Quantitative Biology 2013, 1(1): 9–31
DOI 10.1007/s40484-013-0003-5

QB

REVIEW

Mathematics, genetics and evolution

Warren J. Ewens*

Department of Biology and Statistics, The University of Pennsylvania, Philadelphia, PA 19104, USA
* Correspondence: wewens@sas.upenn.edu

Received October 8, 2012; Revised October 22, 2012; Accepted November 6, 2012

The importance of mathematics and statistics in genetics is well known. Perhaps less well known is the importance of these subjects in evolution. The main problem that Darwin saw in his theory of evolution by natural selection was solved by some simple mathematics. It is also not a coincidence that the re-writing of the Darwinian theory in Mendelian terms was carried largely by mathematical methods. In this article I discuss these historical matters and then consider more recent work showing how mathematical and statistical methods have been central to current genetical and evolutionary research.

INTRODUCTION

A brief description of the Darwinian theory of evolution by natural selection is as follows. In his revolutionary book, generally called *The Origin of Species*, Darwin [1] claimed that biological evolution arises by natural selection, operating on the variation that exists between the individuals in any biological population. The argument has four main components. First, the more fit individuals in the population leave disproportionately more offspring than the less fit individuals. Second, the offspring in large measure inherit the fitness of their parents. Third, and following from the first two points, the offspring generation is on average more fit than the parental generation. Finally, as generation succeeds generation, the population steadily becomes more and more fit, and eventually the more fit types replace the less fit. The changes in the population are often taken as being quite gradual, and may well be thought of as taking place over a time span of many thousands or even hundreds of thousands of years. This is a population level theory. There is no concept of the evolution of the individual. The individual himself does not evolve. It is the population that evolves as generation succeeds generation.

Clearly the existence of variation in the population is crucial to the argument. Without this variation there are no fitness differentials between individuals, and the selective process cannot proceed. Darwin considered variation from one person to another in physical, mental and other characteristics, but when his argument is recast in genetical terms, as it will be below, it will be necessary to measure variation at the genetic level, and to assess the extent to which his theory continues to hold when investigated at the level and in terms of the modern molecular notion of a gene.

The main problem with the theory, as it was presented by Darwin in 1859, was that at that time the hereditary mechanism was unknown. However, the nature of this mechanism is crucial to a complete understanding of the argument. Worse than this, insofar as any idea of a theory of heredity was known in 1859, the most prevalent theory was based on the idea that any characteristic of a child, for example his blood pressure, is a mixture or blending of that characteristic in that child's parents, plus or minus some small deviation deriving from unknown random effects. It is easy to see that under this so-called "blending" theory an effective uniformity of any characteristic among the individuals in the population will soon arise, so that after no more than about ten or twenty generations there will be essentially no individual-to-individual variation in any characteristic available for natural selection to act on. This difficulty was raised soon after the Darwinian theory was put forward, and was recognized immediately as a major criticism of the theory, and (unfortunately) Darwin amended later editions of his book in the light of it. To his dying day Darwin did not know of the resolution of the "variation preserving" difficulty.

Clearly some modification of the argument is necessary, since we do not observe, in present-day populations, the uniformity of characters that the blending theory predicts. However, any modification to the blending

theory would probably require the assumption that the characteristics of children do not closely resemble those of the parents, and would thus remove one of the main underpinnings of the Darwinian theory.

The above discussion brings us to Mendel. Mendel's work [2] appeared in 1866, seven years after the appearance of *The Origin of Species*. It was in effect unread, and its importance unappreciated, until it was rediscovered in 1900. It led, however, to the solution of the maintenance of variation puzzle that Darwin could not solve.

Mendel made the first and basic steps in elucidating the hereditary mechanism. As is well known, he considered seven characters in peas, each of which happened to have a simple genetic basis. For example, he found that seed color, either green or yellow, is determined by the genes at a single gene locus, at which arose the "green" and the "yellow" alleles. We will call the green/yellow gene concept the billiard ball paradigm — a gene is seen under this paradigm as being either a green or a yellow billiard ball, with no known internal structure. In the first half of the 20th century the billiard ball paradigm of the gene was all that was available. The resolution of the variation problem was made using this paradigm for the gene, and it still holds even today when the DNA structure of the gene is known. To see how this was so we turn to some mathematical consequences of the billiard ball paradigm.

MATHEMATICAL IMPLICATIONS OF MENDELISM: THE RANDOM-MATING CASE

Introductory concepts

Our focus throughout is on *diploid* organisms, such as Man, in which any individual receives half his genetic composition from his mother and half from his father. We consider only autosomal loci: the sex-linked case involves minor differences from the analysis below. Consider some specific gene locus on some specific chromosome, which we call locus "A". In general, genes of many different types might occur at this locus. Suppose for the moment that only two types of genes can arise at this locus, which we call the A_1 allele and the A_2 allele. (The words "gene" and "allele" often become confused in the literature. Here we adhere to the concept that a gene is a physical entity while the word "allele" refers to a gene type. Despite this we often use the expression "gene frequency" instead of the more logical "allele frequency", since the term "gene frequency" has become embedded in the literature.) There are three possible genotypes, A_1A_1, A_1A_2 and A_2A_2. As mentioned above, the Darwinian theory is a population-level theory, and it is therefore necessary to consider the population frequencies of these genotypes, and how they change with time. We start with arbitrary frequencies in generation 1, as shown in the table below, and consider what happens in succeeding generations under the assumptions of random mating, no mutation, no selective differences, and indeed no complicating elements of any kind. It is easy to see that the genotype frequencies in generations 1, 2 and 3 are as given below:

	A_1A_1	A_1A_2	A_2A_2
frequency, generation 1	X_{11}	$2X_{12}$	X_{22}
frequency, generation 2	x^2	$2x(1-x)$	$(1-x)^2$
frequency, generation 3	x^2	$2x(1-x)$	$(1-x)^2$

(1)

In this table $x = X_{11} + X_{12}$, so that x is the frequency of the allele A_1 in generation 1. The entries in this table show three important features. First, the genotype frequencies attained in generation 2 are of a binomial form, with x, the frequency of A_1 in generation 1, being the parameter of this distribution. Second, elementary calculations show that the frequency of A_1 in generation 2 is also x. Finally, the genotype frequencies achieved in generation 2 continue to hold in generation 3, and hence also hold in all future generations. This final observation shows that there is no tendency for the variation in the population to be dissipated. These elementary calculations, first made independently by Hardy [3] and Weinberg [4] in 1908, just a few years after Mendelism was re-discovered, show that under a Mendelian hereditary system, Darwinism is saved: the variation needed for the operation of the Darwinian theory is preserved. A little bit of mathematics has gone a long way!

The "Hardy-Weinberg", or binomial, form of the genotype frequencies in the above table, from generation 2 onwards, arise (in a randomly mating population) even when selective differences between genotypes arise, provided that genotype frequencies are taken at the time of conception of any generation. Thus Hardy-Weinberg frequencies will be used in the analysis of the selective case below, it then being understood that all frequencies are taken at this stage of the life cycle. In Genetics courses the emphasis is often placed mainly on the binomial form of the Hardy-Weinberg frequencies, which, conveniently, depend on the single quantity x. The importance of the permanence of these frequencies (at least in cases where there is no selection) to the Darwinian paradigm is often not even mentioned. Yet this permanence is the true implication of the Hardy-Weinberg scheme. A similar result hold for haploid populations: again, since the gene is the unit of transmission, genetic variation is preserved in these populations also.

Of course genetic variation might eventually be lost through the action of selection or random drift (discussed below), but the time-scales for changes brought about by these is much longer than that appropriate to the blending theory.

It is worthwhile asking why variation is preserved in the Mendelian system. The basic reason is a "quantal" one: a parent of genotype A_1A_2 does not pass on, say, two-thirds of an A_1 gene and one-third of an A_2 gene, but an entire A_1 gene or an entire A_2 gene. It is also possible to argue that the "quantal" Mendelian mechanism is the only hereditary system that preserves variation, and thus allows the evolution of superior forms by natural selection. Thus it can be argued that if intelligent, that is evolved, creatures exist elsewhere in the universe, they also will have a Mendelian hereditary mechanism.

Selection and mutation

So far there has been no mention of selection or of mutation. Selection occurs when the various genotypes have different fitnesses. These differences can arise for various reasons, in particular through different viabilities, different fertilities and different mating success probabilities for the various genotypes. The theory for the latter two forms of selection is complex, and here we focus on differential viability fitnesses, that is on different capacities of the various genotypes to survive from the time of conception to the age of reproduction.

Suppose then that the fitness of any individual depends on his genotype at a single locus A, and (as above) that two alleles, A_1 and A_2, and thus three genotypes, are possible at that locus. The fitnesses of individuals of these three genotypes are defined in the following table:

genotype	A_1A_1	A_1A_2	A_2A_2
fitness	w_{11}	w_{12}	w_{22}
frequency	x^2	$2x(1-x)$	$(1-x)^2$

(2)

From this table we are able to calculate several quantities, in particular (i) the mean fitness $\overline{w}$ in the population, (ii) the variance σ^2 in fitness, (iii) the frequency x' of A_1 in the offspring generation and from this the change $\Delta x = x' - x$ in the frequency of A_1 between parental and offspring generations, and finally (iv) the change $\Delta\overline{w}$ in the mean fitness of the population between parental and offspring generations. These values are found by straightforward algebra and are given below:

$$\overline{w} = w_{11}x^2 + 2w_{12}x(1-x) + w_{22}(1-x)^2, \quad (3)$$

$$\sigma^2 = w_{11}^2x^2 + 2w_{12}^2x(1-x) + w_{22}^2(1-x)^2 - \overline{w}^2, \quad (4)$$

$$x' = x\frac{w_{11}x + w_{12}(1-x)}{\overline{w}}, \quad (5)$$

$$\Delta x = x(1-x)\frac{w_{11}x + w_{12}(1-2x) - w_{22}(1-x)}{\overline{w}}, \quad (6)$$

$$\Delta\overline{w} = 2x(1-x)\{w_{11}x + w_{12}(1-2x) - w_{22}(1-x)\}^2 \times\left\{w_{11}x^2 + \left(w_{12} + \frac{1}{2}w_{11} + \frac{1}{2}w_{22}\right)x(1-x) + w_{22}(1-x)^2\right\}\overline{w}^{-2}. \quad (7)$$

Clearly $\Delta\overline{w}$ is non-negative, so we may conclude that natural selection acts so as to increase, or at worst maintain, the mean fitness of the population. This provides a quantification in genetic terms of the Darwinian concept that an "improvement" in the population has been brought about by the action of natural selection.

If the w_{ij} are all close to unity we may write, to a sufficiently close approximation,

$$\Delta\overline{w} \approx 2x(1-x)\{w_{11}x + w_{12}(1-2x) - w_{22}(1-x)\}^2. \quad (8)$$

We return to this approximation later.

Equation (6) gives useful information about the central microevolutionary process, namely the replacement of one allele in a population by another. For example, it is easy to see that if $w_{11} > w_{12} > w_{22}$, the frequency of A_1 steadily increases and asymptotically approaches the value 1. The time $T(x_1, x_2)$ required to increase the frequency of A_1 from some low value x_1 to some higher value x_2 can be found by approximating (6) by the differential equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = x(1-x)\{w_{11}x + w_{12}(1-2x) - w_{22}(1-x)\}, \quad (9)$$

to obtain

$$T(x_1,x_2) = \int_{x_1}^{x_2}\frac{\mathrm{d}x}{x(1-x)\{w_{11}x + w_{12}(1-2x) - w_{22}(1-x)\}}. \quad (10)$$

Note that the denominator $\overline{w}$ in (6) has been replaced by 1 in this calculation, in accordance with the view that, since most population sizes are held at a constant value by extrinsic forces such as food supply, the mean fitness of a population can be taken as 1.

Equation (10) can be solved for $T(x_1, x_2)$, but is best left in the form given, since this is quite informative. For example, the denominator of the integrand on the right-hand side in (10) shows that when the frequency x of A_1 is close to 0 and close to 1, only very slow changes in this frequency will arise through selection.

A second important fitness configuration is that for which

$$w_{12} > w_{11}, \quad w_{12} > w_{22}, \quad (11)$$

that is if there is heterozygote superiority in fitness. Here there is a stable equilibrium at which the frequency of A_1

is x^*, defined by

$$x^* = \frac{w_{12} - w_{22}}{2w_{12} - w_{11} - w_{22}}. \tag{12}$$

When the frequency of A_1 is x^* the mean fitness does not change from one generation to the next. These two examples show that the Mendelian system can explain both evolution (through changes in gene frequencies) or the existence of standing genetic variation, with unchanging gene frequencies.

Mutation is the spontaneous change of a gene from one allele to another. Suppose that the mutation rate from A_1 to A_2 is u and that the mutation rate from A_2 to A_1 is v. It is easy to see that the daughter generation frequency of A_1 is given by

$$\text{daughter generation frequency} = x(1-u) + v(1-x), \tag{13}$$

so that the change in frequency from one generation to the next is

$$\Delta x = v - (u+v)x. \tag{14}$$

This implies that there is an equilibrium where the frequency of A_1 is given by

$$\text{frequency of } A_1 = \frac{v}{u+v}, \tag{15}$$

and it is easy to see that this equilibrium is stable.

When both mutation and selection exist, the change in the frequency of A_1 is, to a first approximation, the sum of the changes given in (6) and (14). Because mutation rates are generally thought of as being smaller than selective differences, the first of these changes is considered the more significant one, except in the neighborhood of an equilibrium point. There is again an equilibrium point of the frequency of A_1 when both selection and mutation exist, and the nature of this point depends on the relative values of the fitnesses w_{ij}. When $w_{11} > w_{12} > w_{22}$, for example, this equilibrium frequency is very close to $1 - u/(w_{11} - w_{12})$. The inferior allele A_2 is maintained in the population only because of the mutation from A_1 to A_2, and this point is called one of selection-mutation balance. When the inequalities (11) hold and mutation exists there will also be a stable equilibrium frequency of A_1. Since mutation rates are very low, this will not differ appreciably from the frequency given in (12).

GENES AND GENOTYPES: THE EVOLUTIONARY "FAULT-LINE"

Despite the fact that the Mendelian hereditary system provides the perfect framework for Darwinian evolution, there remains one serious problem in joining Darwinism with Mendelism. This arises because there is a fault-line in the Mendelian/Darwinian process. A fitness belongs to an individual and depends on that individual's genotype, but an individual passes on a gene, not his entire genotype, to an offspring. This means that for a complete evolutionary investigation of the process it is necessary to isolate that part of the fitnesses of the various genotypes that can be attributed to "genes within genotypes". Doing this was the great achievement of Fisher [5], who showed how to overcome, as far as is possible, the fault-line referred to above. To isolate the "genes within genotypes" component of fitness we approximate the respective fitnesses in (3) by "additive gene (or genetic) fitness" values indicated as follows:

fitness of A_1A_1 fitted by $\overline{w} + 2\alpha_1$,
fitness of A_1A_2 fitted by $\overline{w} + \alpha_1 + \alpha_2$,
fitness of A_2A_2 fitted by $\overline{w} + 2\alpha_2$.

Here α_1 and α_2 can be thought of as additive fitness contributions of the alleles A_1 and A_2 respectively, relative to the mean fitness $\overline{w}$. Of course these approximate values do not represent the true fitnesses of these genotypes, since dominance effects (at the A locus) and epistatic effects (between the A locus and all other loci in the genome) have been ignored. Nevertheless, substantial progress is possible using these approximations, as we now see.

With these approximate fitted values as actual fitnesses, we would compute the population mean fitness by

$$x^2(\overline{w} + 2\alpha_1) + 2x(1-x)(\overline{w} + \alpha_1 + \alpha_2) + (1-x)^2(\overline{w} + 2\alpha_2),$$

which reduces to $\overline{w} + 2[\alpha_1 x + \alpha_2(1-x)]$. This would be equal to $\overline{w}$ if

$$\alpha_1 x + \alpha_2(1-x) = 0. \tag{16}$$

The values of α_1 and α_2 are found by a weighted least squares procedure that is, by minimizing the weighted sum of squares

$$\begin{aligned} &x^2(w_{11} - \overline{w} - 2\alpha_1)^2 \\ &\quad + 2x(1-x)(w_{12} - \overline{w} - \alpha_1 - \alpha_2)^2 \\ &\quad + (1-x)^2(w_{22} - \overline{w} - 2\alpha_2)^2 \end{aligned} \tag{17}$$

with respect to α_1 and α_2. The minimizing values of α_1 and α_2 are found to be

$$\begin{aligned} \alpha_1 &= w_{11}x + w_{12}(1-x) - \overline{w}, \\ \alpha_2 &= w_{12}x + w_{22}(1-x) - \overline{w}. \end{aligned} \tag{18}$$

These values do satisfy (16), so that with the "fitnesses" of the three genotypes given respectively by $\overline{w} + 2\alpha_1$, $\overline{w} + \alpha_1 + \alpha_2$ and $\overline{w} + 2\alpha_2$ and with α_1 and α_2 defined by (18), the true mean fitness is recovered. In other words it was not necessary to impose the condition (16): it is automatically satisfied by the values of α_1 and α_2 defined

by (18). On the other hand equations typified by (16) will be needed later when the whole genome is considered.

The residual sum of squares after these values of α_1 and α_2 are inserted in (17) is called the *dominance variance*, denoted σ_D^2, and is given by

$$\sigma_D^2 = x^2(1-x)^2\{w_{11} - 2w_{12} + w_{22}\}^2. \quad (19)$$

If the actual genotype fitnesses had been of the additive form in which $w_{11} - w_{12} = w_{12} - w_{22}$, a perfect fit to the fitnesses will be obtained by using the three approximating fitnesses $\overline{w} + 2\alpha_1$ for A_1A_1, $\overline{w} + \alpha_1 + \alpha_2$ for A_1A_2 and $\overline{w} + 2\alpha_2$ for A_2A_2. In this case the dominance variance σ_D^2 is zero. In all other cases the fit will not be perfect; that is, one can usually explain some but usually not all of the variation in the fitness values by fitting the α values.

Of far greater importance than σ_D^2 is that part of the variance in fitness that *is* explained by fitting the additive values α_1 and α_2. This is Fisher's [5] *additive genetic variance* and is denoted σ_A^2. It is found as the difference between the total variance (4) and the dominance variance. Straightforward algebra shows that

$$\sigma_A^2 = 2x(1-x)(w_{11}x + w_{12}(1-2x) - w_{22}(1-x))^2. \quad (20)$$

The additive genetic variance thus measures the amount of variation in fitness that *can* be explained by genes within genotypes. It might be expected from the foregoing that the additive genetic variance plays an important role in several evolutionary considerations. This is indeed the case, and we now consider this concept in more detail.

THE ADDITIVE GENETIC VARIANCE

The importance of the additive genetic variance can perhaps best be appreciated by considering some examples of its use. We do this in this section. It is important to remember that the additive genetic variance is a diploid population concept: it has no equivalent for haploid populations.

The Fundamental Theorem of Natural Selection

The Fundamental Theorem of Natural Selection (Fisher [5]) is a very complicated theorem and its full and correct version will be given later. The (incorrect and oversimplified) version of the theorem states that since in a population of constant size we can take $\overline{w}$ to be 1, the change $\Delta\overline{w}$ in mean fitness between parental and daughter generations (as given approximately in (8)) is approximately equal to the additive genetic variance (given in (20)). This is often stated as the Fundamental Theorem of Natural Selection. It is not, however, the correct statement of the theorem. Nevertheless, this misinterpretation of the theorem is interesting, since it shows that Darwinian ideas have to be updated because of these mathematical calculations, which show that because of the Mendelian mechanism it is only the "additive genetic" part of the total variance in fitness that contributes to an increase in mean fitness. Thus it is not accurate to say that evolution by natural selection will occur if there is positive variance in fitness: a more accurate statement is that evolution will occur if and only if there is positive *additive genetic* variance in fitness in the population. Mathematics has played a key role here in up-dating and indeed amending the Darwinian paradigm in the light of Mendelian genetics.

We give the correct version of the Fundamental Theorem of Natural Selection later. For now we note that it is quintessentially a diploid population result. At any given gene locus a parent passes on a single gene to an offspring, not his entire genotype. Thus it is essential to isolate that part of the total variance in fitness that is due to "genes within genotypes". This is precisely what the additive genetic variance does. Evolution by natural selection will not occur if the additive part of the total variance in fitness is zero. This is emphasized in the following section.

Standing genetic variance

The "natural selection" change in gene frequency given in (6) can be contrasted with the situation where the genetic composition of a population does *not* change from one generation to the next. The case where the fitness values are of the form (11) provides one example of this. When the frequency of A_1 is at the value x^* as given in (12) neither of the alleles A_1 and A_2 is more "fit" than the other. At this equilibrium, $\sigma_A^2 = 0$. Thus while there is a positive variance in fitness at this equilibrium point, none of this variation is "additive", or residing within genes. Thus neither allele is superior to the other and no evolution occurs, as is in any event clear since genotype frequencies do not change from one generation to the next when the frequency of A_1 is x^*.

The correlation between relatives

One of the key components of the Darwinian theory is the similarity, or in statistical terms the correlation, between parent and offspring with respect to some measurement, and in particular to fitness. Consider some character determined entirely by the genotype at the locus A, and for which all A_1A_1 individuals have measurement m_{11}, all A_1A_2 individuals have measurement m_{12}, and all A_2A_2 individuals have measurement m_{22}. To find the parent-offspring correlation it is necessary to consider the flow of genes between parent and offspring at the A locus.

Elementary Mendelian arguments show that the genetic combinations between parent and offspring that can arise, with various probabilities, are as indicated in the table below:

		offspring		
	genotype	A_1A_1	A_1A_2	A_2A_2
	A_1A_1	x^3	$x^2(1-x)$	0
parent	A_1A_2	$x^2(1-x)$	$x(1-x)$	$x(1-x)^2$
	A_2A_2	0	$x(1-x)^2$	$(1-x)^3$

(21)

(The values in this table assume that the father and the mother of any individual are unrelated, an assumption that we make for now.) If one considers all possible combinations of parental-offspring measurements and their probabilities, it is found that the correlation between parent and offspring for this measurement is

$$\text{correlation(parent-offspring)} = \frac{1}{2}\frac{\sigma_A^2}{\sigma^2},$$

where σ^2 and σ_A^2 are as given in (4) and (20) respectively, with w_{ij} replaced by m_{ij}. Similar calculations for other relative pairs show that, for example,

$$\text{correlation(sib-sib)} = \frac{1}{2}\frac{\sigma_A^2}{\sigma^2} + \frac{1}{4}\frac{\sigma_D^2}{\sigma^2},$$

$$\text{correlation(uncle-nephew)} = \frac{1}{4}\frac{\sigma_A^2}{\sigma^2},$$

$$\text{correlation(grandparent-grandchild)} = \frac{1}{4}\frac{\sigma_A^2}{\sigma^2}.$$

For our present purposes the main interest in these formulae is in the central role that the additive genetic variance σ_A^2 plays in them. In view of the comments made earlier about the interpretation of the additive genetic variance as describing that part of the variance deriving from genes within genotypes, and the transmission of genes from parent to offspring, it is not surprising that parent-offspring correlation is directly proportional to the additive genetic variance.

A more elegant way of finding these formulae was proposed by Malécot [6]. We consider two individuals X and Y, and define x_f (x_m) as the gene that X received from his father (mother), with a similar definition for y_f and y_m. We say for example that $x_f \equiv y_m$ if the gene that X received from his father is "identical by descent" with the gene that Y received from his mother, where two genes are identical by descent if they can be traced back to a common ancestor gene (for example in a parent or a grandparent). Defining

$$P_{ff} = \text{prob}(x_f \equiv y_f),$$
$$P_{fm} = \text{prob}(x_f \equiv y_m),$$
$$P_{mf} = \text{prob}(x_m \equiv y_f),$$
$$P_{mm} = \text{prob}(x_m \equiv y_m),$$
$$\alpha = \frac{1}{2}(P_{ff} + P_{fm} + P_{mf} + P_{mm}),$$
$$\beta = P_{ff}P_{mm} + P_{fm} + P_{mf},$$

the correlation between any pair of relatives is given by

$$\text{correlation} = \alpha\frac{\sigma_A^2}{\sigma^2} + \beta\frac{\sigma_D^2}{\sigma^2}. \quad (22)$$

This formula can be used to verify the correlations given above. For example, for two sibs, $P_{ff} = P_{mm} = 1/2$, $P_{fm} = P_{mf} = 0$, and insertion of these values into (22) yields the sib-sib correlation given above.

This correlation formula (22) is used frequently in the literature. However it applies only under very restrictive circumstances, namely (i) that the character in question depends only on the genes at one single gene locus, (ii) that there is no contribution to the correlation deriving from the environment, (iii) that the population mates at random, (iv) that stochastic effects have been ignored, and (v) that members of any mating pair are unrelated. Under non-random mating and in cases where the character in question depends on the genes at many loci the formulae become much more complex, as is shown later. All the same, the additive genetic variance plays a key roll in these more complex formulas.

Heritability

The concept of the (narrow) heritability (h^2) is central in plant and animal breeding. In the context described above it is defined as

$$h^2 = \frac{\sigma_A^2}{\sigma^2},$$

where σ_A^2 and σ^2 relate to the additive and total variance in the character in question. This can be regarded as a measure of the extent to which the stock can be improved by breeding. If for example $h^2 = 0$ there is no additive genetic variance in the character and no improvement is possible. In other words the gene frequencies are already at their optimum values as indicated by (12). The larger the heritability the more the stock can be improved by selective breeding.

GENERALIZATION TO THE CASE OF MANY ALLELES

The theory developed above can be generalized in many

ways. In this section we consider one of these, namely the generalization to the case of an arbitrary number of possible alleles at the gene locus A of interest.

Suppose that the number m of alleles that can arise at the locus A exceeds two. There is now a (symmetric) matrix W of fitness values given by

$$\begin{bmatrix} w_{11} & w_{12} & w_{13} & \cdots & w_{1m} \\ w_{21} & w_{22} & w_{23} & \cdots & w_{2m} \\ w_{31} & w_{32} & w_{33} & \cdots & w_{3m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{m1} & w_{m2} & w_{m3} & \cdots & w_{mm} \end{bmatrix},$$

where w_{ij} is the fitness of the genotype A_iA_j. The condition that there be a stable equilibrium with all alleles present at positive frequencies is that W have exactly one positive eigenvalue and at least one negative eigenvalue (Kingman [7]). This is a far less transparent condition than that arising in (11) when $m = 2$, and could not be found other than by mathematical methods.

The simplest example of this result arises in the case where $w_{ii} = 1 - s$, $w_{ij} = 1$, $(i \neq j)$, where s is a (small) positive constant. In this case the eigenvalues of W are $m-s, -s, ..., -s$, the above condition is satisfied, and the equilibrium (where each x_i takes the value m^{-1}) is stable.

If the frequencies of the various alleles are $x_1, x_2, ..., x_m$, the mean fitness $\overline{w}$ of the population is given by $\overline{w} = \sum_i \sum_j w_{ij} x_i x_j$. It can be shown (Kingman [7]) that this increases (or at worst remains unchanged) from one generation to another, in agreement with the corresponding conclusion when $m = 2$.

The additive genetic variance is found by minimizing the expression

$$\sum_i \sum_j \left(w_{ij} - \overline{w} - \alpha_i - \alpha_j\right)^2$$

with respect to parameters $\alpha_1, \alpha_2, ..., \alpha_m$. It is found that the minimizing values of $\alpha_1, \alpha_2, ..., \alpha_m$ are given by

$$\alpha_i = \sum_j w_{ij} x_j - \overline{w}, \ (i = 1, 2, ..., m). \tag{23}$$

These are the direct generalizations of the values given in (18). They also satisfy the generalization of (16), namely that

$$\sum_i x_i \alpha_i = 0. \tag{24}$$

Thus this equation does not have to be imposed extrinsically, as is often done. As with (16) it will have a far greater importance later, when we consider the whole genome.

The additive genetic variance σ_A^2, defined as the sum of squares removed by fitting the values of $\alpha_1, \alpha_2, ..., \alpha_m$, can be expressed in various ways, one of the most useful being

$$\sigma_A^2 = 2 \sum_i x_i \alpha_i^2. \tag{25}$$

Provided that the w_{ij} values are close to each other, the increase in mean fitness is found to be close to the additive genetic variance, thus extending the incorrect and over-simplified form of the Fundamental Theorem of Natural Selection as described above to the case of an arbitrary number of alleles at the locus under consideration.

If we define w_i by $w_i = \sum_j w_{ij} x_j$ the change in the frequency of A_i between successive generations can be conveniently expressed as

$$\Delta x_i = x_i (w_i - \overline{w}) / \overline{w}. \tag{26}$$

This is consistent with the expression given in (6). In general, many results found for the two allele case continue to apply in the multi-allelic case. For example, the correlation between relatives formula (22) continues to hold, with σ_A^2 being defined as in (25) and the total genetic variance being defined appropriately. For this reason, much of the theory focuses on simpler the two-allele case, with the assumption that results found from the analysis will apply exactly, or at least approximately, to the multi-allele case.

THE STOCHASTIC THEORY

All of the above theory is deterministic — there is no component of randomness involved. But random phenomena arise at several levels in evolutionary genetics, from the small-scale intrinsic effect of the random transmission of one of two genes from a parent to an offspring, to large-scale extrinsic random ecological and other events which affect the genetic make-up of an entire population. It is therefore necessary to consider the stochastic theory which takes these random events into account, since among other things it is important to assess the effects of the stochastic behavior on the Darwinian theory.

Before this theory is described, it has to be emphasized that the models used are at best rough approximate descriptions of what happens in reality. The complexity of the real world makes it impossible to formulate models that have the predictive value of many models in physics. For example, we assume in this section that the population of interest mates at random. We later consider more realistic models that take the complexities of nature somewhat more into account.

A key component of the stochastic theory is the size of the population under consideration. The classic stochastic model describing the evolution of the population, the so-called "Wright-Fisher" model (named after its two independent originators Fisher [5] and Wright [8]) is a

Markov chain model and indeed is one of the first examples of these models that was examined in depth. This model is one of simple binomial sampling. Assuming that there is no selection, mutation, geographical dispersal, no complications due to there being two genders, and so on, the model assumes that the genetic composition of an offspring generation is found by random sampling with replacement from the genes of the parental generation until the full offspring generation of genes is obtained.

We adopt the standard notation N for the number of diploid individuals, so that the population total number of genes at the locus A of interest is $2N$. Then if there are i A_1 genes in the parental generation, the probability p_{ij} that there are j such genes in the offspring generation is given by

$$p_{ij} = \binom{2N}{j}\left(\frac{i}{2N}\right)^j\left(\frac{2N-i}{2N}\right)^{2N-j}. \qquad (27)$$

When selection and mutation occur a more general expression, namely

$$\binom{2N}{j}(x')^j(1-x')^{2N-j} \qquad (28)$$

arises, where x' depends on the selective and mutation parameters. In the case where mutation is absent and parental generation fitnesses and genotype frequencies are given in (2) the value of x' is as given in (5), where on the right-hand side of (5) x is replaced by $i/(2N)$. When there is mutation but no selection, x' is given by (13). A more complicated model allows both selection and mutation.

Many other models, for example birth and death models, have been investigated in detail in the literature, but here we consider only models of the binomial type (27) and (28).

The qualitative behavior of the model (27) is that loss of one or other allele will eventually occur, so that genetic variation will no longer exist at the locus in question. It is thus of relevance to the Darwinian theory to assess how long genetic variation may be expected to persist at this locus, assuming that the (very simple) model (27) holds.

Unfortunately, even for the simple model (27), which does not allow for selection, it seems to be impossible to find a simple closed expression for the mean fixation time, that is for the mean time until one or other allele is lost from the population. This problem leads to the approximation of this model, which is in discrete time and discrete space, by a continuous-time continuous-space diffusion process. Approximations of this type have been common in the population genetics literature since the 1920's, and they sometimes have curious mathematical properties not shared by diffusion processes arising in physics. The result obtained by the diffusion process approximation is that if the initial frequencies of the two alleles are p and $1 - p$, the mean time until one or other allele is lost from the population is, to a very close approximation,

$$-4N(p\log p + (1-p)\log(1-p))$$

generations. This implies that although in this model genetic variation will eventually be lost, in a population of any reasonable size the mean time for such loss is generally very large. In practice this time might be so long that other considerations, such as the creation of new variation by mutation, have to be taken into account. In any event the main conclusion, that the "variation preserving" property of the Mendelian system continues to hold, is essentially maintained.

Sometimes interest focuses on conditional mean times. Suppose for example that the condition is made that A_1 is eventually lost from the population. The conditional mean time for this to occur is

$$-4N\frac{p\log p}{1-p}$$

generations. A parallel calculation applies when the condition is made that it is A_2 that is eventually lost.

It is also of interest to find the probability that a selectively favored allele becomes fixed. This probability depends on the population size, the fitness values of the various genotypes and the initial frequency of the favored allele. Again a diffusion approximation (now to the model (28)) is needed. If for example $w_{11} = 1 + s$, $w_{12} = 1 + s/2$ and $w_{22} = 1$, the probability that A_1 becomes fixed depends on s and on the number of genes $2N$ in the population at the locus of interest. The table below gives some typical calculations, assuming that the initial frequency of A_1 is 0.001.

	N	10^4	10^5	10^6
	0.00001	0.001	0.002	0.020
s	0.0001	0.002	0.020	0.181
	0.001	0.020	0.181	0.865
	0.01	0.181	0.865	1.000

(29)

Clearly the larger the population size, the more likely it is that the favored allele becomes fixed. This occurs essentially because in a large population random factors are of less importance than in a small population. Also, of course, the probability of fixation of the favored allele increases with its selective advantage.

Particular interest attaches to the case where there is one single initial A_1 gene, corresponding to a situation where a single initial A_1 mutant arises in an otherwise purely A_2A_2 population. Branching process theory was

employed in the 1920's to calculate the survival probability of this new mutant, and it is found that when the fitnesses are of the form $w_{11} = 1 + s$, $w_{12} = 1 + s/2$, $w_{22} = 1$, this probability is about s. This is clearly independent of the population size.

When mutations from A_1 to A_2 (at rate u) and from A_2 to A_1 (at rate v) both occur, there exists a stationary distribution of the frequency of A_1. When there is no selection, this distribution depends on u, v and the number $2N$ of genes in the population at the A locus. The diffusion theory approximation to this distribution is

$$f(x) = \frac{\Gamma(4Nu + 4Nv)}{\Gamma(4Nu)\Gamma(4Nv)} x^{4Nv-1}(1-x)^{4Nu-1}, \qquad (30)$$

where x is the frequency of A_1. The mean of this distribution is $v/(u+v)$, as might be expected from (15), and the variance is

$$\frac{uv}{(u+v)^2(4Nu + 4Nv + 1)}. \qquad (31)$$

This variance has no analogue in the deterministic analysis, and thus the stochastic analysis provides new information about the likely values of the frequencies of the two alleles under two-way mutation.

The form of the distribution is U-shaped for small population sizes, indicating that for such populations the most likely situation is that one or other allele is quite rare (or even temporarily missing from the population). For large population sizes the distribution is unimodal, and concentrates closely around the mean.

When selection exists the form of the stationary distribution is more complicated, being of the form

$$f(x) = \frac{\Gamma(4Nu + 4Nv)}{\Gamma(4Nu)\Gamma(4Nv)} x^{4Nv-1}(1-x)^{4Nu-1} g(x), \qquad (32)$$

where $g(x)$ depends on the nature of the selective values of the various genotypes.

Generalizations of many of the above results are known when an arbitrary number of alleles m can arise at the locus in question, in particular of the stationary distribution (32). Thus the above results serve only as an introduction to the entire theory that is now known. All of the above ignores many complicating factors that exist in real populations, for example the presence of two sexes, the likely geographical dispersion of the population, changes (either cyclical or monotonic) in the population size, and so on. These complications are often dealt with usefully through the concept of the "effective population size N_e". This is a quite complicated concept and there are several definitions of an effective population size (see for example Ewens [9]). For our purposes here it is sufficient to say that significant properties of the behavior of a population can be found by replacing N in the above formulae by one or other of the definitions of the effective population size. The formula for N_e depends on the complicating factor involved and the definition of N_e used. In the case of two sexes, for example, most definitions lead to an effective population size of about $4N_fN_m/(N_f + N_m)$, where N_f and N_m are, respectively, the number of females and the number of males in the population. This formula implies that, when $N_m = 1$, the effective population size is very close to 4, no matter how many females there are in the population. This happens because half the genes transmitted in any generation derive from a single male, leading to rapid changes in gene frequencies. This implies among other things that there is a rapid loss of genetic variation over succeeding generations, a matter of serious concern to animal breeders.

NON-RANDOM-MATING POPULATIONS

In this section we generalize some of the results described above to the case of non-random-mating populations. Such a generalization is necessary, given the interest in the human population deriving from data from the Human Genome Project (see http://www.ornl.gov/sci/techresources/Human-Genome/project/about.shtml for more on this) and the fact that the human population does not mate at random. Because of the complications involved, we only consider the deterministic theory.

Genotype frequency changes

Suppose that at the time of conception of the parental generation

$$\text{frequency of } A_iA_i = x_{ii}, \ (i = 1, 2, ..., m),$$
$$\text{frequency of } A_iA_j = 2X_{ij}, \ (i \neq j).$$

From this, the allelic frequencies x_i, $(i = 1, 2, ..., m)$ are given by $x_i = \sum_{j=1}^{m} X_{ij}$. (All summations in this analysis are over the range $(1, 2, ..., m)$, so this range is not explicitly mentioned below.)

The mean population fitness at the time of conception of the parental generation is $\overline{w} = \sum_i \sum_j X_{ij} w_{ij}$, where as above the fitness of A_iA_j is w_{ij}, while the variance in fitness is $\sigma^2 = \sum_i \sum_j X_{ij} w_{ij}^2 - \overline{w}^2$, The within-generation change in the frequency of the allele A_i is then

$$\sum_j X_{ij}(w_{ij} - \overline{w})/\overline{w}. \qquad (33)$$

There exist few if any population genetic models in which the frequency of an allele at the time of reproduction in a parental generation differs from the frequency of that allele in the daughter generation at conception. Thus

assuming that this equality holds, the between generation change Δx_i in allelic frequencies is also given by (33), that is,

$$\Delta x_i = \sum_j X_{ij}\left(w_{ij} - \overline{w}\right)/\overline{w}. \tag{34}$$

The additive genetic variance

To find the additive generic variance in the non-random-mating case, it is first necessary to find the average effects $\alpha_1, \alpha_2, ..., \alpha_m$ of the various alleles at the locus of interest. These are found by the natural generalization of the procedure in the random-mating case that is by minimizing the expression

$$\sum_i \sum_j X_{ij}\left(w_{ij} - \overline{w} - \alpha_i - \alpha_j\right)^2$$

with respect to the parameters $\alpha_1, \alpha_2, ..., \alpha_m$, subject to a constraint generalizing (16), namely (in the present notation) $\sum x_i \alpha_i = 0$. (Once again, this constraint is not necessary. However we nevertheless make it because in the multi-locus case considered below a constraint generalizing (16) *is* needed.) The values of $\alpha_1, \alpha_2, ..., \alpha_m$ so found are given implicitly as the solutions of the equations

$$x_i\alpha_i + \sum_j X_{ij}\alpha_j = \overline{w}\Delta x_i, \ (i = 1, 2, ..., m). \tag{35}$$

If we define matrices $\mathbf{D}$ and $\mathbf{X}$ by

$$\mathbf{D} = \text{diag}\{x_1, x_2, ..., x_m\}, \quad \mathbf{X} = \{\mathrm{X}_{ij}\},$$

and vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\Delta}$ by

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_m)',$$
$$\boldsymbol{\Delta} = (\Delta x_1, \Delta x_2, ..., \Delta x_m)',$$

then the various equations in (35) can be written in matrix form as

$$(\mathbf{D} + \mathbf{X})\boldsymbol{\alpha} = \overline{w}\boldsymbol{\Delta}. \tag{36}$$

The additive genetic variance can be written in terms of the average effects as

$$\sigma_A^2 = 2\overline{w}\sum_i \alpha_i \Delta x_i. \tag{37}$$

This differs in general from the "random mating" expression given in (25). However it reduces to (25) when random mating is the case. This occurs because under random mating, $\sum_j X_{ij}\alpha_j = x_i \sum_j \alpha_j x_j = 0$, so that from (35), $\overline{w}\Delta x_i = x_i\alpha_i$. Using this equality in (37), the expression for σ_A^2 becomes $2\sum_i x_i \alpha_i^2$, as in (25).

The Fundamental Theorem of Natural Selection

It was remarked above that the commonly-stated version of the Fundamental Theorem of Natural Selection is not the true version. The incorrect version of the theorem states that for a random-mating population in which fitness depends on the genes at a single locus only, $\Delta\overline{w} \approx \sigma_A^2/\overline{w}$. Thus the incorrect version assumes a random-mating population and provides an approximate statement only. Also, we shall see later that this version of the theorem cannot be true when fitness depends on the genes at two or more loci since in this case mean fitness can decrease, even under random mating. Thus this change in mean fitness then cannot be approximately equal to any form of variance. Thus commonly-stated (but incorrect) version of the theorem has a limited usefulness, since in practice fitness depends on the genes at many loci. Further, many populations of interest, especially the human population, do not mate at random. We now give the correct version of the theorem for the case when fitness depends on the genes at a single locus only, and later generalize the theorem to the case where fitness depends on any collection of genes in the entire genome. Random mating is not assumed.

We rewrite the mean fitness as $\overline{w} = \sum_i \sum_j X_{ij}\left(\overline{w} + \alpha_i + \alpha_j\right)$, and then define the partial change $\Delta_P\overline{w}$ in the mean fitness by

$$\Delta_P\overline{w} = \sum_i \sum_j \Delta X_{ij}\left(\overline{w} + \alpha_i + \alpha_j\right). \tag{38}$$

Then

$$\Delta_P\overline{w} = 2\sum_i \alpha_i \Delta x_i = \sigma_A^2/\overline{w}. \tag{39}$$

This is the correct version of the Fundamental Theorem of Natural Selection when fitness depends on the genes at a single locus only. The statement of the theorem is an exact one, not an approximation, and random mating is not assumed.

The correlation between relatives

Quite apart from its importance in determining the evolution of a population, non-random mating has an important effect on the correlation between relatives. There is a vast literature on this subject, and here only a few comparatively straightforward results will be given.

The fundamental new parameter of interest is ρ, the correlation between mating individuals in the measurement of interest. For a random-mating population, $\rho = 0$. When $\rho \neq 0$ the correlations given in (22) need to be modified. Defining

$$h^2 = \frac{\sigma_A^2}{\sigma^2},$$

$$d^2 = \frac{\sigma_D^2}{\sigma^2},$$

equation (22) implies that in the very simple case considered previously, the correlation between relatives in a random-mating population is of the form $a_1h^2 + a_2d^2$ for some constants a_1 and a_2. However, under non-random mating, we find, for example, that

$$\text{correlation(parent-offspring)} = \frac{1}{2}(1+\rho)h^2,$$

$$\text{correlation(sib-sib)} = \frac{1}{2}(1+\rho h^2)h^2 + \frac{1}{4}d^2,$$

$$\text{correlation(uncle-nephew)} = \left(\frac{1}{2}(1+\rho)h^2\right)^2 h^2 + \frac{1}{8}\rho h^2 d^2.$$

The second and third of these correlations are not of the random-mating form $a_1h^2 + a_2d^2$. Further, a dominance (that is d^2) term now enters uncle-nephew correlations (and many other correlations) whereas under random mating it does not. This implies that conclusions drawn from simple correlations such as those following from (22) can be quite erroneous. Despite this, and maybe through ignorance, correlations as given by (22) and its various generalizations often appear in the literature.

TWO LOCI: THE RANDOM-MATING CASE

We now turn to the case where the fitness and indeed any other characteristic of any individual is assumed to depend on the genes that the individual has two loci, not just one. Although this generalization is only a small step towards considering the case where the individual's fitness depends on his entire genome, it does allow some significant advances to be made beyond the one-locus analysis.

Evolutionary calculations

In this section we assume random mating of the population in question, and restrict attention to the deterministic case. To analyze the joint evolution at two loci, A and B, it is necessary and sufficient (in the case of random mating) to use the frequencies of the various possible *gametes* formed by the alleles at these two loci. For our purposes it is sufficient to assume that the two loci are on the same chromosome, and then to think of a gamete as this chromosome. If there are two alleles (A_1 and A_2) possible at the A locus and two alleles (B_1 and B_2) possible at the B locus, there are four possible gametes, A_1B_1, A_1B_2, A_2B_1 and A_2B_2 defined by these alleles, and we write the frequencies of these in a parental generation as c_1, c_2, c_3 and c_4 respectively.

Because of the phenomenon of crossing over, gametes are not necessarily passed on faithfully from a parent to an offspring. An individual with one chromosome of the gametic form A_1B_1 and the other chromosome of gametic form A_2B_2 will pass on a chromosome of gametic form A_1B_2 or of gametic form A_2B_1 if a crossing over (more exactly, an odd number of crossings over) occurs between the A and the B loci. We denote the probability of such a crossing over by R; in practice it is always the case that $0 \leqslant R \leqslant 1/2$. Assuming no selection at either locus, the recurrence relations defining offspring generation gametic frequencies in terms of parental generation frequencies are

$$c_1' = c_1 + R(c_2c_3 - c_1c_4),$$

$$c_2' = c_2 - R(c_2c_3 - c_1c_4),$$

$$c_3' = c_3 - R(c_2c_3 - c_1c_4),$$

$$c_4' = c_4 + R(c_2c_3 - c_1c_4),$$

where the dash notation refers to the offspring generation frequencies.

One of the important features of a two-locus analysis is that one can assess whether some conclusion found from a one-locus analysis gives results that are consistent with those found for the two-locus analysis. For example, one of the first one-locus analysis results given above is that if there is no selection, gene frequencies remain unchanged from one generation to the next. In the two-locus analysis the offspring generation frequency of A_1 is $c_1' + c_2'$, and the above equations show that this is equal to $c_1 + c_2$, the parental generation frequency of A_1. Thus the one-locus result is confirmed in this more general analysis. However, as shown below, not all one-locus conclusions are confirmed by a two-locus analysis.

Clearly gametic frequencies, unlike gene frequencies, change from one generation to another. To analyze the evolutionary properties of the system governed by the gametic frequency recurrence relations it is useful to introduce the "coefficient of linkage disequilibrium" D, defined by $D = c_1c_4 - c_2c_3$. Then defining η_i by

$$\eta_i = +1, \ (i = 2, 3),$$
$$\eta_i = -1, \ (i = 1, 4),$$

the gametic recurrence relations can be written

$$c_i' = c_i + \eta_i RD, \ (i = 1, 2, 3, 4). \tag{40}$$

The recurrence relation for D itself is easily shown to be $D' = (1-R)D$, so that $D^{(t)} = (1-R)^t D$. Clearly the value of D decreases geometrically fast as time goes on. When $D = 0$, the frequency of any gamete is the product of the frequencies of its constituent alleles, and the two loci evolve in effect independently. A state of "linkage equilibrium" has been reached.

A different picture emerges when selective differences exist between the nine genotypes possible at the two loci. Suppose that the fitnesses are as given in (41), and that at the time of conception of the parental generation the frequencies of the various genotypes are as given in (42).

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	w_{11}	w_{12}	w_{22}
A_1A_2	w_{13}	$w_{14} = w_{23}$	w_{24}
A_2A_2	w_{33}	w_{34}	w_{44}

(41)

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	c_1^2	$2c_1c_2$	c_2^2
A_1A_2	c_1c_3	$2(c_1c_4 + c_2c_3)$	$2c_2c_4$
A_2A_2	c_3^2	$2c_3c_4$	c_4^2

(42)

The mean fitness $\overline{w}$ of the population is then given by

$$\overline{w} = \sum_i \sum_j c_i c_j w_{ij},$$

and it is straightforward to show that the gametic frequency recurrence relations are

$$c_i' = \overline{w}^{-1}(c_i w_i + \eta_i R w_{14} D),$$

where $w_i = \sum_j c_j w_{ij}$ and η_i is as defined above. These recurrence relations have some quite interesting properties. For example, at equilibrium ($c_i' = c_i$),

$$\overline{w} = w_i + c_i^{-1}\eta_i R w_{14} D, \ (i = 1, 2, 3, 4),$$

whereas mean fitness is maximized when

$$\overline{w} = w_i, \ (i = 1, 2, 3, 4).$$

Thus mean fitness is not necessarily maximized at an equilibrium point, and indeed is maximized at such a point only if, at this point, the coefficient of linkage disequilibrium D is zero. This condition rarely holds in practice when selection exists, and an immediate implication of this is that the course of natural selection is such that mean fitness does *not* necessarily increase from one generation to the next. The reason for this is that a parent can pass on a gamete to an offspring that he does not possess himself, and to this extent the genetic make-up of the child does not resemble that of the parent. Thus the simple version of the Fundamental Theorem of Natural Selection given above, that mean fitness increases from one generation to the next, cannot be true.

When this observation was first made it caused consternation among the population genetics community, since it seemed to imply that the (textbook version of the) Fundamental Theorem of Natural Selection is not true for a two-locus system (and, a fortiori, for the entire genome). This quickly led to many further developments, for example an investigation of cases of fitness arrays for which mean fitness did necessarily increase from one generation to the next. It was shown that one such array is the "additive over loci" fitness case, where the fitnesses are of the additive form shown in (43).

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	$\theta_1 + \lambda_1$	$\theta_1 + \lambda_2$	$\theta_1 + \lambda_3$
A_1A_2	$\theta_2 + \lambda_1$	$\theta_2 + \lambda_2$	$\theta_2 + \lambda_3$
A_2A_2	$\theta_3 + \lambda_1$	$\theta_3 + \lambda_2$	$\theta_3 + \lambda_3$

(43)

A further line of enquiry concerns the additive genetic variance in the case where fitness depends on the genes at two loci. This is found by a least-squares procedure, using the fitnesses in (41) and the frequencies in (42). This is not described in detail here since a much more complete and general discussion is given later.

One can also calculate an additive genetic variance estimate found from the marginal fitnesses at the various loci. Thus the marginal fitness of A_1A_1 is

$$\frac{w_{11}c_1^2 + 2w_{12}c_1c_2 + w_{22}c_2^2}{(c_1 + c_2)^2},$$

with a similar definition for other A-locus genotypes. The marginal A-locus additive genetic variance is then found using the procedure that led to (19). A similar calculation can be done for the B locus, and from this one can calculate the sum of the single locus marginal values. It is found that the condition that the true additive genetic variance be equal to the sum of the single locus marginal values is that $D = 0$.

The correlation between relatives

The correlation structure given in (22) and its various generalizations for the case when the character of interest is determined by the genes at one locus, and random mating of the population is assumed, requires considerable amendment when the character of interest is determined by the genes at two loci. We now discuss some of the complications involved.

Suppose that the measurement in question depends on the genes that an individual has at two loci, A and B. There are now nine genotypes and thus nine potentially

different measurements, as shown in the table below.

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	m_{11}	m_{12}	m_{22}
A_1A_2	m_{13}	$m_{14}=m_{23}$	m_{24}
A_2A_2	m_{33}	m_{34}	m_{44}

(44)

A collection of nine measurements implies a total of eight degrees of freedom between measurements. Thus the total variance σ^2 in the measurement can be split up into eight different meaningful components. Assuming for the moment that mating is random and linkage equilibrium ($D = 0$) between loci occurs, σ^2 can be split up into eight components in the form

$$\sigma^2=\sigma_A^2(1)+\sigma_A^2(2)+\sigma_{AA}^2+\sigma_D^2(1)+\sigma_D^2(2)+\sigma_{DD}^2+\sigma_{(AD)}^2+\sigma_{(DA)}^2.$$

In this decomposition $\sigma_A^2(1)$ is the additive genetic variance at the A locus as calculated from marginal measurement values, $\sigma_A^2(2)$ is the corresponding quantity for the B locus, dominance variances are defined similarly, while all other variances relate to various interactions. It is convenient to write

$$\sigma_A^2=\sigma_A^2(1)+\sigma_A^2(2),$$

$$\sigma_D^2=\sigma_D^2(1)+\sigma_D^2(2),$$

$$\sigma_{AD}^2=\sigma_{(AD)}^2+\sigma_{(DA)}^2,$$

so that

$$\sigma^2=\sigma_A^2+\sigma_D^2+\sigma_{AA}^2+\sigma_{DD}^2+\sigma_{AD}^2.$$

When A and B loci are unlinked ($R = 1/2$), it is found that the parent-offspring correlation in the measurement is

$$\left(\frac{1}{2}\sigma_A^2+\frac{1}{4}\sigma_{AA}^2\right)\Big/\sigma^2,$$

and that the sib-sib correlation is

$$\left(\frac{1}{2}\sigma_A^2+\frac{1}{4}\sigma_D^2+\frac{1}{4}\sigma_{AA}^2+\frac{1}{8}\sigma_{AD}^2+\frac{1}{16}\sigma_{DD}^2\right)\Big/\sigma^2.$$

When A and B loci are linked ($R\neq 1/2$) the parent-offspring remains as above, but the sib-sib correlation becomes

$$\left(\frac{1}{2}\sigma_A^2+\frac{1}{4}\sigma_D^2+\frac{1}{8}(3-4R+4R^2)\sigma_{AA}^2+\frac{1}{4}(1-2R+2R^2)\sigma_{AD}^2+\frac{1}{4}(1-2R+2R^2)^2\sigma_{DD}^2\right)\Big/\sigma^2.$$

It is clear that even in the comparatively simple case of the sib-sib correlation, linkage causes complications to the correlation formulae. This becomes increasingly the case for more distant relationships. Thus complications to correlation formulae are evident even under the simplifying assumptions of random mating and linkage equilibrium. In the more general case with linkage disequilibrium ($D\neq 0$), linked loci and with non-random mating, it is almost impossible to find expressions for the various correlations, and theoretical work on this problem has essentially come to a halt. It is even more difficult to use correlations to estimate parameters such as heritability, and one is forced to rely on various approximation formulae. Theory, it appears, is of limited help so far as correlation questions are concerned.

MANY LOCI: RANDOM-MATING AND NON-RANDOM-MATING POPULATIONS

Evolutionary considerations

In this section we consider what can be said about the completely general case where the fitness of any individual depends in an arbitrary way on all the genes in the entire genome, random mating is not assumed, and the recombination structure in the entire genome is completely arbitrary. Unfortunately it is not possible to derive the intra-generational changes in whole genome genotypes in the absence of knowledge of the mating scheme and the recombination structure between loci. On the other hand it is possible to find the intragenerational changes in gene frequencies, and from this to the intergenerational changes in gene frequencies, and then to derive the correct whole-genome Fundamental Theorem of Natural Selection. That is our aim in this section.

Assume that the various (gigantically large number of) possible whole-genome genotypes are listed in some agreed order as genotypes $1, 2, ..., s, ..., S$. The "time of conception" frequency of the typical genotype s in the parental generation is denoted by g_s and the fitness of this genotype by w_s. Thus at this time the parental generation population mean fitness is $\overline{w}=\sum_s g_s w_s$. From this it is possible to find the intragenerational changes in gene frequencies, and from this to derive the whole-genome Fundamental Theorem of Natural Selection. The intra-generational change in the frequency of the genotype s, that is the change from the time of conception to the time of reproduction, is $\Delta g_s = g_s w_s/\overline{w} - g_s$. This implies that the corresponding intra-generational change Δp_{ai} in the frequency p_{ai} of the allele A_i at the typical gene locus A is given by

$$2\Delta p_{ai}=\sum_s c_{ai(s)}\Delta g_s, \tag{45}$$

where the sum is taken over all whole-genome genotypes and $c_{ai(s)} = 1$, 2 or 0 depending on whether A_{ai} arises once, twice or not at all within the genotype g_s. This is also the inter-generational change in the frequency of this allele, where frequencies are taken at the time of conception of the two generations. This is all that can be said about inter-generation frequency changes. It is, however, enough for our purposes.

Average effects

To make further progress it is necessary to define whole-genome average effects. As in the one locus case, the average effects of the various alleles at the various loci in the genome are defined by minimizing the sum of squares

$$\sum_s g_s \left\{ w_s - \overline{w} - \sum c_{ai(s)} \alpha_{ai} \right\}^2. \tag{46}$$

In the expression (46), α_{ai} is the average effect of A_i, the outer sum is taken over all whole-genome genotypes and the inner sum is taken, for each whole-genome genotype, over all alleles at all loci in the genome contained within that genotype, with c_{ai} defined above. The values of the α_{ai} defined in this way are not unique, but become unique if a condition of the form (24) is imposed for the average effects of the alleles at each locus in the genome.

It is not necessary to give explicit formulae for the various α values defined by this least-squares procedure: indeed they can only be expressed implicitly as the (unique) solution of a gigantic set of simultaneous equations. In parallel with the one-locus case analysis, the mean fitness of the population is now thought of as being

$$\sum_g g_s \left\{ \overline{w} + \sum c_{ai} \alpha_{ai} \right\}, \tag{47}$$

where now α_{ai} denotes the minimizing α values. This expression (as with the expression in the corresponding one-locus case) is numerically identical to that given by the standard definition of mean fitness $\sum_s g_s w_s$ as given above.

The whole-genome additive genetic variance σ_A^2 is the sum of squares removed by fitting the α values. Standard least-squares theory shows that

$$\sigma_A^2 = 2\overline{w} \sum_a \sum_i \alpha_{ai} (\Delta p_{ai}),$$

the sum being over all loci (a) in the genome and all alleles (i) at each locus.

The Fundamental Theorem of Natural Selection: the correct whole-genome statement

Again in parallel with the one-locus case, the partial change $\Delta_P \overline{w}$ in mean fitness is defined as the change in the expression (23) derived solely from the changes Δg_s in the various whole-genome genotype frequencies and ignoring the changes in the α values, namely

$$\sum_s \Delta g_s \left\{ \sum c_{ai} \alpha_{ai} \right\}. \tag{48}$$

The resulting expression can be shown to be equal to $\sigma_A^2 / \overline{w}$, where σ_A^2 now denotes the whole-genome additive genetic variance, defined in a manner extending that for the one-locus case. This simple and exact result is the modern interpretation of the whole-genome Fundamental Theorem of Natural Selection. It is true whatever the mating scheme and the recombination structure might be. It is inconceivable that this result could have been obtained by anything other than a mathematical treatment.

Two comments are in order. First, fitness is considered here (and throughout this article) as a parameter, that is, as an intrinsic property of a genotype. It is in practice unknown, as is any parameter in statistical theory. Second, it is not clear that the FTNS makes a satisfactory statement — the concept of a partial change in mean fitness appears to some to be arbitrary.

The investigation of multi-locus properties in non-randomly-mating populations outlined above has hardly begun, and yet it will be essential to an investigation of the evolution of non-randomly-mating populations, in particular the human population, using whole-genome data.

MOLECULAR POPULATION GENETICS

Introduction

Multi-locus evolutionary theory arises by going outwards from the single gene locus to the whole genome. In considering molecular genetics we go in the other direction, and consider the internal nature of the gene. The investigation of evolutionary population genetics at the molecular level was initiated by Kimura [10], and it is without doubt the most important development of population genetics theory in the last fifty years. We therefore devote substantial attention to it.

Classical population genetics considered processes going forward in time, motivated by the need of rewriting and validating the Darwinian theory in terms of Mendelian genetics. In this task it was completely successful. Current research, by contrast, is largely focused on using current genetic data and assessing the evolutionary processes that led to them. The ensuing theory is thus retrospective rather than prospective, and it involves many statistical inference procedures. In carrying these procedures out it is essential to use the actual genetic material and not abstract quantities like “the allele

A_1" so prevalent in the classical theory. Thus the entire retrospective theory and all the inferential procedures must be carried out using the theory of molecular population genetics.

In molecular population genetics we view the gene for what it is, namely a sequence of the four nucleotides A, G, C and T. For a gene with 1000 nucleotides there are 4^{1000} different possible nucleotide sequences. This number is so large that little accuracy is lost in taking it to be infinite. Each one of these sequences corresponds (in our previous terminology) to some allele, so in using the word "allele" in the following section we mean some specific DNA sequence. We therefore assume an effective infinity of different possible alleles. This viewpoint motivates the theory in the following sections.

The infinitely many alleles model

The theory of molecular genetics is a stochastic one, and the most frequently used analysis assumes a diploid population of size N and a Wright-Fisher model of evolution. (Several other models have, of course, been considered in the literature.) In this model one assumes, since there is an effective infinity of possible alleles as just described, that when a gene mutates, for example by a change A $\rightarrow$ T at some place in its nucleotide sequence, it changes to a gene of an entirely new allele, not so far seen in the population. In the simplified version of the theory considered here, any gene of any allelic type is assumed to mutate at the same rate, denoted u. Further, all allelic types are assumed to be selectively equivalent. (A more advanced theory of course drops these assumptions.)

These various assumptions imply that, if in generation t there are X_i genes of allelic type A_i ($i=1,2,3,...$), then the probability that in generation $t+1$ there will be Y_i genes of allelic type A_i, together with Y_0 new mutant genes, all of different novel allelic types, is

$$\text{prob}\{Y_0, Y_1, Y_2, ...|X_1, X_2, ...\} = \frac{(2N)!}{\prod Y_i!}\prod \pi_i^{Y_i}, \qquad (49)$$

where $\pi_0 = u$ and $\pi_i = X_i(1-u)/(2N)$, $i=1,2,3,....$

This model differs fundamentally from previous mutation models in that since each allele will sooner or later be lost from the population, there can exist no nontrivial stationary distribution for the frequency of any allele. Nevertheless we are interested in stationary behavior, and it is thus important to consider what concepts of stationarity exist for this model. To do this we consider delabeled configurations of the form $\{a, b, c, ...\}$, where such a configuration implies that there exist a genes of one allele, b genes of another allele, and so on. The specific alleles involved are not of interest. The possible configurations can be written down as $\{2N\}$, $\{2N-1, 1\}$, $\{2N-2, 2\}$, $\{2N-2, 1, 1\}$, …, $\{1, 1, 1, ..., 1\}$ in dictionary order: the number of such configurations is $p(2N)$, the number of partitions of $2N$ into positive integers. The quantity $p(2N)$ has been extensively studied in the mathematical literature.

It is clear that (49) implies certain transition probabilities from one configuration to another. Although these probabilities are extremely complex and the Markov chain of configurations has an extremely large number of states, nevertheless standard theory shows that there exists a stationary distribution of configurations. We first consider one simple property of this stationary distribution, namely the probability that two genes drawn at random are of the same allelic type. For this to occur neither gene can be a mutant and, further, both must be descended from the same parent gene (probability $(2N)^{-1}$) or different parent genes which were of the same allelic type. Writing $F_2^{(t)}$ for the desired probability in generation t, we get

$$F_2^{(t+1)} = (1-u)^2((2N)^{-1} + \{1-(2N)^{-1}\}F_2^{(t)}). \qquad (50)$$

At equilibrium, $F_2^{(t+1)} = F_2^{(t)} = F_2$ and thus

$$F_2 = \{1-2N+2N(1-u)^{-2}\}^{-1}, \qquad (51)$$

where $\theta = 4Nu$. This is the simplest property of the stationary distribution of the configuration process.

Consider next the probability $F_3^{(t+1)}$ that three genes drawn at random in generation $t+1$ are of the same allele. These three genes will all be descendants of the same gene in generation t, (probability $(2N)^{-2}$), of two genes (probability $3(2N-1)((2N)^{-2})$) or of three different genes (probability $(2N-1)(2N-2)((2N)^{-2})$). Further, none of the genes can be a mutant, and it follows that

$$F_3^{(t+1)} = (1-u)^3(2N)^{-2}(1+3(2N-1)F_2^{(t)} + (2N-1)(2N-2)F_3^{(t)}). \qquad (52)$$

By equating $F_3^{(t)}$ and $F_3^{(t+1)}$ and using the value calculated above for F_2 we can find the stationary probability F_3 that three genes drawn at random are of the same allelic type. Clearly a continuation of this process is cumbersome and does not lead to any revealing results. We consider some approximate results below.

The above arguments do, however, lead to simple closed form for the partition process Markov chain eigenvalues. Equation (50) can be written in the form

$$F_2^{(t+1)} - F_2^{(\infty)} = (1-u)^2\{1-(2N)^{-1}\}\{F_2^{(t)} - F_2^{(\infty)}\}, \qquad (53)$$

and this implies that $(1-u)^2\{1-(2N)^{-1}\}$ is an eigenvalue of the Markov chain configuration process discussed above. A similar argument using (52) shows that a second eigenvalue is $(1-u)^3\{1-(2N)^{-1}\}\{1-2(2N)^{-1}\}$. Simi-

lar arguments can be used for all of the eigenvalues, and it is found that

$$\lambda_i = (1-u)^i\{1-(2N)^{-1}\}\{1-2(2N)^{-1}\} \cdots \{1-(i-1)(2N)^{-1}\} \quad (54)$$

is an eigenvalue of the configuration process matrix and that its multiplicity is $p(i)-p(i-1)$, where $p(i)$ is the partition number given above. This provides a complete listing of all the eigenvalues. For details see Ewens and Kirby [11].

An important aspect of the theory concerning the infinitely many alleles model is that it is used for inferential procedures. This implies that it is necessary to derive properties of a sample of genes taken from the population. The elements of this sampling theory are now outlined. We call the limiting process $N \to \infty$, $u \to 0$ with $\theta = 4Nu$ held constant, the "asymptotic limit". Then (51) shows that in this limit

$$F_2 = \frac{1}{1+\theta}. \quad (55)$$

Similarly, (52) shows that in this limit

$$F_3 = \frac{2}{(1+\theta)(2+\theta)}. \quad (56)$$

It is possible to continue in this way to find the limiting ($N \to \infty$) probability distribution of any allelic partition of a sample of n genes. This partition is best described by the vector $(a_1, a_2, ..., a_n)$, where a_j denotes the number of alleles in the sample that are represented by exactly j genes (so that $\sum ja_j = n$). The probability distribution of this vector is given by

$$P(a_1, a_2, ..., a_n; \theta) = \frac{n!}{S_n(\theta)} \prod_{j=1}^{n} \frac{\theta^{a_j}}{a_j! j^{a_j}}, \quad (57)$$

(Ewens [12]). Here $S_n(\theta) = \theta(\theta+1)(\theta+2)\cdots(\theta+n-1) = |S_n^1|\theta + |S_n^2|\theta^2 + \cdots + |S_n^n|\theta^n$, so that $|S_n^j|$ is the absolute value of a Stirling number of the first kind.

A particular case of this formula arises when $a_n = 1$, that is there is only one allele represented in the sample. The probability of this is

$$\frac{(n-1)!}{(1+\theta)(2+\theta)\cdots(n-1+\theta)}. \quad (58)$$

The total number of alleles observed in the sample is $\sum a_j$. Denoting this sum by K, it is found from (57) that the probability distribution of K is

$$P(K=k;\theta) = \frac{|S_n^k|\theta^k}{S_n(\theta)}, \ (k=1,2,...,n). \quad (59)$$

From this it follows that the mean of K is

$$E(K) = \sum_{j=1}^{n} \frac{\theta}{\theta+j-1}. \quad (60)$$

We return to these calculations below, when considering inference procedures in population genetics and also when considering the coalescent process.

The infinitely many sites model

The infinitely many alleles model refers to a situation (obtaining in the 1970's) where it was assumed that it is possible to assess whether two different genes were of the same or of different alleles. However that was all that was assumed, since the actual nucleotide sequence of any gene was not, at that time, known. A more refined model is the "infinitely many sites" model, (Kimura [13]), applicable in cases where the nucleotide sequence *is* known. The various assumptions made above for the infinitely many alleles model are retained in the infinitely many sites model, in particular that the gene mutation rate is u. It is assumed that there is no recombination within any gene, so that recombination does not create new alleles. The DNA sequence of the gene is assumed to be sufficiently long that any mutation occurs at a site at which no mutation has previously occurred. (As a less extreme and equivalent assumption so far as the mathematical theory is concerned, we assume any mutation occurs at a site at which only one nucleotide currently exists in the population.) Thus as with the infinitely many alleles model, all mutations are to new alleles, not so far seen in the population. Of course this assumption might not be a reasonable one in view of the properties of protein folding, and this has led to a "finitely many sites" theory not discussed here.

Most of the relevant theory was developed by Watterson [14]. Because exact expressions are very cumbersome, all of the results given here are approximate formulas deriving from the asymptotic limit $N \to \infty$. We define a "segregating site" in a sample (population) as a site where two different nucleotides exist in the sample (population). Watterson found that the mean of the number S of segregating sites in a sample of n genes is approximately

$$E(S) = g_1\theta, \quad (61)$$

where $g_1 = \sum_{j=1}^{n-1} j^{-1}$. We return to this equation later when considering inferential procedures. For $n > 10$ the variance of S is very close to

$$V(S) = g_1\theta + g_2\theta^2, \quad (62)$$

where $g_2 = \sum_{j=1}^{n-1} j^{-2}$. Watterson also found that the probability that there are no segregating sites in this sample is

$$P(S=0)=\frac{(n-1)!}{(1+\theta)(2+\theta)\cdots(n-1+\theta)}. \quad (63)$$

This is identical to the probability that $K = 1$ in the infinitely many alleles model as given in (58) above. This provides a confirming link between the infinitely many alleles model and the infinitely many sites model, since if there are no segregating sites there is only one DNA sequence, or allele, in the sample of n sequences. Unfortunately it is extremely difficult to find other links between the two models, and it is also very difficult to find the joint probability distribution of K and S.

An important case arises when $n = 2$. Watterson found that the probability that there s segregating sites in a sample of this size is

$$\frac{1}{\theta+1}\left(\frac{\theta}{\theta+1}\right)^s.$$

This distribution has mean θ and variance $\theta+\theta^2$. More generally it was shown by Tavaré [15] that in a sample of n sequences the probability distribution of S is

$$\text{prob}(S=s)=\frac{n-1}{\theta}\sum_{j=1}^{n-1}(-1)^{j-1}\binom{n-2}{j-1}\left(\frac{\theta}{j+\theta}\right)^{s+1}. \quad (64)$$

Inferential procedures: estimating θ and testing for selective neutrality

In this section we consider two inferential operations: estimating the parameter θ referred to above, and testing the hypothesis that the frequencies observed in our data are the result of purely random genetic drift and are not, for example, the result of natural selection. We consider these procedures for the two models outlined above, the infinitely many alleles model and the infinitely many sites model. In both models we use the simple approximate results such as (55), (57) deriving from the asymptotic limit theory, and Eq. (64) (given below) for the infinitely many sites model.

The infinitely many alleles model: inference methods

We start with Eqs. (57) and (59), both of which assume selective neutrality among the alleles observed. The conditional distribution of $(a_1, a_2, ..., a_n)$ given that $K = k$ is found from (57) and (59) to be

$$P(a_1,a_2,...,a_n\,|\,k)=\frac{n!}{|S_n^k|\prod_{j=1}^n a_j!j^{a_j}}. \quad (65)$$

The form of this expression shows that K is a sufficient statistic for θ: all the information in the sample relevant to θ is embodied in the observed value k of K. Thus any statistical inference concerning θ using the sample information must be carried out entirely by using the observed value k of K. This implies that θ should be estimated by using only the observed value k. In doing this we regard Eq. (59) as providing the likelihood of θ, given the observed value k of K. It is found that the maximum likelihood estimate $\hat{\theta}$ is the implicit solution of the equation of

$$k=\sum_{j=1}^{n}\frac{\hat{\theta}}{\hat{\theta}+j-1}. \quad (66)$$

Given the observed value of k, this equation has to be solved numerically for $\hat{\theta}$.

We now turn to tests of selective neutrality. These are based on Eq. (65), which is a "neutral theory" distribution and which does not contain any unknown parameters. It is therefore possible (at least in principle — in practice the computations involved are quite complicated) to consider any reasonable test statistic, evaluate its probability distribution from (65), and thus assess whether the observed value of this test statistic is a reasonable one, given this distribution. The most popular form of this test was devised by Watterson [16]: we do not provide any details here.

The infinitely many sites model: inference methods

Suppose that we have data consisting of n DNA sequences and that in these data there are s segregating sites. Equation (61) leads to two possible estimates of θ. The first (Ewens [17], Watterson [14]) follows directly from (61): given the observed value s of S, the estimate of $\hat{\theta}_S$ of θ is

$$\hat{\theta}_S=s/g_1. \quad (67)$$

This is clearly an unbiased estimate of θ. Assuming no recombination between the sites in the gene of interest, the variance of this estimate is

$$g_1\theta+g_2\theta^2/g_1^2. \quad (68)$$

A second estimate follows from the fact that in the case $n = 2$ the mean of S is θ. This implies that if all $\binom{n}{2}$ combinations of pairs of sequences in the data are taken, and the average T of the number of sites at which two sequences differ is taken, then the mean of T is given by

$$\text{Mean of } T=\theta. \quad (69)$$

Thus T could be used as an estimator of θ, and because of

this we write it as $\hat{\theta}_T$, and we re-write (69) as

$$\text{Mean of } \hat{\theta}_T = \theta. \tag{70}$$

Although this is an unbiased estimator of θ, it suffers because of the fact that its variance, which is

$$\frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2, \tag{71}$$

does not approach 0 as the sample size n increases. However our interest here in this estimator is that it forms part of the neutrality hypothesis testing procedure and not as a possible estimator of θ.

The fact that there are two unbiased estimators of θ under the hypothesis of selective neutrality forms the basis of a test first put forward by Tajima [18]. The Tajima procedure is carried out in terms of the statistic D, defined by

$$D = \frac{\hat{\theta}_T - \hat{\theta}_S}{\sqrt{\hat{V}}}, \tag{72}$$

where $\hat{V}$ is an unbiased estimate of the variance of $\hat{\theta}_T - \hat{\theta}_S$ and which we do not give here. The logic in choosing the test statistic is that if selective neutrality is the case (which in statistical terms is the null hypothesis) $\hat{\theta}_T$ and $\hat{\theta}_S$ both have the same mean (θ), so that D is similar to a standardized normal statistic.

The next problem is to find the null hypothesis distribution of D. Under the null hypothesis of neutrality, D does not have a mean of zero, nor does it have a variance of 1, since the denominator of D involves a variance estimate rather than a known variance. Further, the distribution of D depends on the value of θ, which is in practice unknown. Thus there is no null hypothesis distribution of D invariant over all θ values.

The Tajima procedure approximates the null hypothesis distribution of D in the following way. First, it is possible to find both the smallest value that D can take and also the largest value that D can take. These are functions of n and we denote them by a and b respectively. Second, it is assumed, as an approximation, that the mean of D is 0 and the variance of D is 1. Finally, it is also assumed that the density function of D is the generalized beta distribution over the range (a, b), defined by

$$f(D) = \frac{\Gamma(\alpha+\beta)(b-D)^{\alpha-1}(D-a)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)(b-a)^{\alpha+\beta-1}}, \tag{73}$$

with the parameters α and β chosen so that the mean of D is indeed 0 and the variance of D is indeed 1. This leads to the choice

$$\alpha = -\frac{(1+ab)b}{b-a},$$

$$\beta = \frac{(1+ab)a}{b-a}.$$

This approximate null hypothesis distribution is then used to assess whether any observed value of D is significant.

There are various approximations involved in the above procedure. Various authors have assessed the importance of these and have attempted to devise improved tests. We do not provide the details here.

THE COALESCENT

Coalescent theory

The concept of the coalescent is now discussed at length in many textbooks, and entire books (for example Hein, Schierup and Wiuf [19] and Wakeley [20]) and book chapters (for example Marjoram and Joyce [21] and Nordborg [22]) have been written about it.

The aim of the coalescent is to describe the common ancestry at various times in the past of the sample of n genes at some given gene locus through the concept of an equivalence class. (Essentially the same arguments apply to describe the common ancestry of all the $2N$ genes at a given locus in the population (of size N) at various times in the past.) To do this we introduce the notation τ, indicating a time τ in the past (so that if $\tau_1 > \tau_2$, time τ_1 is further in the past than time τ_2). The sample of n genes is assumed taken at time $\tau = 0$.

Two genes in the sample of n are in the same equivalence class at time τ if they have a common ancestor at this time. Equivalence classes are denoted by parentheses: thus if $n = 8$ and at time τ genes 1 and 2 have one common ancestor, genes 4 and 5 a second, and genes 6 and 7 a third, and none of the three common ancestors are identical and none is identical to the ancestor of gene 3 or of gene 8 at time τ, the equivalence classes at time τ are

$$\{(1,2),(3),(4,5),(6,7),(8)\}. \tag{74}$$

We call any such set of equivalence classes an equivalence relation, and denote any such equivalence relation by a Greek letter. As two particular cases, at time $\tau = 0$ the equivalence relation is $\phi_1 = \{(1),(2),(3),(4),(5),(6),(7),(8)\}$, and at the time of the most recent common ancestor of all eight genes, the equivalence relation is $\phi_n = \{(1,2,3,4,5,6,7,8)\}$. The Kingman coalescent process (Kingman [23,24]) is a description of the details of the ancestry of the n genes moving from ϕ_1 to ϕ_n. For example, given the equivalence relation in (74), one possibility for the equivalence relation following a coalescence is $\{(1,2),(3),(4,5),(6,7,8)\}$, so that equivalence classes $(6,7)$ and (8) have just amalgamated. Such an amalgamation is called a coalescence, and the process of successive amalgamations is called the coalescence

process.

Coalescences are assumed to take place according to a Poisson process, but with a rate depending on the number of equivalence classes present. Suppose that there are j equivalence classes at time τ. The probability that the process moves from one nominated equivalence class (at time τ) to some nominated equivalence class which can be derived from it is $\delta\tau$. (Here and throughout we ignore terms of order $(\delta\tau)^2$.) Since there are $j(j-1)/2$ possible choices for pairs of equivalence classes, a coalescence takes place in this time interval with probability $\frac{1}{2}j(j-1)\delta\tau$, since all of the $j(j-1)/2$ amalgamations possible at time τ are equally likely to occur. This implies that no coalescence takes places between time τ and time $\tau+\delta\tau$ with probability $1-\frac{1}{2}j(j-1)\delta\tau$.

In order for this process to describe the "random sampling" evolutionary model described above, it is necessary to scale time so that unit time corresponds to $2N$ generations. With this scaling, the time T_j between the formation of an equivalence relation with j equivalence classes to one with $j-1$ equivalence classes has an exponential distribution with mean $2/(j(j-1))$.

The (random) time $T_{\mathrm{MRCAS}}=T_n+T_{n-1}+T_{n-2}+\cdots+T_2$ until all genes in the sample first had just one common ancestor has mean

$$E(T_{\mathrm{MRCAS}})=2\sum_{j=2}^{n}\frac{1}{j(j-1)}=2\left(1-\frac{1}{n}\right). \qquad (75)$$

(The suffix "MRCAS" stands for "most recent common ancestor of the sample.) This is, of course close to 2 coalescent time units, or $4N$ generations, when n is large. Kingman (Kingman [23,24]) showed that for large populations, many population models (including the Wright-Fisher model considered in this paper) are well approximated in their sampling attributes by the coalescent process. The larger the population the more accurate is this approximation. Calculations such as this are relevant to the time back to "Eve" and to the "Out of Africa" migration.

These calculations provide a population equivalent of the sample expression (61). The mean time for which these are $j-1$ equivalence classes is $2/(j(j-1))$, as shown above. The mean number of mutations along the $j-1$ arms of the coalescent tree is then $\sum_{j=1}^{2N-1}2/(j(j-1))\times 2Nu(j-1)=\sum_{j=1}^{2N-1}\theta/j$. This is the population form of the sample expression (61).

We have just introduced mutation into the coalescent, and we now consider the effects of mutation in more detail. Suppose that the probability that any particular ancestral gene mutates in the time interval $(\tau+\delta\tau,\tau)$ is $\frac{\theta}{2}\delta\tau$. All mutants are assumed to be of new alleles (the infinitely many alleles assumption). If at time τ in the coalescent there are j equivalence classes, the probability that either a mutation or a coalescent event had occurred in $(\tau+\delta\tau,\tau)$ is

$$j\frac{\theta}{2}\delta\tau+\frac{j(j-1)}{2}\delta\tau=\frac{1}{2}j(j+\theta-1)\delta\tau. \qquad (76)$$

We call such an occurrence a defining event, and given that a defining event did occur, the probability that it was a mutation is $\theta/(j+\theta-1)$ and that it was a coalescence is $(j-1)/(j+\theta-1)$.

The probability that k different alleles are seen in the sample is then the probability that k of these defining events were mutations. The above reasoning shows that this probability must be proportional to $\theta^k/S_n(\theta)$, where $S_n(\theta)$ is defined below Eq. (57), the constant of proportionality being independent of θ. This argument leads directly to the expression (59) for the probability distribution of the number of alleles in the sample.

Using these results and combinatorial arguments counting all possible coalescent paths from a partition $(a_1,a_2,...,a_n)$ back to the original common ancestor, Kingman (Kingman [23,24]) was able to derive the more detailed sample partition probability distribution (57), and deriving this distribution from coalescent arguments is perhaps the most pleasing way of arriving at it.

"Age" and "time" results

Many further results are available from coalescent theory (or can be derived by special arguments). The ones listed in this section are "age" results and "time" results (such as (75)). They all relate to the Wright-Fisher infinitely many alleles model introduced above. Some are sample results and some are population results. In view of the nature of the coalescent in which we look backward into time, some of these results concern properties of the oldest allele in a sample (or in the population).

Kelly [25] showed that the probability that the oldest allele in the sample is represented j times in the sample is

$$\frac{\theta}{n}\binom{n}{j}\binom{n+\theta-1}{j}^{-1},\ (j=1,2,...,n). \qquad (77)$$

The case $j=n$ is of particular interest. If an allele is represented n times in a sample, it must be the oldest allele in the sample. Thus the expression (77) for the case $j=n$ should reduce to the expression (58). It is easy to see that this happens, so that (77) may be regarded as a generalization of (58). These calculations also show that the probability that a gene seen j times in the sample is of the oldest allele in the sample is j/n.

Perhaps the most important sample distribution concerns the frequencies of the alleles in the sample when ordered by age. This distribution was found by Donnelly

and Tavaré [26], who showed that the probability that the number of alleles in the sample takes the value k, and that the age-ordered numbers of these alleles in the sample are, in age order, $n_{(1)}, n_{(2)}, ..., n_{(k)}$, is

$$\frac{\theta^k (n-1)!}{S_n(\theta) n_{(k)} (n_{(k)} + n_{(k-1)}) \cdots (n_{(k)} + n_{(k-1)} + \cdots + n_{(2)})}, \tag{78}$$

where $S_j(\theta)$ is defined below (57). This formula can be found in several ways, one being as the size-biased version of (57).

These are many other interesting results concerning the oldest allele in a sample, and further results connecting the oldest allele in the sample to the oldest allele in the population, but the above examples give the general flavor of these.

We now turn to population results and their relation with sample data. Kelly [25] has shown that the probability that the oldest allele in the population is represented by j genes in the population is

$$\frac{\theta}{2N} \binom{2N}{j} \binom{2N+\theta-1}{j}^{-1}. \tag{79}$$

This is the population analogue of (77). There is also a sample result connected with this, namely that in the limiting case $N \to \infty$, the probability that the oldest allele in the population is observed in a sample of size n is $n/(n+\theta)$.

Donnelly [27] showed, more generally, that the probability that the oldest allele in the population is observed j times in the sample is

$$\frac{\theta}{n+\theta} \binom{n}{j} \binom{n+\theta-1}{j}^{-1}, \ (j = 0, 1, 2, ..., n). \tag{80}$$

For the case $j = 0$ the probability (80) is $\theta/(n+\theta)$, confirming the complementary probability $n/(n+\theta)$ found above. Conditional on the event that the oldest allele in the population does appear in the sample, the probability that it arises j times in the sample is given by the expression (77).

We now consider "age" questions. It is found that the mean time, into the past, that the oldest allele in the population entered the population (by a mutation event) is

$$\text{mean age of oldest allele} = \sum_{j=1}^{2N} \frac{4N}{j(j+\theta-1)} \text{generations}. \tag{81}$$

It was shown by Watterson and Guess [28] and Kelly [25] that not only the mean age of the oldest allele, but indeed the entire probability distribution of its age, is independent of its current frequency and indeed of the frequency of all alleles in the population.

If an allele is observed in the population with frequency p, its mean age is

$$\sum_{j=1}^{2N} \frac{4N}{j(j+\theta-1)} (1-(1-p)^j) \text{generations}. \tag{82}$$

This is a generalization of the expression in (81), since if $p = 1$ only one allele exists in the population, and it must then be the oldest allele.

A further calculation concerns the mean age of the oldest allele in a sample of n genes. This mean age is

$$4N \sum_{j=1}^{n} \frac{1}{j(j+\theta-1)} \text{generations}. \tag{83}$$

Except for small values of n, this is close to the mean age of the oldest allele in the population, given in (81). In other words, unless n is small, it is likely that the oldest allele in the population is represented in the sample. In fact we have seen above that the probability that the oldest allele in the population is represented in the sample is $n/(n+\theta)$.

We conclude by discussing two very important "population" probability distributions. For both distributions we consider the limiting $N \to \infty$ case and thus consider frequencies of than numbers of genes of some allele, and density functions rather than discrete probability distributions. All results considered relate to the infinitely many alleles model.

The first of these distributions is Kingman's [29] Poisson-Dirichlet distribution. This is the joint density function of the allele frequencies in a population when ordered by size (that is, the order statistics of the allele frequencies). Unfortunately this distribution is not "user-friendly" and few explicit results are known. Perhaps the most important is that of Watterson [30], which gives the joint density function of the first r order statistics $x_{(1)}$, $x_{(2)}, ..., x_{(r)}$ in the Poisson-Dirichlet distribution. This joint density function is

$$f(x_{(1)}, x_{(2)}, ..., x_{(r)}) = \theta^r \Gamma(\theta) e^{\gamma\theta} g(y) \{x_{(1)} x_{(2)} \cdots x_{(r)}\}^{-1} x_{(r)}^{\theta-1}, \tag{84}$$

where $y = (1 - x_{(1)} - x_{(2)} - \cdots - x_{(r)})/x_{(r)}$, γ is Euler's constant 0.57721…, and $g(y)$ is best defined through the Laplace transform equation (Watterson and Guess [28])

$$\int_0^\infty e^{-ty} g(y) \mathrm{d}y = \exp\left(\theta \int_0^1 u^{-1} (e^{-tu} - 1) \mathrm{d}u\right). \tag{85}$$

The expression (84) simplifies to

$$f(x_{(1)},...,x_{(r)}) = \theta^r\{x_{(1)}\cdots x_{(r)}\}^{-1}(1-x_{(1)}-\cdots-x_{(r)})^{\theta-1}, \quad (86)$$

when $x_{(1)}+x_{(2)}+\cdots+x_{(r-1)}+2x_{(r)}\geqslant 1$, and in particular,

$$f(x_{(1)}) = \theta(x_{(1)})^{-1}(1-x_{(1)})^{\theta-1}, \quad (87)$$

when $\frac{1}{2}\leqslant x_{(1)}\leqslant 1$.

Equation (87) provides two interesting results. First, population geneticists are interested in the probability of "population monomorphism", defined in practice as the probability that the most frequent allele arises in the population with frequency in excess of 0.99. Equation (87) implies that this probability is close to $1-(0.01)^{\theta}$. Second, if there is an allele with frequency between 1/2 and 1 it must be the most frequent, and thus the mean frequency of the most frequent allele must exceed

$$\int_{1/2}^{1} x_{(1)}\theta x_{(1)}(1-x_{(1)})^{\theta-1} = \left(\frac{1}{2}\right)^{\theta}.$$

This is close to 1 when θ is small. In other words, if the mutation rate and the population size are jointly small enough, we are likely to see one allele at high frequency in the population, together with a small number of low-frequency alleles.

Consideration of the most frequent allele in a population and the oldest allele in the population led to the following question (Crow [31]): "What is the probability that the most frequent allele in a population at any time is also the oldest allele in the population at that time?" A nice application of reversibility arguments for suitable population models allowed Watterson and Guess [28] to obtain a simple answer to this question. In models where all alleles are equally fit, the probability that any nominated allele will survive longest into the future is (by a simple symmetry argument) its current frequency. For time reversible processes, this is also the probability that it is the oldest allele in the population. Thus conditional on the current allelic frequencies, the probability that the most frequent allele is also the oldest is simply its frequency $x_{(1)}$. Thus the answer to Crow's question is simply the mean frequency of the most frequent allele. A formula for this mean frequencθy, as a function of the mutation parameter θ, together with some numerical values, were given in Watterson and Guess [28]. This leads to the numerical values in the first row of Table 1 below.

The fact that the Poisson-Dirichlet distribution is not user-friendly makes it all the more interesting that a *size-biased* distribution closely related to it, namely the GEM distribution, named for Griffiths [32], Engen [33] and McCloskey [34], who established its salient properties, is both simple and elegant. This distribution gives the joint density function of the frequencies of the alleles in the population when ordered by age (and not, as with the Poisson-Dirichlet distribution, when ordered by frequencies).

Suppose that a gene is taken at random from the population. The probability that this gene will be of an allele whose frequency in the population is x is just x. This allele was thus sampled by this choice in a size-biased way. It can be shown from properties of the Poisson-Dirichlet distribution that the probability density of the frequency of the allele determined by this randomly chosen gene is

$$f(x) = \theta(1-x)^{\theta-1}, 0<x<1. \quad (88)$$

The GEM distribution can then be defined as follows. Suppose that $x_1, x_2, x_3, ...$ are independently and identically distributed random variables, each having the probability distribution (88). Then define new random variables $w_1, w_2, w_3, ...$ by $w_1 = x_1$, and for $j = 2, 3, 4, ...$,

$$w_j = (1-x_1)(1-x_2)\cdots(1-x_{j-1})x_j. \quad (89)$$

The random vector $(w_1, w_2, ...)$ then has the GEM distribution. We then identify w_1 as the random variable describing the frequency of the oldest allele in the population, w_2 as the random variable describing the frequency of the second oldest allele in the population, and so on. The GEM distribution shows that the mean population frequency of the oldest allele in the population is

$$\theta\int_0^1 x(1-x)^{\theta-1}\mathrm{d}x = \frac{1}{1+\theta}. \quad (90)$$

This mean value of w_1 is given in the second line of Table 1 for various values of θ.

This table provides interesting comparisons between properties of the most frequent allele and the oldest allele in a population. If θ is small the means in the two rows of

Table 1. Mean frequency of (a) the most frequent allele, (b) the oldest allele, in a population as a function of θ. The probability that the most frequent allele is the oldest allele is also its mean frequency.

θ	0.1	0.2	0.5	1.0	2.0	5.0	10.0	20.0
Most frequent	0.936	0.882	0.758	0.624	0.476	0.297	0.195	0.122
Oldest	0.909	0.833	0.667	0.500	0.333	0.167	0.091	0.048

the table are close, and this corresponds to the fact that for small θ the most frequent allele in the population is very likely to be the oldest allele in the population.

Given the focus on retrospective questions arising in molecular population genetics, it is natural to ask further questions about the oldest allele in the population. The mean frequency $(1/(1+\theta))$ of the oldest allele in the population given above implies that when θ is very small, this mean frequency is approximately $1-\theta$. It is interesting to compare this with the mean frequency of the most frequent allele when θ is small, found in effect from the Poisson-Dirichlet distribution to be approximately $1-\theta\log 2$. More generally, the mean population frequency of the jth oldest allele in the population is

$$\frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^{j-1}.$$

The probability that a gene drawn at random from the population is of the type of the oldest allele is the mean frequency of the oldest allele, namely $1/(1+\theta)$, as just shown, since "size-biased" sampling is equivalent to "sampling by ages". More generally the probability that n genes drawn at random from the population are all of the type of the oldest allele in the population is

$$\theta\int_0^1 x^n(1-x)^{\theta-1}\mathrm{d}x=\frac{n!}{(1+\theta)(2+\theta)\cdots(n+\theta)}. \quad (91)$$

From this result and the expression (58) we confirm that if all genes in a sample of n are of the same allele, the probability that this is the oldest allele in the population is $n/(n+\theta)$.

The elegance of many of these formulas make molecular population genetics an attractive field for research. However, this elegance arises because of the simplifying assumptions that have been made, in particular that the gene locus of interest can be considered in isolation from linked gene loci, and that there is no selection. Modern theory considers realistic problems taking these complications into account, and (unfortunately) much of the elegance of the above theory is then lost. Tavaré [35], Durrett [36] and Etheridge [37] give accounts of this more recent theory.

REFERENCES

1. Darwin, C. (1859) On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life. London: John Murray.
2. Mendel, G. (1866) Versuche über pflanzenhybriden (Experiments relating to plant hybridization). Verh. Naturforsch. Ver. Brunn, 4, 3–17.
3. Hardy, G. H. (1908) Mendelian proportions in a mixed population. Science, 28, 49–50.
4. Weinberg, W. (1908) Über den Nachweis der Vererbung beim Menschen. (On the detection of heredity in man). Jahreshelfts. Ver. Vaterl. Naturf. Württemb., 64, 368–382.
5. Fisher, R. A. (1930) The Genetical Theory of Natural Selection. Oxford: Clarendon Press.
6. Malécot, G. (1948) Les Mathématiques de l'Hérédité. Paris: Masson.
7. Kingman, J. F. C. (1961) A mathematical problem in population genetics. Proc. Camb. Philol. Soc., 57, 574–582.
8. Wright, S. (1931) Evolution in Mendelian populations. Genetics, 16, 97–159.
9. Ewens, W. J. (2004) Mathematical Population Genetics. New York: Springer.
10. Kimura, M. (1971) Theoretical foundation of population genetics at the molecular level. Theor. Popul. Biol., 2, 174–208.
11. Ewens, W. J. and Kirby, K. (1975) The eigenvalues of the neutral alleles process. Theor. Popul. Biol., 7, 212–220.
12. Ewens, W. J. (1972) The sampling theory of selectively neutral alleles. Theor. Popul. Biol., 3, 87–112.
13. Kimura, M. (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics, 61, 893–903.
14. Watterson, G. A. (1975) On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol., 7, 256–276.
15. Tavaré, S. (1984) Lines of descent and genealogical processes, and their applications in population genetic models. Theoret. Pop. Biol., 26, 119–164.
16. Watterson, G. A. (1977) Heterosis or neutrality? Genetics, 85, 789–814.
17. Ewens, W. J. (1974) A note on the sampling theory for infinite alleles and infinite sites models. Theor. Popul. Biol., 6, 143–148.
18. Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 123, 585–595.
19. Hein, J., Schierup, M. H. and Wiuf, C. (2005) Gene Genealogies, Variation and Evolution. Oxford: Oxford University Press.
20. Wakeley, J. (2009) Coalescent Theory. Greenwood Village, Colorado: Roberts and Company.
21. Marjoram, P. and Joyce, P. (2009) Practical implications of coalescent theory. In Lenwood, L. S. and Ramakrishnan, N. (eds.), Problem Solving Handbook in Computational Biology and Bioinformatics. New York: Springer.
22. Nordborg, M. (2001) Coalescent theory. In Balding, D. J., Bishop, M. J. and Cannings, C. (eds.), Handbook of Statistical Genetics. Chichester, UK: Wiley.
23. Kingman, J. F. C. (1982) The coalescent. Stoch. Proc. Appl., 13, 235–248.
24. Kingman, J. F. C. (1982) On the genealogy of large populations. J. Appl. Probab., 19, 27–43.
25. Kelly, F. P. (1977) Exact results for the Moran neutral allele model. J. Appl. Probab., 14, 197–201.
26. Donnelly, P. J. and Tavaré, S. (1986) The ages of alleles and a coalescent. Adv. Appl. Probab., 18, 1–19.
27. Donnelly, P. J. (1986) Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles. Theor. Popul. Biol., 30, 271–288.
28. Watterson, G. A. and Guess, H. A. (1977) Is the most frequent allele the oldest? Theor. Popul. Biol., 11, 141–160.
29. Kingman, J. F. C. (1975) Random discrete distributions. J. R. Stat. Soc. [Ser. A], 37, 1–22.
30. Watterson, G. A. (1976) The stationary distribution of the infinitely-many neutral alleles model. J. Appl. Probab., 13, 639–651.
31. Crow, J. F. (1972) The dilemma of nearly neutral mutations: how important are they for evolution and human welfare? J. Hered., 63, 306–

316.
32. Griffiths, R. C. (1980) Unpublished notes.
33. Engen, S. (1975) A note on the geometric series as a species frequency model. Biometrika, 62, 697–699.
34. McCloskey, J. W. (1965) A model for the distribution of individuals by species in an environment. Unpublished PhD. thesis. Michigan State University.
35. Tavaré, S. (2004) Ancestral inference in population genetics. In Picard J. (ed.), École d'Été de Probabilités de Saint-Fleur XXX1-2001, 1-188, Berlin: Springer-Verlag.
36. Durrett, R. (2008) Probability Models for DNA Sequence Evolution. Berlin: Springer-Verlag.
37. Etheridge, A. (2011) Some Mathematical Models from Population Genetics. Berlin: Springer-Verlag.