biol. 128:11–45.

Kessler, D. A., H. Levine, D. Ridgway, and L. Tsimring, 1997. Evolution of a smooth landscape. J. Stat. Phys. 87:519–544.

Kimura, M., 1983. The neutral theory of molecular evolution. Cambridge University Press, The Pitt Building, Trumpington Street, Cambridge CB2 1RP.

Laird, S. and H. J. Jensen, 2006. The tangled nature model with inheritance and constraint: Evolutionary ecology constricted by a conserved resource. Ecological Complexity 3:253–262.

Lawson, D. J. and H. J. Jensen, 2007. Neutral evolution in a biological population as diffusion in phenotype space: Reproduction with local mutation but without selection. Phys. Rev. Lett. 98:098102.

Ralf Metzler and Joseph Klafter, 2000. The random walk's guide to anomalous diffusion: a fractional dynamics approach. Phys. Rep. 339:1–77.

Moran, P. A. P., 1962. The Statistical Processes of Evolutionary Theory. Clarendon Press, Oxford.

Nei, M., 2005a. Selectionism and neutralism in molecular evolution. Mol. Biol. and Evol. 22:2318–2342.

———, 2005b. Selectionism and neutralism in molecular evolution. Mol. Biol. Evol. 22:2318–2342.

van Nimwegen, E., 2006. Influenza escapes immunity along neutral networks. Science Pp. 1884–1886.

van Nimwegen, E. and J. P. Crutchfield, 2000. Metastable evolutionary dynamics: Crossing fitness barriers or escaping via neutral paths? Bul. Math. Biol. 62:799–848.

van Nimwegen, E., J. P. Crutchfield, and M. Huynen, 1999. Neutral evolution of mutational robustness. Proc. Nat. Acad. Sci. 96:9716–9720.

Ohta, T., 1997. The meaning of near-neutrality at coding and non-coding regions. Gene 205:261{267.

|||, 2002. Near-neutrality in evolution of genes and gene regulation. Proc. Nat. Acad. Sci. 99:16134{16137.

|||, 2003. Origin of the neutral and nearly neutral theories of evolution. J. Biosci. 28:371{377.

Rand, D. M., M. Dorfsman, and L. M. Kann, 1994. Neutral and non-neutral evolution of drosophila mitochondrial dna. Genetics 138:741{756.

Schuster, P., 1997. Genotypes with phenotypes: Adventures in an RNA toy world. Biophysical Chemistry 66:75{110.

Stephens, M., 2007. Inference under the coalescent in D. J. Balding, M. Bishop, and C. Cannings, eds. Handbook of Statistical Genetics, 3rd Ed. John Wiley & Sons, Inc.

Tachida, H., 1991. A Study on a Nearly Neutral Model in Finite Populations. Genetics 128:183{192.

Traulsen, A., J. C. Claussen, and C. Hauert, 2006. Coevolutionary dynamics in large, but finite populations. Phys. Rev. E 74:011901.

Waxman, D. and S. Gavrilets, 2005. 20 questions on adaptive dynamics. J. Evol. Biol. 18:1139{1154.

William, J., O. Ballard, and M. Kreitman, 1995. Is mitochondrial dna a strictly neutral marker? Trends in Ecology and Evolution 10:485{488.

Yi-Cheng Zhang, Maurizio Serva, and Mikhail Polikarpov, 1990. Diffusion Reproduction Processes. J. Stat. Phys. 58:849{861.

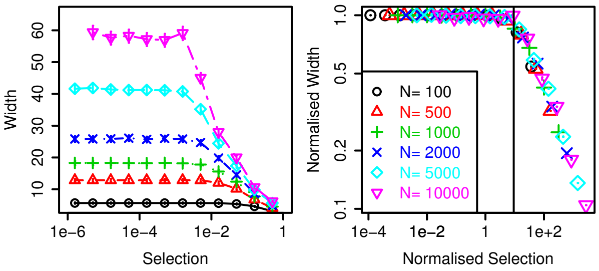

Figure 1: Ensemble average width against selection for fitness landscape 1 (random and uncorrelated) as described by Eq. 1 with mutation probability $p_m = 0.5$, for a range of population sizes N. Left: Ensemble averaged width against selection. Each curve is flat for $s < s_w^*(N)$ i.e. $w(s < s_w^*) = w(0)$. Right: The normalised width against normalised selection for the same data, collapsed using the method from the Appendix. The critical selection $s_w^* \propto N^{-0.94}$ is shown as a vertical line as an aid to the eye. Each datapoint is the time average (over 50000 generations) of a simulation after it has reached equilibrium, ensemble averaged over 200 independent runs with standard deviations calculated using statistical bootstrapping.

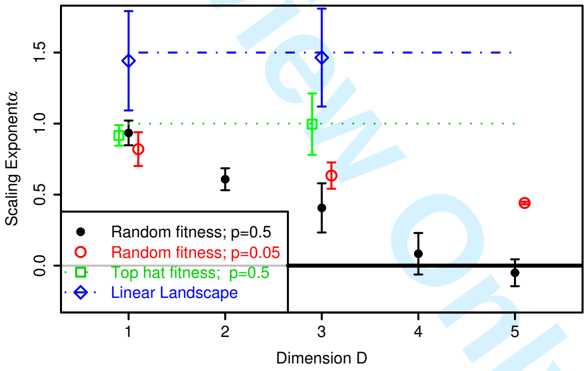

Figure 2: Exponent α for the population size dependence $s_w^* \propto N^{-\alpha}$ as a function of dimension. Error bars are 95% confidence intervals for the regression fit for α (linear regression on a log-log scale for selection s against population N). Shown are the data the for random fitness landscape (Eq. 1) at $p_m = 0.5$ and $p_m = 0.05$, the 'top-hat' correlated fitness landscape (Eq. 2), and the linear fitness landscape (Eq. 3). Dashed lines correspond to theoretical values. Horizontal perturbations to the dimension have been made for visibility and do not reflect fractal dimensions.

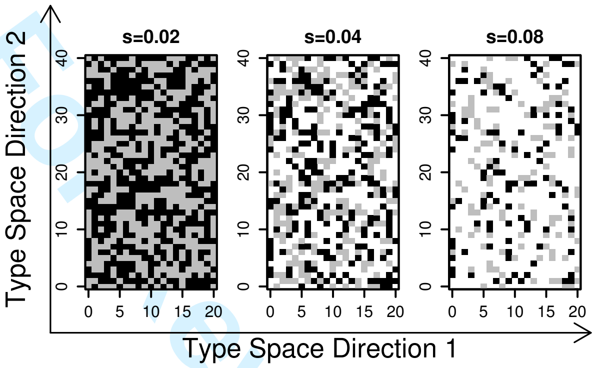

Figure 3: Illustration of 'connectivity' in the random landscape model (Eq. 1) for $D = 2$ and $N = 100$. The fitness landscape itself is shown. Selection values are (left) $s = 0.02$, (middle) $s = 0.04$ and (right) $s = 0.08$. Types with fitness in the range $(0.99, 1)$ compete truly neutrally and are shown in black. Types in grey have fitness in the range $(0.98, 0.99)$ which is high enough to survive by chance for moderate times at low population levels. The low mutation rate neutral regime (left) is characterised by a connected network of neutrally competing types (coloured black). However, the neutral regime considered in this paper (middle) allows linking of neutrally competing types by slightly less fit types (coloured grey). At higher selection, connectivity breaks down into isolated clusters and non-neutral dynamics are observed (right).