

Chapter 2

Genetic Background

2.1 Basic Genetic Forces: Drift, Mutation, Recombination and Selection

Genome is constituted by all heritable or genetic material coded in the DNA. In humans, it is organized in 23 pairs of **chromosomes** in each cell's nucleus consisting of the total of twice 3×10^9 bases or symbols, as well as of thousands of copies of the relatively small circular mitochondrial genomes each consisting of about 16,600 symbols. Human nuclear genome is **diploid**, because the two sets of chromosomes are separately inherited from the two parents. The inheritance pattern follows Mendel's Laws, whereby the sex cells called gametes contain one of the two sets of parental chromosomes, and offspring are formed by a fusion of two parental gametes, sperm and egg. In some species such as bacteria only one set of chromosomes, indeed a single chromosome, exists. Such species are called **haploid**. Human mitochondrial genome is haploid and is inherited from individual's mother. A **gene** is a specific region of DNA that codes for a single protein. The position of a gene on a chromosome is known as its **locus**. More generally, a locus is frequently understood as a fixed point of reference in the genome. Variants of the DNA sequence at a locus are called **alleles**. Diploid individuals who have two identical alleles at a given locus are called **homozygotes**, whereas those who have two different alleles are called **heterozygotes**. If the total of k alleles exist in the population, there may exist at most k different homozygotes and $k(k - 1)/2$ different heterozygotes.

Random **genetic drift**, or simply **drift**, results from random undirected changes in allele frequency that occur by chance in all populations. Drift is caused by a chance loss of chromosomes, which fail to be transmitted to individuals of a descendant generation. This process comes about because populations are not infinitely large but rather are finite, or limited in size. The breeding individuals of any one generation produce a potentially infinite pool of gametes. For the basic case, it may be assumed that allele frequencies among gametes equal the allele frequencies among adults. However, because of the finite size of the population, the chromosomes of the descendant generation are sampled with replacement from the gamete pool, and as a result, some chromosomes may be passed more than once and some may be not passed at all to descendant generations. Let us notice that genetic drift in its pure form

is causing a reduction of the within-population variability, since no new variants are supplied while existing variants are continually lost. Finally, only descendants of a single ancestral chromosome remain in the population; this phenomenon is called **fixation**.

Mutation is a random event, which results in change of the allele at a given locus, to another allele, either preexisting in the population or new. Mutation is “attempting” to restore the genetic variation continually reduced by drift. We frequently assume that all individuals have the same genotype at a locus, which is called the **wildtype allele homozygote**, until a new variant is introduced to the population as a result of mutation. Mutants will then spread in the population according to the Mendel’s Laws. They usually become extinct, but some of them can reach higher allele frequency over time. There exist many different kinds of mutations, but all we are concerned with is that a mutation substitutes an allele with another allele. The probability that a mutation occurs at a locus is called the **mutation rate** at this locus. Mutation rate can differ from locus to locus. If we consider a single nucleotide as a locus, the mutation rate is below 10^{-8} per locus, per generation. Microsatellite markers have a higher mutation rate, around 10^{-4} – 10^{-3} per locus, per generation. The wildtype allele can be mutated to a **variant allele**, and a variant allele can be mutated back to the wildtype allele, which is called **back mutation**. However, this latter usually occurs at a much lower rate, so that it is frequently considered impossible.

Genetic **recombination**, also called **crossing over**, refers to a genetic event that occurs during the formation of sperm and egg cells. During the early stages of cell division in meiosis, two chromosomes of a homologous pair may exchange segments, producing genetic variations in germ cells. For example, if a chromosome has haplotype (vector of variants on these two loci) AB, and its homolog has haplotype ab, one of the gamete cells, because of recombination, may receive a chromosome with genotype Ab. Such gametes are called recombinant. The proportion of recombinants is called the **recombination rate** between these two loci, which reaches 1/2 if the two loci are located on two different chromosomes, and thus segregate independently. The **genetic distance** (also called map distance) between two loci is defined as the average number of crossovers between the loci per meiosis. The unit of genetic distance is the **centimorgan (cM)**. Two loci are 1 cm apart if on average there is one crossover occurring between these two loci on a single strand each 100 meioses. Because of uneven recombination rates across the chromosomes, the map distance does not necessarily reflect the true physical distance, which is measured in base pairs (bp). One important fact is that if two loci are close to each other, they tend to co-segregate during meiosis because of the low probability of crossing over.

Natural **selection** is a process that favors or induces survival and perpetuation of one kind of organism over other. Selection can be positive (or advantageous) or negative (or purifying) and has a profound impact on the evolution of the population.

Usually, the term “**fitness**” is used to describe the selective value of the phenotype. Fitness can be understood as the average number of progeny produced by the individual endowed with a particular phenotype. In the case when only 2 alleles exist at a locus, natural selection can be modeled by the relative fitnesses w_{AA} , w_{Aa} and w_{aa} , of genotypes AA , Aa and aa . We call a fitness (selection) model additive, if the fitnesses of genotypes AA , Aa and aa are equal to 1, $1 + s/2$ and $1 + s$, where $s \in (-1, 1)$, or recessive, if the fitnesses are equal to 1, 1 and $1 + s$. Other models include heterozygous advantage (overdominance or balancing selection), in which case the fitness of heterozygote Aa is higher than those of homozygotes AA and aa . (See e.g. [1] for a detailed discussion of these subjects.) Mean fitness in the population can be expressed by the formula

$$\bar{w} = w_{AA}p_A^2 + 2w_{Aa}p_Ap_a + w_{aa}p_a^2, \quad (2.1)$$

where $p_A + p_a = 1$. Consequently, at the descendant generation, the frequencies of alleles A and a are respectively equal to

$$\begin{aligned} p'_A &= p_A(w_{AA}p_A + w_{Aa}p_a)/\bar{w}, \\ p'_a &= p_a(w_{Aa}p_A + w_{aa}p_a)/\bar{w}. \end{aligned} \quad (2.2)$$

Under different hypotheses regarding fitnesses w_{AA} , w_{Aa} and w_{aa} , the expressions above lead to different patterns of evolution and different equilibria of allele frequencies. (We refer again to [1].) However, these considerations are valid only in absence of genetic drift, the condition satisfied for large population sizes N and for large copy numbers (Np_A and Np_a) of alleles. In situations such as when a single new mutant invades a wild-type population drift plays a major role and the above expressions do not apply.

This remark applies in particular to **Hardy–Weinberg Equilibrium**, which reflects binomial sampling of sex chromosomes in diploid organisms under assumption of no selective pressure [1]. According to Hardy–Weinberg law, in case of two alleles, the population frequencies of heterozygotes and homozygotes are equal to $p_{AA} = p_A^2$, $p_{aa} = p_a^2$, and $p_{Aa} = 2p_Ap_a$, respectively. Hardy–Weinberg Equilibrium can be affected by genetic drift but also by the so called meiotic drive (segregation distortion). For an example related to human genetic disease see [2], and for more recent reviews see [3, 4].

Interestingly, selection pressure does not have to be constant over the life time of an individual, or over the evolutionary history of the population. For example, some diseases such as **Alzheimer’s** only show decreased fitness at the later part of human lives. Because such diseases may not affect the fitness before the mating age, they may show no overall selection disadvantage. Another example is that a disease may be advantageous at first, but at a cost of deteriorated fitness later. Such a model is called **antagonistic pleiotropy**. The selection pressure on a disease allele may also change because of environmental and/or social changes.

Of the four major population genetic forces discussed above, the model the book is devoted to takes into account only two: mutations and drift. While incorporating recombination is possible (see [5, 6]), incorporating selection would lead to a much more complicated analysis and substantial non-linearity of the model. Hence, our findings apply merely to selectively neutral loci. To be sure, there are many loci of this type, including most examples of microsatellites (see Sect. 2.5).

2.2 Cannings, Wright–Fisher and Moran Models and the Coalescent

2.2.1 Cannings and Wright–Fisher Models

These are examples of discrete population models in discrete time, usually considered under the hypothesis of finite, constant (in time) population size. Our presentation follows the comprehensive review paper [7]. We set the constant population size to $2N$, where N is a positive integer. Population is treated as a “gametic urn” in which a diploid individual contributes two possibly different gametes.

2.2.1.1 Cannings Model

At each time-step, the $2N$ individuals are randomly labelled $i = 1, \dots, 2N$. Generation $n + 1$ consists of the offspring of individuals from generation n . For any i , individual i from generation n begets a number η_i of offspring, so that

$$\sum_i \eta_i = 2N.$$

The joint distribution of the $2N$ -tuple $(\eta_i)_{i=1,\dots,2N}$ is permutation invariant (exchangeable).

We observe a subpopulation $(Y_n; n \geq 0)$, which at time 0 is of given size: $Y_0 = y$, and Y_n denotes the number of descendants of this subpopulation at time n . As it transpires, $(Y_n; n \geq 0)$ is a discrete-time Markov chain, with two absorbing states 0 and $2N$. For any integer $0 \leq y \leq 2N$, we write $P_y[\cdot]$ to denote the conditional probability measure $P[\cdot | Y_0 = y]$. Let τ denote the absorption time

$$\tau = \inf\{n : Y_n = 0 \text{ (extinction) or } Y_n = 2N \text{ (fixation)}\}.$$

If we exclude the trivial case in which each individual has exactly 1 progeny, then exchangeability and constancy of the population size imply $\tau < \infty$ with probability 1. Indeed, $P_y[\text{fixation}] = y/(2N)$ and $P_y[\text{extinction}] = 1 - y/(2N)$.

2.2.1.2 Wright–Fisher Model

The Wright–Fisher (WF) model is a special case of the Cannings model, where the $2N$ -tuple $(\eta_i)_{i=1,\dots,2N}$ follows the multinomial distribution with parameters $(2N; 1/2N, \dots, 1/2N)$. As for the associated Markov chain Y , conditional on $Y_n = y$, $Y_{n+1} = \text{Bin}(2N, \frac{y}{2N})$, that is, Y_{n+1} follows the binomial distribution with number of trials $2N$ and success probability $\frac{y}{2N}$. Put otherwise, each individual from generation $n + 1$ picks its (one) parent at random, uniformly among the individuals of generation n , and these $2N$ samplings are independent.

2.2.2 Kingman–Tajima Coalescent

We discuss the **coalescent** introduced by J.F.C. Kingman in the papers [8–10], which offer a more mathematical approach. Tajima [11] introduced the same object using a more intuitive population genetics approach.

Individuals in the Wright–Fisher model are not independent: Tracing back their genealogical lines, we discover that some of them descend from of a single common ancestor, from whom they inherit most of their genetical make-up. This is to say that the structure of dependence is coded in random genealogical trees. It is described by a mathematical object, named Kingman’s n -coalescent, the main subject of this subsection.

To define it, first we discuss a related pure death process. We consider the Wright–Fisher population of size $M = 2N$, and observe n individuals sampled from generation 0. We are interested in the number $X_M(k)$; $k \geq 1$; of ancestors of this sample k generations back; we assume that the process is well-defined for all $k \geq 0$; i.e. that the population has evolved according to the Wright–Fisher rules for an indefinitely long time. $X_M(k)$; $k \geq 0$; is a discrete-time Markov chain with values in $\{1, \dots, n\}$ and transition probabilities

$$p_{i,j} = p_{i,j}(M) = M^{-i} \left\{ \begin{matrix} i \\ j \end{matrix} \right\} \binom{M}{j} j!,$$

where $\left\{ \begin{matrix} i \\ j \end{matrix} \right\}$ is the Stirling number of the second kind [12, 13]. Indeed, M^i is the number of all possible ways i members may choose their parents, and the number of ways exactly j parents may be chosen is the product of three numbers. The first of them is the number of ways the set of i elements may be partitioned into j subsets, i.e. the Stirling number of the second kind. The second is the number of ways j parents may be chosen from the population of M individuals—the binomial coefficient $\binom{M}{j}$, and the third is the number of possible assignments of j parents to subsets.

The process $X_M(k)$; $k \geq 0$; is a pure death process in that its paths are non-increasing sequences. As shown by Kingman, when $M \rightarrow \infty$,

$$X_M([tM]); \quad t \geq 0 \quad (2.3)$$

converges to a continuous-time (pure death) process with intensity matrix $Q = (q_{ij})$, where

$$q_{ii} = -\binom{i}{2} \quad i = 1, \dots, n, \quad q_{i,i-1} = \binom{i}{2}, \quad i = 2, \dots, n,$$

and $q_{ij} = 0$ otherwise. In other words

$$M[(p_{ij})_{1 \leq i, j \leq n} - I] \longrightarrow Q$$

componentwise.

A more comprehensive analysis allows tracing of the whole genealogy of a sample. To this end, for a sample of n individuals we consider the Markov chain $\mathcal{R}_M(k)$, $k \geq 0$, of equivalence relations in $\{1, \dots, n\}$; the pair (i, j) belongs to the equivalence relation $\mathcal{R}_M(k)$ iff the individuals i and j have a common ancestor k generations ago. Each equivalence class corresponds to a member of a population that lived k generations ago, yet the opposite statement is not true because some members of this generation may have not have descendants. $\mathcal{R}_M(0)$ is the main diagonal in the square $\{(i, j) | 1 \leq i, j \leq n\}$ and by the above analysis, $\mathcal{R}_M(k)$ eventually reaches the full equivalence relation, i.e. the whole square (see Fig. 2.1 and Table 2.1). The corresponding continuous-time Markov chain has intensity matrix Q given by

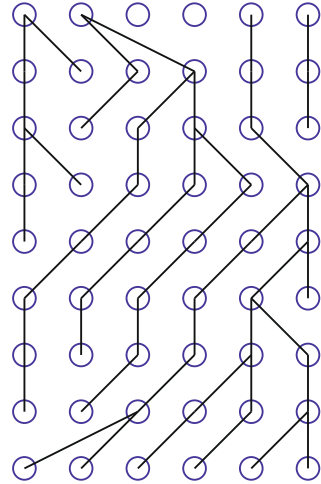
$$q_{\mathcal{E}, \mathcal{E}'} = \begin{cases} -\binom{|\mathcal{E}|}{2}, & \text{if } \mathcal{E} = \mathcal{E}', \\ 1, & \text{if } \mathcal{E} \prec \mathcal{E}', \\ 0, & \text{otherwise,} \end{cases}$$

where $|\mathcal{E}|$ denotes the number of equivalence classes in an equivalence relation $\mathcal{E}$ and we write $\mathcal{E} \prec \mathcal{E}'$ if $\mathcal{E} \subset \mathcal{E}'$ and $\mathcal{E}'$ is formed by amalgamating (exactly) two equivalence classes of $\mathcal{E}$. The Markov chain with the above intensity matrix is called the n -coalescent of Kingman.

There are two fundamental features of the coalescent. First of all, here merely two genealogical lines may merge at a time: no multiple merges are possible. This is in contrast to the approximating genealogies in the Wright–Fisher model, where many individuals may have a single ancestor in the preceding generation. In fact, Kingman’s combinatorial analysis leading to coalescent proves that under suitable assumptions such multiple merges may be disregarded. The second fact to be noted is that the time (to repeat: counted backwards) to the most recent ancestor of two individuals is exponential with parameter 1. Taking into account the scaling used in (2.3), we see that in a large Wright–Fisher population, the time T to the most recent common ancestor of two individuals is approximately exponential with parameter $\frac{1}{2N}$

Table 2.1 Equivalence relations in Fig. 2.1

Equivalence relation	Equivalence classes
$\mathcal{R}_6(0)$	$\{1\}\{2\}\{3\}\{4\}\{5\}\{6\}$
$\mathcal{R}_6(1)$	$\{1, 2\}\{3\}\{4\}\{5, 6\}$
$\mathcal{R}_6(2)$	$\{1, 2\}\{3, 4\}\{5, 6\}$
$\mathcal{R}_6(3)$ and $\mathcal{R}_6(4)$	$\{1, 2\}\{3, 4, 5, 6\}$
$\mathcal{R}_6(5)$ and consecutive ones	$\{1, 2, 3, 4, 5, 6\}$

Fig. 2.1 (Untangled) genealogies in the Wright–Fisher model

$$P(T > t) = e^{-\frac{t}{2N}}. \quad (2.4)$$

In particular, the smaller the population the shorter is τ .

The latter statement may be deduced also directly. Denoting by T_{2N} the time to the most recent common ancestor of two individuals in a Wright–Fisher model, we see that

$$P(T_{2N} > i) = \left(\frac{2N-1}{2N}\right)^i,$$

i.e. that τ_{2N} has a geometric distribution (shifted by one). Put otherwise, if finding a common ancestor of two individuals in a preceding generation is a success, then the event $\{T > i\}$ is that of i failures in i consecutive trials. Now, for each $t > 0$,

$$P\left(\frac{T_{2N}}{2N} > t\right) = P(T_{2N} > 2Nt) = P(T_{2N} > [2Nt]) = \left(1 - \frac{1}{2N}\right)^{[2Nt]}$$

where $[\cdot]$ denotes the integer part. It follows that

$$\lim_{2N \rightarrow \infty} P\left(\frac{T_{2N}}{2N} > t\right) = e^{-t},$$

as desired.

2.3 The Master Equation and the Moran Model

In this section, we introduce the **Moran model** of population genetics, which provides one way in which our master equation can be defined. This model is frequently used by theoreticians instead of the Wright–Fisher model, since it yields exact or at least asymptotic results both in the neutral case and in the case of selection. We are loosely following the approach in Durrett’s book on models of evolution of DNA sequences [14], see also [15].

Moran model is usually defined in the time-discrete and time-continuous version. The verbal definitions in both cases are almost the same:

- Constant population of N individuals
- Periodically, a randomly chosen individual dies and at the same moment, another randomly chosen individual proliferates (can be the same individual) see Fig. 2.2
- In the model with directional selection, there are individuals of two types: wildtype (W) and mutant (M) and the choice of individual that proliferates is biased. The odds that a wildtype proliferates are $(1 - s)(N - i)/i$, $s \in (0, 1)$.

The difference is in the exact formulation.

2.3.1 Discrete Case

Let us denote the number of mutants by i . There are four possibilities

- W (a wild type individual) dies; this happens with probability $\frac{N-i}{N}$
 - W proliferates; this happens with probability $\frac{(1-s)(N-i)}{(1-s)(N-i)+i}$
 - M (a mutant individual) proliferates; this happens with probability $\frac{i}{(1-s)(N-i)+i}$

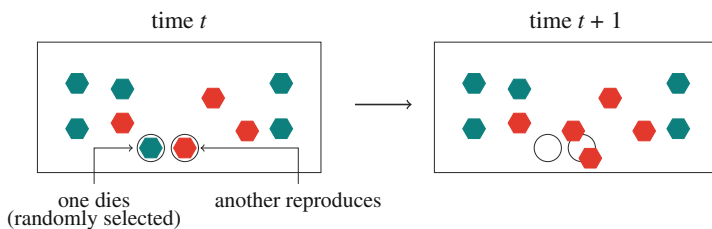


Fig. 2.2 Moran model with discrete time and directional selection

- M dies; this happens with probability $\frac{i}{N}$
 - W proliferates, with probability $\frac{(1-s)(N-i)}{(1-s)(N-i)+i}$
 - M proliferates, with probability $\frac{i}{(1-s)(N-i)+i}$

Only the WM and MW options lead to change in number of mutants

$$p_{i,i+1} = \frac{N-i}{N} \frac{i}{[(1-s)(N-i)+i]}, \quad p_{i,i-1} = \frac{i}{N} \frac{(1-s)(N-i)}{[(1-s)(N-i)+i]},$$

the MM and WW options jointly contribute to $p_{i,i}$. States $\{0\}$ and $\{N\}$ are absorbing. The probability of being eventually absorbed in $\{N\}$, if at time 0 there are i mutants, is equal to

$$P[T_N < T_0] = \frac{1 - (1-s)^i}{1 - (1-s)^N}$$

in the case with selection which leads to (take $s \rightarrow 0+$)

$$P[T_N < T_0] = i/N$$

in the neutral case. Here, T_0 and T_N are times of absorption at $\{0\}$ and $\{N\}$, respectively.

2.3.2 Continuous Case

Usually, it is defined by transition intensities

$$q_{i,i+1} = (N-i) \frac{i}{N}, \quad q_{i,i-1} = i \frac{(1-s)(N-i)}{N},$$

which have different denominators than the transition probabilities in the discrete version. However, despite this, the probability of fixation of the mutant is identical in the continuous and discrete case. The expected time to absorption in $\{N\}$ (fixation of the mutant) has asymptotics

$$E_1(T_N | T_N < T_0) \sim \frac{2}{s} \ln(N)$$

as $N \rightarrow \infty$, in the case with selection, and

$$E_i(T_N | T_N < T_0) \sim \frac{-N[1 - (i/N)]}{(i/N)} \ln[1 - (i/N)]$$

in the neutral case.

2.3.3 Connection with the Master Equation

In the neutral case of the continuous-time Moran model, the time to the most recent common ancestor of a pair of individuals is exponentially distributed with the parameter $1/N$ (or $1/(2N)$, considering N diploid individuals), if the expected lifetime of the individual is exponential with parameter 1. The demonstration follows directly from the definition of the Moran model and lack of memory of the exponential distribution. Our master equation can be understood as being derived directly from the time-continuous Moran model in the neutral case. Modification involving selection are also possible, although they may become complicated (see [16, 17]).

The advantage over the Wright–Fisher model is that in the time-continuous Moran model the exponential distribution is exact even for small N . In the Wright–Fisher model, which is defined in the terms of discrete non-overlapping generations, exponential distribution is obtained as a continuous approximation. Moreover, as already discussed, in the Wright–Fisher model, more than 2 lines of descent may merge in a past generation which leads to substantial difficulties in analysis of the full model (see e.g. [18], see also [19, 20]), while here such multiple merges are impossible.

2.4 Derivation of the Master Equation

As we have seen, the time-continuous Moran model assumes the population is composed of a constant number of $2N$ haploid individuals. Each individual undergoes death/birth events according to a Poisson process with intensity 1 (mean length of life of each individual is equal to 1). Upon a death/birth event, a genotype for the individual is sampled with replacement from the $2N$ chromosomes present at this moment, including the chromosome of the just-deceased individual. The following is the equivalent coalescent formulation of the Fisher–Wright–Moran model for a population of $2N$ haploid individuals under genetic drift and mutations following a general time-continuous Markov chain:

- *Coalescent with independent branch lengths with exponential distribution with parameter $1/(2N)$.* The interpretation is that for any two individuals from the population, the time to their common ancestor is a random variable T with exponential distribution, see (2.4) ([21, 22]).
- *Markov model of mutations with transition probabilities $P_{ij}(t)$ and intensities Q_{ij}* (see also Sects. 2.5 and 4.3). The interpretation is that if the allele state of an individual is i at time 0, then his/her allele state at time t (or the allele state of his/her descendant at time t) is equal to j with probability $P_{ij}(t)$. In the finite-dimensional case, the transition matrix $P(t) = \exp(Qt)$, where Q is the **intensity matrix** satisfying the following conditions: (a) $Q_{ij} \geq 0$, $i \neq j$, (b) $\sum_j Q_{ij} = 0$, all i .

We will use the coalescent model of genetic drift, modified to allow for the varying population size, i.e. $N = N(t)$, which will be represented by time-dependent hazard rate of the time to coalescence:

- The time T to the common ancestor of two individuals from the sample taken at time t is a random variable with hazard rate $[2N(t - \tau)]^{-1}$, i.e. $\Pr[T > \tau] = \exp[-\int_0^\tau [2N(t - u)]^{-1} du]$ (comp. (2.4)).
- The model of mutation is the same as above.

Let $R_{jk}(t) = \Pr[X_1 = j, X_2 = k]$, where X_1 and X_2 are randomly selected chromosomes. If the common ancestor of X_1 and X_2 was of allele type i and it existed τ units of time ago, then $R_{jk}(t) = P_{ij}(\tau)P_{ik}(\tau)$. The allele type of the common ancestor is the state of the Markov chain associated with the mutation process and so it is equal to i with probability $\pi(t) = \Pr[X_1(t) = i]$ defined by this process. Taking this into account, we obtain,

$$R_{jk}(t) = \int_0^\infty \sum_i \pi_i(t - \tau) P_{ij}(\tau) P_{ik}(\tau) \left(\frac{1}{2N(t - \tau)} e^{-\int_0^\tau \frac{du}{2N(t-u)}} \right) d\tau. \quad (2.5)$$

In matrix notation, following a change of variables $\sigma = t - \tau$,

$$R(t) = \int_{-\infty}^t P^T(t - \sigma) \Pi(\sigma) P(t - \sigma) \left(\frac{1}{2N(\sigma)} e^{-\int_\sigma^t \frac{du}{2N(u)}} \right) d\sigma, \quad (2.6)$$

where $\Pi(t) = \text{diag}[\pi_i(t)]$. Let us notice that $\sum_{jk} R_{jk}(t) = 1 - \exp\{-\int_0^\infty [2N(t - u)]^{-1} du\}$, so the distribution $R(t)$ may be improper if $\int_0^\infty [2N(t - u)]^{-1} du < \infty$. This would mean that X_1 and X_2 do not have a common ancestor.

Also, the above formulation requires that the Markov chain be extendable indefinitely into the past, i.e. that $\Pi(\sigma)$ exist for all $\sigma \leq t$. Not getting into conditions that might ensure this, let us carry out a formal transformation of (2.6), by splitting the integration interval into two parts

$$\begin{aligned} R(t) &= \left(\int_{-\infty}^0 + \int_0^t \right) P^T(t - \sigma) \Pi(\sigma) P(t - \sigma) \left(\frac{1}{2N(\sigma)} e^{-\int_\sigma^t \frac{du}{2N(u)}} \right) d\sigma \\ &= P^T(t) \left[\int_{-\infty}^0 P^T(-\sigma) \Pi(\sigma) P(-\sigma) \left(\frac{1}{2N(\sigma)} e^{-\int_\sigma^0 \frac{du}{2N(u)}} \right) d\sigma \right] P(t) e^{-\int_0^t \frac{du}{2N(u)}} \\ &\quad + \int_0^t P^T(t - \sigma) \Pi(\sigma) P(t - \sigma) \left(\frac{1}{2N(\sigma)} e^{-\int_\sigma^t \frac{du}{2N(u)}} \right) d\sigma \\ &= P^T(t) R(0) P(t) e^{-\int_0^t \frac{du}{2N(u)}} \\ &\quad + \int_0^t P^T(t - \sigma) \Pi(\sigma) P(t - \sigma) \left(\frac{1}{2N(\sigma)} e^{-\int_\sigma^t \frac{du}{2N(u)}} \right) d\sigma. \end{aligned} \quad (2.7)$$

The latter expression could be derived independently by assuming that if the coalescent time is longer than t , the two individuals do not coalesce, but that their allele statuses have joint distribution $R(0)$ and marginal distributions $\pi(0)$. Let us note that if $R(0)$ is proper, then $R(t)$ is proper.

It can be demonstrated using differentiation of the above expression with respect to t that $R(t)$ given by (2.7) satisfies the following matrix differential equation,

$$\dot{R}(t) = [Q^T R(t) + R(t) Q] - \frac{1}{2N} R(t) + \frac{1}{2N} \Pi(t) \quad (2.8)$$

with a given initial condition $R(0)$. This latter statement is exactly correct in the finite-dimensional case. In the infinite-dimensional case, a more thorough analysis is needed (see Chaps. 4 and 5).

Equation (2.8), which is our Master Equation, is a modification of a matrix differential equation known as the Lyapunov equation [23]. It was first derived by O'Brien [24, 25], then in a specific setup by Kimmel in [26] and then re-derived and published with comprehensive qualitative analysis in [27]. In Chaps. 4 and 5 will explain its form and provide insight into asymptotic behavior of its solutions. As we shall see, in particular, the second term on the right-hand side reflects genetic drift, while the first term speaks of independent processes of mutations on two individuals after the time of split of genealogical lines.

2.5 Examples of Markov Mutations

Before completing this chapter, we need to provide a population genetics intuitions leading to the notion of a **Markov mutation**. Mathematical point of view will be given in Sect. 4.3.

2.5.1 Microsatellite DNA and the Stepwise Mutation Model

Microsatellite repeat loci are stretches of repeated DNA motifs of length of 2–6 nucleotides. An example is a triplet repeat (motif of length 3) with allele size $X = 4$ (motif repeated 4 times)

$$\dots |ACG|ACG|ACG|ACG| \dots$$

Mutations in such loci usually have the form of expansions or contractions occurring at a high rate, $\nu \sim 10^{-3}$ – 10^{-4} per generation. More specifically,

$$X \longrightarrow X + U \quad (2.9)$$

where U is an integer-valued random variable, at time epochs of a Poisson process with intensity ν . This **Stepwise Mutation Model** (SMM), mathematically is an **unrestricted random walk** (see e.g. [28]).

Microsatellites are highly abundant in the genome (GDB database lists ca 5,000 dinucleotides). They are also highly polymorphic (variable). Applications of microsatellites include: forensics (identification), mapping (locating genes), and evolutionary studies.

A microsatellite locus can be considered to have a denumerable set of alleles indexed by integers. Two statistics can summarize the variability at a microsatellite locus in a sample of n chromosomes: The estimator of the genetic variance

$$\hat{V}/2 = \sum_{i=1}^n (X_i - \bar{X}) / (n-1), \quad (2.10)$$

where $X_i = X_i(t)$ is the size of the allele in the i th chromosome present and $\bar{X}$ is the mean of the X_i

$$V(t) = E(\hat{V}) = E[(X_i - X_j)^2], \quad (2.11)$$

and X_i and X_j are the sizes of two alleles from the population [29]; and the estimator of homozygosity

$$\hat{P}_0 = \left(n \sum_{k=1}^K p_k^2 - 1 \right) / (n-1), \quad (2.12)$$

where p_k denotes the relative frequency of allele k in the sample

$$P_0(t) = E(\hat{P}_0) = \Pr[X_i(t) = X_j(t)]. \quad (2.13)$$

Random variables X_i considered here are exchangeable but not independent.

2.5.2 Mitochondrial DNA and the Infinitely Many Sites Model

Many loci have the form of long sequences of DNA nucleotides, e.g. $\cdots ACGTG \cdots$, with any single residue mutating independently and very infrequently (10^{-5} – 10^{-9} per generation) by base substitution, e.g. $A \rightarrow G$. Since it is highly unlikely that a mutation “hit” occurs more than once at some residue, it can be effectively assumed that the locus has an infinite number of sites and that mutations occurring at times defined by a Poisson process “select” a new site each time. This is the Infinitely Many Sites Models (IMSM) [15]. It is considered that sequences of the hypervariable (HV) region of the human mitochondrial genome conform to the IMSM [30].

Let us consider two chromosomes ($n = 2$) and an IMSM locus. We compute the theoretical distribution of the number of mismatches between two sequences (loci). Assume mutation rate ν per generation per locus:

“Infinitely” long DNA sequences

$\Rightarrow$ Each mutation occurs at a new site in sequence

$\Rightarrow X = \#\{\text{mismatches between 2 sequences}\}$
 $= \#\{\text{mutations in both branches of coalescent}\}$
 $\sim \text{Poisson}(2 \cdot \nu \cdot T)$, conditional on T

$\Rightarrow$ Probability generating function (probability generating function) of X , $\alpha(s) = E(s^X)$ satisfies (see [31])

$$\alpha(s) = \int_0^\infty e^{2\nu\tau(s-1)} p(\tau) d\tau \quad (2.14)$$

where

$$p(\tau) = \frac{\exp\left[-\int_0^\tau \frac{du}{2N(u)}\right]}{2N(\tau)}, \quad \tau \geq 0, \quad (\text{timebackwards}). \quad (2.15)$$

is the distribution density of T .

$\alpha(s)$ is the probability generating function of the number of segregating sites. Introducing the mutational time $t = 2\nu\tau$, and coalescence intensity function in the mutational time scale, $\pi(t) = \frac{1}{2\nu} p\left(\frac{t}{2\nu}\right)$, we obtain

$$\alpha(s) = \int_0^\infty e^{t(s-1)} \pi(t) dt. \quad (2.16)$$

With $z = -(s - 1)$ the probability generating function $\alpha(1 - z)$ is the Laplace transform $\alpha(1 - z) = \tilde{\pi}(z)$ of the coalescence intensity function $\pi(t)$.

Infinitely many sites model cannot be conveniently modeled using the Master Equation in the form discussed in the current monograph. However, an extension based on stochastic point processes has been developed in [32].

References

1. Graur D, Li W.-H (2000) Fundamentals of molecular evolution, 2nd edn. Sinauer Associates, Sunderland
2. Polański A, Chakraborty R, Kimmel M, Deka R (1998) Dynamic balance of segregation distortion and selection maintains normal allele sizes at the myotonic dystrophy locus. *Math Biosci* 147(1):93–112
3. Labbe A, Huang LO, Infante-Rivard C (2013) Transmission ratio distortion: a neglected phenomenon with many consequences in genetic analysis and population genetics. In: Naumova AK, Greenwood CMT (eds) *Epigenetics and complex traits*. Springer, New York, pp 265–285 (English)

4. Zöllner S, Xiaquan W, Hanchard NA, Herbert MA, Ober C, Pritchard JK (2004) Evidence for extensive transmission distortion in the human genome. *Am J Hum Genetics* 74(1):62–72
5. Bobrowski A, Kimmel M (2003) A random evolution related to a Fisher-Wright-Moran model with mutation, recombination and drift. *Math Methods Appl Sci* 2003(26):1587–1599
6. Bobrowski A, Kimmel M, Wojdyła T (2010) Asymptotic behavior of a Moran model with mutations, drift and recombinations among multiple loci. *J Math Biol* 61:455–473
7. Lambert A (2008) Population dynamics and random genealogies. *Stoch Models* 24(suppl. 1):45–163
8. Kingman JFC (1982) The coalescent. *Stoch Proc Appl* 13:235–248
9. Kingman JFC (1982) Exchangeability and the evolution of large populations. In: Koch G, Spizzichino F (eds) *Exchangeability in probability and statistics*. North-Holland, Amsterdam, pp 97–112
10. Kingman JFC (1982) On the genealogy of large populations. *J Appl Prob* 19A:27–43
11. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460
12. Graham RL, Knuth DE, Patashnik O (1994) *Concrete mathematics*, 2nd edn. Addison-Wesley, Reading
13. Riordan J (1958) *An introduction to combinatorial analysis*. Wiley Publications in Mathematical Statistics. Wiley, Chapman & Hall Ltd, New York, London
14. Durrett R (2008) *Probability models for DNA sequence evolution*. Springer, New York (See also the 1st edn, published in 2002)
15. Ewens WJ (2004) *Mathematical population genetics*, 2nd edn. Springer, New York (See also the 1st edn, published in 1979)
16. Krone SM, Neuhauser C (1997) Ancestral processes with selection. *Theor Popul Biol* 51:210–237
17. Peng B, Kimmel M, Amos CI (2012) *Forward-time population genetics simulations*. Wiley, Hoboken
18. Yun-Xin Fu (2006) Exact coalescent for the wrightfisher model. *Theor Popul Biol* 69:385–394
19. Bobrowski A, Kimmel M (2004) Asymptotic behavior of joint distributions of characteristics of a pair of randomly chosen individuals in discrete-time fisher-wright models with mutations and drift. *Theor Popul Biol* 66(4):355–367
20. Bobrowski A, Kimmel M, Kubalińska M (2010) Non-homogeneous infinitely many sites discrete-time model with exact coalescent. *Math Methods Appl Sci* 33(6):713–732
21. Griffiths RC, Tavaré S (1994) Sampling theory for neutral alleles in a varying environment. *Philoso Trans R Soc Lond* 344:403–410
22. Tavaré S (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor Popul Biol* 26:119–164
23. Gajic Z, Qureshi MTJ (1995) *Lyapunov matrix equation in system stability and control, mathematics in science and engineering*, vol 195. Academic Press Inc, San Diego
24. O'Brien P (1982) Allele frequencies in a multidimensional Wright-Fisher model with general mutation. *J Math Biol* 15(2):227–237
25. O'Brien P (1985) Homozygosity in a population of variable size and mutation rate. *J Math Biol* 22(3):279–291
26. Kimmel M, Chakraborty R, King JP, Bamshad M, Watkins WS, Jorde LB (1998) Signatures of population expansion in microsatellite repeat data. *Genetics* 148:1921–1930
27. Bobrowski A, Kimmel M, Chakraborty R, Arino O (2001) A semigroup representation and asymptotic behavior of the Fisher-Wright-Moran coalescent. In: Rao CR, Shanbhag DN (eds) *Handbook of statistics 19: stochastic processes: theory and methods*, Chapter 8. Elsevier Science, Amsterdam
28. Chung KL (1960) *Markov chains with stationary transition probabilities*. Springer, Berlin
29. Kimmel M, Chakraborty R, Stivers DN, Deka R (1996) Dynamics of repeat polymorphisms under a forward-backward mutation model: within-and between-population variability at microsatellite loci. *Genetics* 143:549–555

30. Cyran KA, Kimmel M (2010) Alternatives to the Wright-Fisher model: the robustness of mitochondrial Eve dating. *Theor Popul Biol* 78(3):165–172
31. Polański A, Chakraborty R, Kimmel M (1998) Application of a time-dependent coalescence process for inferring the history of population size changes from dna sequence data. *Proc Nat Acad Sci* 95(10):5456–5461
32. Bobrowski A, Wang N, Chakraborty R, Kimmel M (2002) Non-homogeneous infinite sites model under demographic change: mathematical description and asymptotic behavior of pairwise distributions. *Math Biosci* 175(2):83–115

An Operator Semigroup in Mathematical Genetics

Bobrowski, A.; Kimmel, M.

2015, VI, 88 p. 9 illus., 8 illus. in color., Softcover

ISBN: 978-3-642-35957-6