The basic science and mathematics of random mutation and natural selection

Alan Kleinman*†

The mutation and natural selection phenomenon can and often does cause the failure of antimicrobial, herbicidal, pesticide and cancer treatments selection pressures. This phenomenon operates in a mathematically predictable behavior, which when understood leads to approaches to reduce and prevent the failure of the use of these selection pressures. The mathematical behavior of mutation and selection is derived using the principles given by probability theory. The derivation of the equations describing the mutation and selection phenomenon is carried out in the context of an empirical example. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: random mutation; natural selection; mathematics; probability theory

1. The mathematical model of mutation and natural selection

The mutation and natural selection phenomenon consists of two components. The mutation component is a random stochastic event that occurs in the replication of a genome. Natural selection, on the other hand, can be either random or non-random. Natural selection is random when there are a variety of causes of death or impaired reproduction of members of a population. For example, some members of the population are killed by accident, others by dehydration, others by starvation, others by predation, and so on. This type of selection is characteristic in genetic drift and is not significant to the evolution of drug resistance, herbicide resistance, pesticide resistance and less than durable cancer treatments and therefore will not be discussed in this paper. On the other hand, non-random selection such as the use of antimicrobial agents, herbicides, pesticides and cancer treatments, which cause the death or impaired reproduction across entire populations in a non-random manner, will be described here.

In particular, the derivation of the equations, which describe mutation and non-random selection (from here on will be simply termed ‘mutation and selection’), will be carried out in the context of a well-measured empirical example of mutation and selection. This example describes the mutations required for a bacterial population to become resistant to an antibiotic. While this example is of particular importance to the use of selection pressures in the practice of treatment of infectious diseases, the principle is more general and can be applied to the evolution of herbicide-resistant weeds, pesticide-resistant insects and failure of cancer treatments.

The mathematical principles used to derive the equations of mutation and selection are obtained from the text, Advanced Engineering Mathematics [1] by Erwin Kreyszig. Governing axioms and principles will be repeated in this paper in order to clarify the derivation of the governing equations.

We start the derivation of the equations in the next section with the empirical example of mutation and selection, which will form the context for the mathematics.

2. The empirical example of mutation and selection

The empirical example that we will use to frame the derivation of the equations, which describe mutation and selection, was published in Science and is titled Darwinian Evolution Can Follow Only Very

PO Box 1240, Coarsegold, CA 93614, U.S.A.
*Correspondence to: Alan Kleinman, PO Box 1240, Coarsegold, CA 93614, U.S.A.
†E-mail: kleinman@sti.net

The copyright line for this article was changed on 22 January 2019 after original online publication. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium provided the original work is properly cited and is not used for commercial purposes.
Abstract: Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins

Five point mutations in a particular ß-lactamase allele jointly increase bacterial resistance to a clinically important antibiotic by a factor of \( \sim 100,000 \). In principle, evolution to this high-resistance ß-lactamase might follow any of the 120 mutational trajectories linking these alleles. However, we demonstrate that 102 trajectories are inaccessible to Darwinian selection and that many of the remaining trajectories have negligible probabilities of realization, because four of these five mutations fail to increase drug resistance in some combinations. Pervasive biophysical pleiotropy within the ß-lactamase seems to be responsible, and because such pleiotropy appears to be a general property of missense mutations, we conclude that much protein evolution will be similarly constrained. This implies that the protein tape of life may be largely reproducible and even predictable.

What this empirical example demonstrates is that the sequence of mutations must occur in an order of ever increasing fitness in order for the evolutionary process to have a reasonable chance of occurring. In addition, this example demonstrates that there is more than a single sequential order, which can occur. In other words, not every member of the population must have the same sequence of mutations in order to evolve resistance to the antibiotic selection pressure. The population of bacteria has subdivided into subpopulations, each taking their own trajectory to achieve resistance to this particular selection pressure.

If we label one subpopulation ‘1’, that subpopulation must get mutation A1 followed by mutation B1, in turn followed by mutation C1, then D1 and finally E1 in order to evolve resistance to the antibiotic selection pressure. If we label another subpopulation ‘2’, that subpopulation must get a different set of mutations, which we can label A2 followed by mutation B2, in turn followed by mutation C2, then D2 and finally E2 in order to evolve resistance to the antibiotic selection pressure. Each of the subpopulations that Weinreich and his co-authors describe has its own set of mutations, which lead to the evolution of a high-resistance ß-lactamase allele. Each of the subpopulations are evolving independently of the other subpopulations. Once a particular subpopulation starts on an evolutionary trajectory, the replication of members from that subpopulation do not contribute to trials for the next beneficial mutation in a different subpopulation on a different evolutionary trajectory.

In order to evolve a high-resistance ß-lactamase allele, a bacterial subpopulation must accumulate five mutations in a particular sequence. We will now address how a subpopulation can accumulate these five mutations. We will drop the numeric and call these five mutations A, B, C, D and E, and the five mutations must occur in that order. The first mathematical question is to identify the trial and possible outcomes for that trial in this stochastic process. We describe this mathematics in the following section.

3. The mathematics of the empirical example of mutation and selection

To derive the mathematical behavior of the empirical example of the mutation and selection phenomenon, we first define some terms.

- \( n \) is the total population size.
- \( n_A \) is the number of members in the subpopulation with mutation A.
- \( n_{GA} \) is the number of generations that the members of the total population reproduce for the probability that mutation A will occur.
- \( n_{GB} \) is the number of generations that the members of the population with mutation A reproduce for the probability that mutation B will occur.
- \( \mu \) is the probability (frequency) that an error in replication will occur at a particular site in a single member in one replication.
- \( \text{P(Beneficial}_A\text{)} \) is the probability that, of all the possible mutations that can occur at the particular site, it will be the beneficial mutation A.
- \( \text{P(Beneficial}_B\text{)} \) is the probability that, of all the possible mutations that can occur at the particular site, it will be the beneficial mutation B.
- \( \text{P(A)} \) is the probability that beneficial mutation A will occur at a particular site.
- \( \text{P(Ac)} \) is the probability that beneficial mutation A will not occur at a particular site.
- \( \text{P(B)} \) is the probability that beneficial mutation B will occur at a particular site.
- \( \text{P(Bc)} \) is the probability that beneficial mutation B will not occur at a particular site.

With these terms defined, we can determine the probability that mutation A will occur on some member of the population. The first step is to identify the trial (the random event) and the possible outcomes for
that trial. In the mutation and selection phenomenon, the mutation is the trial, and the possible outcomes for that trial are the substitution of a base, Ad, Cy, Gu or Th, where Ad represents adenine, Cy cytosine, Gu guanine and Th thymine, and we can denote the probabilities for each of these outcomes as $P(\text{Ad})$, $P(\text{Cy})$, $P(\text{Gu})$ and $P(\text{Th})$, respectively. However, there are other possible outcomes from a mutation. For example, a base can be inserted at the site. Let an ‘i’ preceding the base denote an insertion. Then, $P(i\text{Ad})$ means the probability that an Ad base will be inserted at the particular site and so on for the other bases. If there is a deletion of a base at that site, then $P(\text{del})$ means the probability that a deletion of a base will occur at that site. Additional terms can be added to include any other forms of mutation that can occur.

The outcomes from a mutation are mutually exclusive events. That is, if a substitution of Ad occurs, no other outcomes such as a substitution of a different base, insertion or deletion can occur and so on. Then, to compute the outcomes for a mutation, we can use the addition rule of probabilities for mutually exclusive events. This rule is repeated here from reference [1], p. 715.

If $E_1, \ldots, E_m$ are mutually exclusive events, then

$$P(E_1 \cup E_2 \cup \ldots \cup E_m) = P(E_1) + P(E_2) + \ldots + P(E_m)$$

Then, for our case, the mathematical expression for the possible outcomes for a mutation is

$$P(-\infty < X < +\infty) = P(\text{Ad}) + P(\text{Cy}) + P(\text{Gu}) + P(\text{Th}) + P(i\text{Ad}) + P(i\text{Cy}) + P(i\text{Gu}) + P(\text{del}) + \ldots = 1$$

where ‘…’ represents any other mutation such as the probability of a double deletion or double insertion of the base at that site and any other forms of mutation you might imagine could be included in the sample space.

One could think of these outcomes as equivalent to an unfair die (with more than six sides) where a roll of this die is more likely to give a substitution mutation rather than an insertion or deletion of a base. Then, the probability for a beneficial mutation $A$ occurring on a single member in a single replication is

$$P(A) = P(\text{Beneficial}_A)$$

where $P(\text{Beneficial}_A)$ has a value between 0 and 1.

If the beneficial mutation is the substitution of the Ad base, then $P(\text{Beneficial}_A) = P(\text{Ad})$; if the beneficial mutation is the substitution of a Gu base, then $P(\text{Beneficial}_A) = P(\text{Gu})$; if the deletion of the base is the beneficial mutation, then $P(\text{Beneficial}_A) = P(\text{del})$; and so on. With certainty, we know that $0 \le P(\text{Beneficial}_A) \le 1$.

Equation (2) is the probability that, in any given member of the population in a single replication, the beneficial mutation $A$ will occur at the specific site to improve fitness for that member. The $P(\text{Beneficial}_A)$ term takes into account that not all the possible mutations that might occur at a particular site will be beneficial.

Now, to compute the probability that mutation $A$ will occur at that specific site in a population size ‘n’, we must use the complementary rule of probabilities. This theorem is given on p. 715 of reference [1] and is repeated here.

The probability of an event $E$ and its complement $E^C$ in a sample space $S$ is related by the formula

$$P(E) = 1 - P(E^C)$$

Then,

$$P(A^c) = 1 - P(\text{Beneficial}_A)$$

Equation (3) is the probability that, in any given member of the population in a single replication, the beneficial mutation $A$ will not occur at the specific site to improve fitness for that member.

Then, to compute the probability that a beneficial mutation $A$ will not occur in some member of a population size ‘n’ at a particular site in a single generation, we use the multiplication rule of probabilities.
Again from reference [1], p. 715 and repeated here, we obtain the following theorem for the multiplication rule:

If A and B are events in a sample space S and P(A) ≠ 0 and P(B) ≠ 0, then

\[ P(A \cap B) = P(A)P(B|A) = P(B)P(A|B) \]

If the events A and B are such that P(A \cap B) = P(A)P(B), then they are called independent events.

Equation (3) gives the probability that, in a single generation in a single member of the population, the beneficial mutation will not occur. To determine the probability that the beneficial mutation will not occur in a single generation in a population size 'n' raises the value in equation (3) to the 'n' power, by using the multiplication rule of probabilities, which gives

\[ P(\text{Ac}) = (1 - P(\text{Beneficial}_A)\mu)^n \]  

(4)

Equation (4) gives the probability that, in a population size 'n', the beneficial mutation A will not occur in a single generation of 'n' replications.

Now, to compute the probability that the beneficial mutation A will not occur in a population size 'n' in multiple generations \( n_{GA} \) raises the result in equation (4) to the \( n_{GA} \) power again, by using the multiplication rule, which gives

\[ P(\text{Ac}) = ((1 - P(\text{Beneficial}_A)\mu)^n)^{n_{GA}} = (1 - P(\text{Beneficial}_A)\mu)^{n \cdot n_{GA}} \]  

(5)

Equation (5) gives the probability that the mutation A will not occur in any member in population size 'n' in \( n_{GA} \) generations. To compute the probability that mutation A will occur in some member in the population size 'n' in \( n_{GA} \) generations, we again use the complementary rule and equation (5) to obtain

\[ P(A) = 1 - (1 - P(\text{Beneficial}_A)\mu)^{n \cdot n_{GA}} \]  

(6)

The next step of the calculation of this evolutionary process is to compute the probability that mutation B will occur at the correct site in some member of the population that already has mutation A. The computation just performed for mutation A is analogous except now that the population size for mutation B is limited to those members that already have mutation A. We start this part of the computation by computing the probability that our beneficial mutation B will occur on some member of the population that already has mutation A.

\[ P(B) = P(\text{Beneficial}_B)\mu \]  

(7)

where \( P(\text{Beneficial}_B) \) is defined in the same manner as was \( P(\text{Beneficial}_A) \).

Again, we recognize with certainty that \( 0 \leq P(\text{Beneficial}_B) \leq 1 \).

Equation (7) is the probability that, in any given member of the subpopulation 'n_A' (members of the population that already have mutation A) in a single replication, the beneficial mutation B will occur at the specific site to improve fitness for that member.

Now, to compute the probability that mutation B will occur at the specific site in a population size 'n_A', we must again use the complementary rule of probabilities to compute the probability that mutation B will not occur on some member of the population 'n_A'. Using equation (7) and the complementary rule, we obtain

\[ P(\text{Bc}) = 1 - P(\text{Beneficial}_B)\mu \]  

(8)

Equation (8) gives the probability that, in a single generation in a single member of the population, the beneficial mutation B will not occur. To determine the probability that the beneficial mutation will not occur in a single generation in a population size 'n_A' raises the value in equation (8) to the 'n_A' power, by using the multiplication rule, which gives

\[ P(\text{Bc}) = (1 - P(\text{Beneficial}_B)\mu)^{n_A} \]  

(9)

Equation (9) gives the probability that, in a population size 'n_A', the beneficial mutation B will not occur in a single generation.
Now, to compute the probability that the beneficial mutation \( B \) will not occur in a population size \( n \) in multiple generations \( n \) raises the result in equation (9) to the \( n \) power again, by using the multiplication rule, which gives

\[
P(Bc) = ( (1 - P(Beneficial_B) \mu)^{n \rightarrow n_{GB}} = (1 - P(Beneficial_B) \mu)^{n_A \rightarrow n_{GB}}
\] (10)

Equation (10) gives the probability that the mutation \( B \) will not occur in any member in population size \( n \) in \( n \) generations. Again, using the complementary rule and equation (10), we obtain the probability that mutation \( B \) will occur in at least one member of a population size \( n \) in \( n \) generations:

\[
P(B) = 1 - (1 - P(Beneficial_B) \mu)^{n_A \rightarrow n_{GB}}
\] (11)

Equation (11) gives the probability that the beneficial mutation \( B \) will occur on some member of the subpopulation \( n \) in \( n \) generations. The next step of the calculation of this evolutionary process is to compute the joint probability that mutations \( A \) and \( B \) will occur at the correct sites in some member of the population. Recognizing that mutations \( A \) and \( B \) are independent events, we compute the joint probability that mutation \( B \) will occur on some member with mutation \( A \) using the multiplication rule using equations (6) and (11):

\[
P(A)P(B) = \left(1 - (1 - P(Beneficial_A) \mu)^{n_{A \rightarrow n_{GA}}}\right) \left(1 - (1 - P(Beneficial_B) \mu)^{n_A \rightarrow n_{GB}}\right)
\] (12)

Equation (12) gives the joint probability of mutation \( B \) occurring on a member of the population, which already has mutation \( A \) as a function of population sizes and generations.

This computational scheme can be continued for mutations \( C, D \) and \( E \), and we would obtain the following joint probability equation:

\[
P(A)P(B)P(C)P(D)P(E) = \left(1 - (1 - P(Beneficial_A) \mu)^{n_{A \rightarrow n_{GA}}}\right) \left(1 - (1 - P(Beneficial_B) \mu)^{n_A \rightarrow n_{GB}}\right) \left(1 - (1 - P(Beneficial_C) \mu)^{n_{AB \rightarrow n_{GC}}}\right) \left(1 - (1 - P(Beneficial_D) \mu)^{n_{ABC \rightarrow n_{GD}}}\right) \left(1 - (1 - P(Beneficial_E) \mu)^{n_{ABCD \rightarrow n_{GE}}}\right)
\] (13)

where \( n_{AB} \) is the subpopulation size with mutations \( A \) and \( B \); \( n_{ABC} \) is the subpopulation size with mutations \( A, B \) and \( C \); \( n_{ABCD} \) is the subpopulation size with mutations \( A, B, C \) and \( D \); \( n_{GC} \) is the number of generations for mutation \( C \) to occur; \( n_{GD} \) is the number of generations for mutation \( D \) to occur; \( n_{GE} \) is the number of generations for mutation \( E \) to occur; and \( P(Beneficial_C), P(Beneficial_D) \) and \( P(Beneficial_E) \) are the corresponding probabilities for the beneficial mutations \( C, D \) and \( E \), which will occur at a particular site when a mutation does occur.

Equation (13) gives the probability of accumulation of the five mutations, \( A, B, C, D \) and \( E \), as a function of subpopulation size, number of generations and mutation rate for the evolutionary process to occur. This mathematical behavior of mutation and selection gives rise to specific requirements for an evolutionary process to have a reasonable probability of occurring. These requirements are described in the discussion section, which follows.

4. Discussion of the mathematics of mutation and selection

Equation (13) gives rise to specific requirements for an evolutionary process by mutation and selection to have a reasonable probability to occur. Each step in the evolutionary process occurs in a cyclical manner. Mutation \( A \) occurs somewhere in the population and forms a new subpopulation of members with that mutation \( A \). However, until that subpopulation size increases over generations, the probability that mutation \( B \) will fall on some member with mutation \( A \) will be small. There are simply not enough trials occurring until the number of members in that subpopulation increases over the generations. The number of trials for the particular mutation increases as the population size increases and the number
of generations that the subpopulation is able to replicate. Each of the components of equation (13) is of the form:

\[ P(X) = (1 - (1 - P(Beneficial)\mu)^{n_G}) \]  

(14)

\( P(Beneficial)\mu \) will always be a small number making the value of \( (1 - P(Beneficial)\mu) \) close to but slightly less than 1, and for small values of \( n \) and \( n_G \), \( (1 - P(Beneficial)\mu)^{n_G} \) will be close to 1 and the probability for that component will be close to 0. As \( n \) and \( n_G \) increase, \( (1 - P(Beneficial)\mu)^{n_G} \) will decrease approaching 0, and the probability for that component will approach 1. The following graph illustrates this where the probability of a mutation occurring is plotted as a function of number of members and number of generations for various values of \( P(Beneficial)\mu \) (Figure 1).

When \( P(beneficial)\mu \) is less than \( 1 \times 10^{-7} \), there must be at least 100,000 replications of members who would benefit from the particular mutation before the probabilities of the beneficial mutation occurring starts to increase significantly. Even when there are 1,000,000 replications, the probability of that beneficial mutation occurring is still only about 0.1.

This cycle of beneficial mutation, followed by amplification of the beneficial mutation, must repeat itself over and over in order for the evolutionary process to have a reasonable probability to occur. Mutation B will not have a reasonable probability of occurring on a member with mutation A until the number of members with mutation A increases and/or the number of generations that members with mutation A can replicate becomes large. Only when the number of members with mutation A and the number of generations that members with mutation A can replicate reach a sufficient amount, there will be a reasonable probability that mutation B will occur on some member with mutation A. And mutation C will not have a reasonable probability of occurring on a member of the subpopulation with mutations A and B until those members with mutation A and B can increase in number sufficiently and/or replicate for a sufficient number of generations for the mutation C event to occur.

This cyclical process of beneficial mutation, followed by amplification of beneficial mutation, is different than the evolutionary process discussed by Haldane in his classic paper *The Cost of Natural Selection* [3]. In this paper, Haldane proposes that ‘The principle unit process in evolution is the substitution of one gene for another at the same locus’. The Weinreich example from reference [2] demonstrates that evolutionary processes can lead to multiple different alleles. And the evolutionary process does not consist of a substitution of one allele for another but is an amplification process where a beneficial mutation (giving a more beneficial allele) must amplify (the members with that allele must increase in number) for there

![Figure 1](image-url)
to be a reasonable probability that another beneficial mutation occurs on a member with the previous beneficial mutation.

Any disruption of this beneficial mutation/amplification of beneficial mutation cycle will stifle the evolutionary process. Introducing a second selection pressure, which targets a different genetic locus than the first selection pressure along with the first selection pressure, forces the population to attempt to take two evolutionary trajectories simultaneously and will disrupt the evolutionary cycle. It is this principle, which has led to the success of combination therapy for the treatment of HIV [4]. Likewise, if Weinreich and his co-authors had used two antibiotics where the second drug targets a different genetic locus instead of his single beta-lactam drug, the evolutionary process that these authors measured in their experiment would have been disrupted. Even if a beneficial mutation for the first drug were to occur, the second drug would disrupt the amplification process for that beneficial mutation for the first drug. Likewise, amplification of a beneficial mutation for the second drug would be disrupted by the selection pressure of the first drug. This is a consequence of the basic science and mathematics of the mutation and selection phenomenon and the multiplication rule of probabilities.

5. Glossary

Selection or selection pressure. Stressors on a population that kills or impairs the reproduction of some or all members of a population.

Mutation. An error in the replication of a genetic sequence (DNA/RNA). This error can be a substitution of a base, insertion of a base or bases, deletion of a base or bases or any other errors in replication at a particular location in the genetic sequence. Most of these errors are detrimental; that is, they reduce the fitness to reproduce of that member, but occasionally, one of the kinds of mutations will improve the fitness to reproduce.

Mutation rate. The frequency or probability that an error will occur at a particular site in the genetic sequence during a single replication.

References

1. Kreyszig E. Advanced Engineering Mathematics 3rd ed. John Wiley and Sons Inc.: New York, London, Sydney, Toronto, 1972.
2. Daniel MW. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 7 2006; 312 (5770): 111–114.
3. Haldane JBS. The Cost of Natural Selection. In Journal of Genetics, Vol. 55. Blackwell Publishing (Wiley-Blackwell): New York, London, Sydney, Toronto, 1957; 511–524, (Available from: http://www.blackwellpublishing.com/ridley/classictexts/haldane2.pdf) [Accessed on 12 September 2014].
4. David H. FRONTLINE Interview. (Available from: http://www.pbs.org/wgbh/pages/frontline/aids/interviews/ho.html) [Accessed on 12 September 2014].