J. theor. Biol. (2000) **204,** 629–638
doi:10.1006/jtbi.2000.2032, available online at http://www.idealibrary.com on IDEAL®

Ancestral Processes in Population Genetics—the Coalescent

M. MÖHLE*

Johannes Gutenberg-Universität Mainz, Fachbereich Mathematik, Saarstraße 21, 55099 *Mainz, Germany*

(*Received on* 25 *January* 1999, *Accepted in revised form on* 1 *March* 2000)

A special stochastic process, called the coalescent, is of fundamental interest in population genetics. For a large class of population models this process is the appropriate tool to analyse the ancestral structure of a sample of n individuals or genes, if the total number of individuals in the population is sufficiently large. A corresponding convergence theorem was first proved by Kingman in 1982 for the Wright–Fisher model and the Moran model. Generalizations to a large class of exchangeable population models and to models with overlying mutation processes followed shortly later. One speaks of the "robustness" of the coalescent, as this process appears in many models as the total population size tends to infinity. This publication can be considered as an introduction to the theory of the coalescent as well as a review of the most important "convergence-to-the-coalescent-theorems". Convergence theorems are not only presented for the classical exchangeable haploid case but also for larger classes of population models, for example for diploid, two-sex or non-exchangeable models. A review-like summary of further examples and applications of convergence to the coalescent is given including the most important biological forces like mutation, recombination and selection. The general coalescent process allows for simultaneous multiple mergers of ancestral lines.

1. Introduction

Since Fisher (1958) and Wright (1969) or even since Darwin and Mendel there is no doubt that modelling population systems in biology or particle systems in physics is of basic interest in modern science. Many attempts have been made to model such systems, most of them based on either deterministic approaches (difference and differential equations, dynamical systems), on stochastic approaches (stochastic processes) or mixings of both (stochastic dynamical systems Arnold, 1994). Time-discrete and time-continuous models are distinguished. In most cases, it depends on the specific problem or on the structure of the population whether a time-discrete or a time-continuous model seems to be more appropriate to describe and to analyse the population. In some cases, time-discrete processes can be approximated using time-continuous processes. The time-continuous limit processes are usually constructed via diffusion approximations (Ethier & Kurtz, 1986) which are in principle based on first- and second-order Taylor-like expansions of certain moments or transition functionals. Population genetics applications of such approximations can be found in Ethier & Nagylaki (1980), Nagylaki (1990) and references therein. The n-coalescent turns out to be a time-continuous approximation for the ancestral structure of a certain class of time-discrete population models. In stochastics the evolution

* E-mail: moehle@mathematik.uni-mainz.de

0022–5193/00/120629 + 10 $35.00/0

of time-discrete populations is usually modelled via a family of random variables $\{\nu_i^{(r)}\}_{i,r}$, where $\nu_i^{(r)}$ is the number of offspring (children) of the i-th individual (particle) alive in the r-th generation. Simple models assume that the population is haploid having non-overlapping generations, i.e. the parents are removed from the population and the children form the population of the next generation.

Without further assumptions on the joint distribution of the offspring variables $\nu_i^{(r)}$ it is not easy to derive any interesting or detailed results for such a population model. In order to imagine how large this class of models is, the following two cases are considered. Assume that the offspring variables $\nu_i^{(r)}$ are independent. Then this model is a typical Bienaymé–Galton–Watson branching process. Many scientists have been working in this field (Athreya & Ney, 1972; Harris, 1989; Jagers, 1975). Another interesting scenario appears when the considered population is in some biological equilibrium. This is usually modelled by the assumption that the population size is fixed to some constant N in each generation. In this case, for fixed r the offspring variables $\nu_i^{(r)}$ tend to be negatively correlated. The coalescent is the appropriate tool to analyse the ancestral structure for models of this type.

This article gives a basic introduction to and can be considered as well as a review about the theory of the coalescent. It starts with the above-mentioned models with fixed population size and collects the convergence-to-the-coalescent theorems so far known. The author is aware that such a review cannot be complete in the sense that everything about the "coalescent" is covered. Other excellent reviews on this topic are available (Donnelly & Tavaré, 1995; Hudson, 1991; Li & Fu, 1999), most of them focusing on the genetics impact of the coalescent. This article aims to illuminate the dynamical and mathematical aspects of the coalescent.

2. Haploid Population Models and their Ancestral Structure

We consider first the neutral haploid population models with discrete, non-overlapping generations $r \in \mathbb{N}_0 := \{0, 1, 2, \dots\}$ and fixed population size $N \in \mathbb{N} := \{1, 2, \dots\}$ introduced by Cannings (1974, 1975). In these models $\nu_i^{(r)}$, $i \in \{1, \dots, N\}$ denotes the number of descendants of the i-th individual alive in the r-th generation, $r \in \mathbb{N}$. As usual in ancestral population genetics the generations are labelled backward in time, i.e. if for example the r-th generation is the parent generation then the $(r-1)$-th generation is the children generation. As the population size is assumed to be fixed it follows for each $r \in \mathbb{N}$ that the offspring variables $\nu_i^{(r)}$, $i \in \{1, \dots, N\}$, have to satisfy the condition

$$\sum_{i=1}^{N} \nu_i^{(r)} = N. \tag{1}$$

The behaviour of the ancestral structure of such models has been first studied by Kingman (1982a–c), later by many other authors (see, for example Dannelly & Tavaré, 1995; Griffiths & Tavaré, 1994; Marjoram, 1992) for the class of exchangeable neutral models. For a finite sequence of random variables $\nu_1, \dots, \nu_N$ the exchangeability is defined by the property that the joint distribution of these variables is invariant under permutation, i.e. the distribution of $(\nu_{\pi 1}, \dots, \nu_{\pi N})$ does not depend on the special choice of the permutation π of the indices $1, \dots, N$. A typical example is the well-known Wright–Fisher model, where the offspring variables $\nu_1^{(r)}, \dots, \nu_N^{(r)}$ are symmetrical multinomially distributed, i.e.

$$P(\nu_1^{(r)} = k_1, \dots, \nu_N^{(r)} = k_N) = \frac{N!\, N^{-N}}{k_1! \cdots k_N!} \tag{2}$$

as long as $k_1 + \cdots + k_N = N$. Later, the results have been extended to more general models, for example for models where the exchangeability is not assumed any more (see for example Möhle, 1998a, 1999). Here we assume that

1. the offspring vectors $(\nu_1^{(r)}, \dots, \nu_N^{(r)})$, $r \in \mathbb{N}$ are independent and identically distributed for different generations and that
2. conditioned on the $\nu_i^{(r)}$ all "legitimate" assignments of offspring to parents (i.e. consistent with the $\nu_i^{(r)}$) are equally likely.

These two conditions are the key properties for all that follows. We shall see soon that the first condition ensures that the so-called ancestral processes considered later has the Markov property. The second assumption ensures that the transition probabilities of the ancestral process have a certain structure [see eqn (3)]. Note that the second condition is weaker than the assumption that the offspring variables are exchangeable. For example, it allows for ordered family sizes $\nu_1^{(r)} \geqslant \cdots \geqslant \nu_N^{(r)}$ within each fixed generation. Write ν_i for $\nu_i^{(1)}$ for convenience. Fix $n \leqslant N$ and sample n individuals at random from the current generation. For $r \in \mathbb{N}_0$ let $\mathscr{R}_r$ denote the equivalence relation which contains the pair (i, j) if and only if the i-th and the j-th individual of this sample have a common ancestor in the r-th generation backward in time. The above first condition ensures that the so-called ancestral process $(\mathscr{R}_r)_{r \in \mathbb{N}_0}$ is a time-homogeneous Markov chain. The state space is $\mathscr{E}_n$, the set of all equivalence relations on $\{1, \ldots, n\}$ and the initial value is $\mathscr{R}_0 = \Delta := \{(i, i) \mid i \in \{1, \ldots, n\}\}$. For $\xi, \eta \in \mathscr{E}_n$ let $p_{\xi\eta} := P(\mathscr{R}_r = \eta \mid \mathscr{R}_{r-1} = \xi)$ denote the probability for a transition of the ancestral process from ξ to η. Obviously, $p_{\xi\eta} = 0$ for $\xi \not\subseteq \eta$. Assume now that $\xi \subseteq \eta$. As in Kingman (1982b) let $C_1, \ldots, C_a$ denote the equivalence classes of η and let $C_{\alpha\beta}$, $\alpha \in \{1, \ldots, a\}$, $\beta \in \{1, \ldots, b_\alpha\}$ denote the equivalence classes of ξ such that $C_\alpha = \bigcup_{\beta=1}^{b_\alpha} C_{\alpha\beta}$. From the second condition, it follows by a combinatorial "putting balls into boxes" argument (see, for example, Kingman, 1982b; Möhle, 1998a) that the transition probability for the case $\xi \subseteq \eta$ is given by

$$p_{\xi\eta} = \frac{1}{(N)_b} \sum_{\substack{i_1, \ldots, i_a = 1 \\ \text{all distinct}}}^{N} E((\nu_{i_1})_{b_1} \cdots (\nu_{i_a})_{b_a}), \qquad (3)$$

where $b := |\xi|$ denotes the number of equivalence classes of ξ, $b_1, \ldots, b_a$ are the group sizes of merging equivalence classes of ξ and the notation $(x)_b := x(x-1)\cdots(x-b+1)$ is used.

3. Convergence Results for the Coalescent

The n-coalescent $(R_t)_{t \geqslant 0}$ (also called Kingman's coalescent) is a time-continuous Markov process with state-space $\mathscr{E}_n$, initial state Δ and infinitesimal generator $Q = (q_{\xi\eta})_{\xi, \eta \in \mathscr{E}_n}$ given by

$$q_{\xi\eta} := \begin{cases} -b(b-1)/2, & \text{if } \xi = \eta, \\ 1, & \text{if } \xi \prec \eta, \\ 0, & \text{otherwise,} \end{cases} \qquad (4)$$

where $b := |\xi|$ denotes the number of equivalence classes of ξ and $\xi \prec \eta$ if and only if $\xi \subseteq \eta$ and $|\xi| = |\eta| + 1$, i.e. η is obtained from ξ by combining (Latin: coalescere = to merge, to unite) two equivalence classes of ξ.

In 1982, Kingman (1982a–c) published the following first convergence result for the case of exchangeable reproduction. If the variance $\sigma_N^2 := \mathrm{Var}(\nu_1)$ converges to some constant $\sigma^2 \in (0, \infty)$ as N tends to infinity and if $\sup_N E(\nu_1^k) < \infty$ for all $k \in \mathbb{N}$, then the finite-dimensional distributions of the time-scaled ancestral process $(\mathscr{R}_{[N\sigma^{-2}t]})_{t \geqslant 0}$ converge to those of the n-coalescent. Kingman's proof is based on the expansion

$$P_N = I + N^{-1}\sigma_N^2 Q + O(N^{-2})$$

for the transition matrix $P_N := (p_{\xi\eta})_{\xi, \eta \in \mathscr{E}_n}$ of the ancestral process. This "Taylor expansion" is then used to verify that $\lim_{N \to \infty} P_N^{[N\sigma^{-2}t]} = \mathrm{e}^{tQ}$. The main fact of this convergence result is that one has to measure time in units of N generations in order to reach convergence to the coalescent.

The standard example is the Wright–Fisher model where the offspring variables $(\nu_1, \ldots, \nu_N)$ have the symmetrical multinomial distribution (2). In this case, $\mathrm{Var}(\nu_1) = 1 - 1/N$ converges to $\sigma^2 := 1$. The supremum condition is also satisfied as the descending factorial moments are given by $E((\nu_1)_k) = (N)_k N^{-k} \leqslant 1$. Unfortunately, the above convergence result is not applicable for some very important population models. For example, for the Moran model ($\nu_1 := 2$, $\nu_2 = \cdots = \nu_{N-1} := 1$, $\nu_N := 0$) it was first shown separately that the finite-dimensional distributions of the process $(\mathscr{R}_{[N^2 t/2]})_{t \geqslant 0}$ converge to those of the n-coalescent. Here the time has to be measured in units of $N^2/2$ generations. In comparison with the Wright–Fisher model this time-scaling is of order N higher. This is reasonable as under the Moran model the reproduction from generation

to generation is (more or less) of order N slower than in the standard Wright–Fisher model.

In recent years, it was a challenge in biology and mathematics to extend these convergence results to larger classes of population models. Obviously,

$$c(N) := \frac{1}{(N)_2} \sum_{i=1}^{N} E((\nu_i)_2) \tag{5}$$

is the probability that two individuals, chosen randomly without replacement from some generation, have a common ancestor one generation backward in time. This probability is called the *coalescence probability*. It is assumed that $c(N) > 0$, i.e. the trivial model $\nu_1 = \cdots = \nu_N := 1$ is avoided. The following more general convergence theorem is known since 1998 and shows that the coalescence probability is the basic quantity and very important to understand the theory about the coalescent.

Theorem 3.1. *Assume that the following conditions are satisfied.*

1. *Limit condition*: $\lim_{N\to\infty} c(N) = 0$.
2. *Moment conditions*:

$$\text{(a)} \quad \lim_{N\to\infty} \frac{1}{N^3 c(N)} \sum_{i=1}^{N} E((\nu_i)_2 \nu_i^k) = 0$$

$$\text{for all } k \in \mathbb{N}$$

$$\text{(b)} \quad \lim_{N\to\infty} \frac{1}{N^4 c(N)} \sum_{i,j=1}^{N} E((\nu_i)_2 \nu_j^2) = 0.$$

Then the process $(\mathscr{R}_{[t/c(N)]})_{t\geqslant 0}$ *converges weakly in* $D_{\mathscr{E}_n}([0, \infty))$ *to the n-coalescent as N tends to infinity.*

The proof of the convergence of the finite-dimensional distributions is given in Möhle (1998a) (in an even more general, non-time-homogeneous context) and the proof for the convergence in $D_{\mathscr{E}_n}([0, \infty))$ can be found in Möhle (1999). The theorem is for example applicable to the Moran model, where the coalescence probability is given by $c(N) = 2/(N(N-1)) \sim 2/N^2$ and further for all models with $\lim_{N\to\infty} N^2 c(N) = \infty$ and

$$\sup_N \frac{1}{N} \sum_{i=1}^{N} E(\nu_i^k) < \infty \quad \forall k \geqslant 2. \tag{6}$$

This includes especially the standard Wright–Fisher model. If the offspring variables $\nu_1, \ldots, \nu_N$ are identically distributed, eqn (6) reduces to the classical condition "$\sup_N E(\nu_1^k) < \infty$" used by Kingman (1982a, b). Theorem 3.1 has been extended to non-time-homogeneous models with deterministic variable population sizes (Möhle, 1998a, 1999). Unfortunately, for the general case of stochastic variable population size the literature is rather sparse and not a lot is known about the connection to the branching process models, where the offspring variables are assumed to be independent (Athreya & Ney, 1972; Harris, 1989; Jagers, 1975). Models with infinite population size have been studied originally by Fleming & Viot (1979), later by many other authors (see, for example, Donnelly & Kurtz, 1996; Ethier & Kutz, 1993).

The limit condition $\lim_{N\to\infty} c(N) = 0$ is in fact necessary to ensure convergence to the coalescent as otherwise no time-continuous process will appear as N tends to infinity. The moment conditions (a) and (b) are too strong. In 1999, the following convergence theorem was shown. It presents in some sense the "minimal" condition which is necessary and sufficient for convergence to the n-coalescent.

Theorem 3.2. *The process* $(\mathscr{R}_{[t/c(N)]})_{t\geqslant 0}$ *converges weakly in* $D_{\mathscr{E}_n}([0, \infty))$ *to the n-coalescent if and only if*

$$\lim_{N\to\infty} \frac{d(N)}{c(N)} = 0, \tag{7}$$

where $d(N) := ((N)_3)^{-1} \sum_{i=1}^{N} E((\nu_i)_3)$.

This means [see also eqns (3) and (5)] that the n-coalescent appears in the limit as the population size tends to infinity if and only if triple mergers of ancestral lines are asymptotically negligible in comparison with binary mergers. A proof of Theorem 3.2 for the case of exchangeable

reproduction is given in Möhle & Sagitov (1999b). It can be extended easily to the more general class of models considered here. Note that eqn (7) ensures that $\lim_{N\to\infty} c(N) = 0$. If the offspring variables are identically distributed, then eqn (7) reduces simply to

$$\lim_{N\to\infty} \frac{E((\nu_1)_3)}{N\, E((\nu_1)_2)} = 0. \tag{8}$$

4. Generalizations for More Complex Population Models

In recent years, the number of publications about the coalescent increased enormously. Theoretical and mathematical aspects are covered as well as applied statistical and biological topics and even numerical questions in computer science (see, for example, Beerli & Felsenstein, 1999 or Griffiths & Tavaré, 1996). The book of Donnelly & Tavaré (2000) gives some historical perspectives about recent research around the coalescent. For more details see also the reviews already mentioned in the introduction (Section 1). The theory of the coalescent has been extended to more complex models, for example for models with

- mutation (Donnelly & Tavaré, 2000; Tavaré, 1984), for
- diploid and two-sex population models (Möhle, 1998c; Möhle & Sagitov, 1999b), models with
- self-fertilization or partial selfing (Fu 1997; Möhle, 1998b; Nordborg & Donnelly, 1997), further for
- sub-divided population models and geographically structured models (Bahlo & Griffiths, 2000a, b; Beerli & Felsenstein, 1999; Herbots, 1994; Notohara, 1990; Wilkinson-Herbots, 1998), for models with
- recombination (Griffiths & Marjoram, 2000; Hey & Wakeley, 1997; Hudson & Kaplan, 1988)
- selection (Kaplan *et al.*, 1988; Krone & Neuhauser, 1997a, b)
- models with changing population size (Donnelly & Tavaré, 2000; Griffiths & Tavaré, 1994, 1996; Möhle, 1998a; Tajima, 1989).

In the following, the above models and connections between them are discussed in more detail. As far as possible citations are avoided for the rest of this section as they are already given in the list above. The interesting and fascinating result is that in all cases the coalescent (or some generalization of the coalescent) arises as the population size becomes large. This property is called the *robustness* of the coalescent. Robustness results are well known in probability theory. For example, the central limit theorem ensures convergence to the normal distribution for a large class of sums of random variables. In this sense, the author likes to think about Kingman's coalescent as the "genetics normal distribution". We now discuss the above models in more detail.

The haploid models have been first generalized by the assumption that mutations occur with probability μ per gene per generation independently of the underlying reproduction mechanism caused by the offspring variables $\nu_i^{(r)}$. In other words, the reproduction process is overlaid independently, i.e. in a neutral way, by a Bernoulli mutation process. If in addition to condition (7) the mutation rate $\theta := \lim_{N\to\infty} 2\mu/c(N)$ exists, then the time-scaled ancestral process $(\mathscr{R}_{[t/c(N)]})_{t\geqslant 0}$ converges weakly to a limit process $(R_t)_{t\geqslant 0}$, the so-called n-coalescent with mutation rate θ. In the limit independently of the underlying genealogy mutations occur on the branches of the tree according to a Poisson process with rate $\theta/2$. This Poisson process is the time-continous limit of the Bernoulli mutation process acting in the time-discrete model. The coalescent with mutation turned out to be a breakthrough in population genetics. For example, it led to a simple proof of the well-known Ewens sampling formula. Furthermore, in biological statistics it is extensively used to estimate the mutation rate θ.

Tracing back the ancestry of a sample of n genes in diploid or two-sex populations is indeed more complicated as genes chosen within an individual have another probability to have a common ancestor than genes belonging to different individuals or different couples. Nevertheless, under weak and realistic assumptions—after going backward in time a finite number of generations—the ancestral genes belong to different couples. From this generation

on the ancestral tree looks again like a coalescent. Thus, Kingman's assumption that the population has to be haploid is not essential and the coalescent theory was extended successfully to diploid and two-sex population models.

Besides the mutation rate θ several other parameters have been introduced to describe more complex population models. For example, the mechanism of reproduction in diploid plant populations often involves a mixture of self-fertilization and random mating. This is usually modelled by the assumption that with some probability s (the selfing rate) an individual is the offspring of a self-fertilization and with probability $1-s$ it is the offspring of a random mating. The coalescent theory has been extended to such population models in order to estimate simultaneously the mutation rate θ and the selfing rate s using approaches mostly based on frequency data or DNA sequence data.

In many cases, the population is not panmictic, for example it can be subdivided into several regions (colonies) due to geographic structure. Such populations can be modelled via subpopulation size parameters c_i and migration parameters m_{ij} controlling the migration between the colonies i and j. Under certain conditions on these migration parameters which essentially say that migration is rare in the sense that the m_{ij} are of order $1/c(N)$, the ancestry of a sample of genes chosen from such a subdivided population can be approximated via the so-called structured coalescent, a generalization of Kingman's coalescent. It keeps track of the location of the ancestors of the sample at each time. The generator of this structured coalescent process is described via the subpopulation size parameters c_i and the scaled migration parameters $M_{ij} := \lim_{N\to\infty} m_{ij}/c(N)$.

Populations with recombination have for example been modelled assuming that a gene, thought of as a length of DNA, is represented by the unit interval $[0,1]$. In the discrete Wright–Fisher model with probability $1-r$ a gene chooses one parental gene and with probability r two parental genes, when a recombination event occurs. In case of a recombination a break point position Z in the unit interval is chosen according to some probability measure and the gene is formed by combining the parts $[0,Z]$ and $(Z,1]$ from the first and the second parental gene. In most applications, Z has a discrete uniform distribution taking the values $1/m,\dots,(m-1)/m$ which leads to an m-locus model, but also continuous distributions of Z are important. It is assumed that $\rho = \lim_{N\to\infty} r/c(N)$ exists. The parameter ρ is called the recombination rate. In the limit $N\to\infty$ looking backward in time besides the usual mergers of ancestral lines also branches appear corresponding to a recombination event. The number of ancestors of the sample backward in time is a birth and death process with rates $\mu_k = k(k-1)/2$ and $\lambda_k = k\rho/2$. This limit process is called the coalescent with recombination. The genealogy of the sample is embedded in a graph with a coalescing and branching structure. This graph is called the "ancestral recombination graph". In a (at first glance) similar way an "ancestral selection graph" has been introduced in order to discuss population models evolving according to random reproduction with selection (and mutation). Here a continuous Moran model has been studied extensively, but it is mentioned that the results should not depend on the particular model. For simplicity, focus on a single locus having a finite number of possible types (alleles) $1,\dots,K$. The K allele model with selection and mutation can be modelled introducing the following three sets of parameters.

1. An individual of type $i\in\{1,\dots,K\}$ gives birth to a new offspring with rate λ_i,
2. this offspring is a mutant with probability $p_N(i)$ and
3. given that the offspring is a mutant, a transition from type i to j occurs with probability γ_{ij}.

Without loss of generality, assume that $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_K$, i.e. the type $i+1$ has a selective advantage in comparison to the type i such that $s_N(i)$ defined by

$$\lambda_i = \lambda_1(1+s_N(i)), \quad i\in\{2,\dots,N\}$$

satisfies $s_N(i)\geqslant 0$. For this situation the following convergence theorem is available. If $\lambda_1 = N/2$, $Ns_N(i)\to\sigma_i$ and if $Np_N(i)\to\theta_i$ with a rate higher than some negative power of N, then as N tends to infinity a coalescing, branching object is

obtained called the "ancestral selection graph". In the limit mutations occur with rate $\theta_i/2$ on the branches and the types change according to the mutation transition probability matrix $\Gamma = (\gamma_{ij})$. The proof of Krone & Neuhauser (1997a, b) for this convergence result uses the concept of duality which is a well known tool to analyse processes in the theory of interacting particle systems (Liggett, 1985).

Kingman's coalescent theory was mainly developed for populations of constant size. It has been also extended to the case of deterministically changing population size in which all the generations are assumed to be large. More precisely, let $N := M_0$ denote the population size of the current generation and for $r = 1, 2, \ldots$ let $M_r = M_r(N)$ denote the population size r generations backward in time, i.e. $M_{r-1} = \sum_{i=1}^{M_r} \nu_i^{(r)}$. A convergence theorem similar to Theorem 3.1 is available which ensures that $(\mathscr{R}_{\tau_N(t)})_{t \geqslant 0}$ converges weakly to the n-coalescent, where τ_N is a properly chosen integer-valued time-scaling function. The limit condition of Theorem 3.1 generalizes to

$$\lim_{N \to \infty} \sum_{r=1}^{\tau_N(t)} c_\tau = t \quad \forall t \geqslant 0, \tag{9}$$

where

$$c_r := \frac{1}{(M_{r-1})_2} \sum_{i=1}^{M_r} E((\nu_i^{(r)})_2)$$

is the coalescence probability in generation r, i.e. the probability that two individuals, chosen randomly without replacement from generation $r-1$, have a common ancestor in generation r. For the constant population size case the time-scaling function is given by $\tau_N(t) = [t/c(N)]$ and eqn (9) reduces to the limit condition of Theorem 3.1. It remains open to derive a generalization of Theorem 3.2 for the variable population size case. Another open and very challenging question is to study models with random population size and in this context connections and differences to branching processes.

5. Convergence to the General Coalescent

In this last section the author wants to introduce the general n-coalescent. In order to make things easy we go back to the simple haploid models without any biological forces like mutation, recombination or selection. Convergence to Kingman's coalescent takes place if and only if condition (7) is satisfied. We will see in this section that this condition is not always satisfied, for example when some of the offspring variables ν_i tend to be large in the sense that some of them have the same order as the total population size N. Biological examples for populations where the ancestral structure might differ from Kingman's coalescent are fish populations or populations with artificial insemination.

From eqn (3) it follows that there exists a global constant L such that $\|P_N - I\|/c(N) \leqslant L$ for all sufficiently large N, where P_N denotes as before the transition matrix of the ancestral process. Thus, there exists a subsequence $(N_k)_{k \in \mathbb{N}}$ with $\lim_{k \to \infty} N_k = \infty$ such that $c := \lim_{k \to \infty} c(N_k)$ exists and such that

$$Q = (q_{\xi\eta})_{\xi,\eta \in \mathscr{E}_n} := \lim_{k \to \infty} \frac{P_{N_k} - I}{c(N_k)} \tag{10}$$

exists. Further, $q_{\xi\eta} = 0$ for $\xi \not\subseteq \eta$, $q_{\xi\eta} \in [0, 1]$ for $\xi \subset \eta$ and $q_{\xi\xi} = -\sum_{\eta \neq \xi} q_{\xi\eta}$. If $c > 0$ then the finite-dimensional distributions of the process $(\mathscr{R}_r)_{r \in \mathbb{N}_0}$ converge as $k \to \infty$ to those of a discrete time Markov process $(R_r)_{r \in \mathbb{N}_0}$ with initial state $R_0 = \Delta$ and transition matrix $I + cQ$.

Assume now that $c = 0$. Then the process $(\mathscr{R}_{[t/c(N_k)]})_{t \geqslant 0}$ converges as $k \to \infty$ weakly in $D_{\mathscr{E}_n}([0, \infty))$ to a time-continuous Markov process $(R_t)_{t \geqslant 0}$ with initial state $R_0 = \Delta$ and transition matrix e^{tQ}. The matrix Q is equal to the generator of the n-coalescent if and only if eqn (7) is satisfied.

In what follows, the so-called method of mixing is used to construct a population model, where the limit-process is not equal to the n-coalescent. The basic idea is to mix two population models in a certain way. First, the two population models are described. Then the mixing procedure is explained.

5.1. MODEL 1

Assume that each individual has N offspring with probability $1/N$ and no offspring with probability $1 - 1/N$. The coalescence probability for

this simple model is just $c^{(1)} = 1$ and the transition matrix of the ancestral process is given by

$$P^{(1)} = \begin{pmatrix} 0 & 1 \\ & \vdots \\ & 1 \end{pmatrix}.$$

Thus, the infinitesimal generator of this model has the form

$$Q^{(1)} = P^{(1)} - I = \begin{pmatrix} -1 & & & 1 \\ & \ddots & & \vdots \\ & & -1 & 1 \\ & & & 0 \end{pmatrix}.$$

5.2. MODEL 2

This population model is even more simple. Assume that each individual has exactly one offspring, i.e. $\nu_1 \equiv 1$. Then the coalescence probability is given by $c^{(2)} = 0$ and the transition matrix $P^{(2)} = I$ is the identity.

The idea is now to mix these two independent models with some probability p, i.e. in each generation with probability p reproduction occurs according to the first model and with probability $1 - p$ the population evolves according to the other model. For our purpose a good choice for p is $p := 1/N$. This ends up in a population model with coalescence probability $c(N) = pc^{(1)} + (1-p)c^{(2)} = p$, backward transition matrix $P_N = pP^{(1)} + (1-p)P^{(2)} = p(P^{(1)} - I) + I$ and generator $Q := \lim_{N\to\infty}(P_N - I)/c(N) = P^{(1)} - I = Q^{(1)}$. This matrix Q is obviously *not* equal to the generator of Kingman's coalescent given by eqn (4). The transition matrix of the limit process has the form

$$\Pi(t) = \begin{pmatrix} e^{-t} & & & 1 - e^{-t} \\ & \ddots & & \vdots \\ & & e^{-t} & 1 - e^{-t} \\ & & & 1 \end{pmatrix},$$

which corresponds obviously to a star-shaped ancestral tree and not to Kingman's coalescent process.

Obviously, the set of all possible Q's arising in eqn (10) is convex, i.e. if one chooses two population models with corresponding generators Q_1 and Q_2, respectively, then for each $p \in [0, 1]$ there exists a (mixed) population model with corresponding generator $pQ_1 + (1-p)Q_2$. In Möhle & Sagitov (1999a) a full classification of all possible generators Q is given as follows. The limit $Q = \lim_{N\to\infty}(P_N - I)/c_N$ exists if and only if the limits

$$\phi_a(b_1, \ldots, b_a) := \lim_{N\to\infty} \frac{E((\nu_1)_{b_1} \cdots (\nu_a)_{b_a})}{N^{b_1 + \cdots + b_a - a} c(N)},$$

$1 \leqslant a \leqslant b := b_1 + \cdots + b_a \leqslant n$ exist and in this case Q has the entries

$$q_{\xi\eta} := \begin{cases} -\sum_{i=1}^{b-1} i\phi_i(2, 1, \ldots, 1), & \text{if } \xi = \eta, \\ \phi_a(b_1, \ldots, b_a), & \text{if } \xi \subset \eta, \\ 0, & \text{otherwise,} \end{cases}$$

where the connection between the equivalence relations ξ and η and the integers $a, b, b_1, \ldots, b_a$ is given at the end of Section 2. The process corresponding to this generator is called the general coalescent. It allows for simultaneous multiple mergers of ancestral lines. The sub-class of the processes with multiple mergers of ancestral lines has been studied in more detail by Pitman (1999) and Schweinsberg (1999). Recently, Schweinsberg (2000) discussed also coalescent processes with simultaneous multiple collisions and "infinite" sample size.

The author wishes to thank the organizers of the fifth international conference on mathematical population dynamics, especially Ellen Baake for organizing the plenary talks of the population genetics and evolutionary dynamics sessions. Many thanks go to Serik Sagitov, Adam Bobrowski, Ali Falahati, Marek Kimmel and to all participants of the conference for sharing their scientific experiences. The author acknowledges helpful comments of two anonymous referees on the first version of the manuscript. Last but not least, the author would like to thank his wife Norma Regina Soares Laurindo Möhle for her patience and support during the preparation of this article.

REFERENCES

ARNOLD, L. (1994). *Zufällige dynamische Systeme*, *Jahresbericht der Deutschen Mathematiker-Vereinigung* (DMV), Band 96, Heft 3, pp. 85–100, Stuttgart: Teubner.

ATHREYA, K. B. & NEY, P. E. (1972). *Branching Processes*. Berlin: Springer.

BAHLO, M. & GRIFFITHS, R. C. (2000a). Inference from gene trees in a subdivided population. Preprint, Department of Statistics, University of Oxford, UK. Available via http://www.stats.ox.ac.uk/mathgen/publications.html (submitted to Theor. Popul. Biol.)

BAHLO, M. & GRIFFITHS, R. C. (2000b). Coalescence time for two genes from a subdivided population. Preprint, Department of Statistics, University of Oxford, UK. Available via http://www.stats.ox.ac.uk/mathgen/publications.html (submitted to *J. Math. Biol.*)

BEERLI, P. & FELSENSTEIN, J. (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152,** 763–773.

CANNINGS, C. (1974). The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. *Adv. Appl. Probab.* **6,** 260–290.

CANNINGS, C. (1975). The latent roots of certain Markov chains arising in genetics: a new approach, II. Further haploid models. *Adv. Appl. Probab.* **7,** 264–282 (1975).

DONNELLY, P. & KURTZ, T. G. (1996). A countable representation of the Fleming–Viot measure-valued diffusion. *Ann. Probab.* **24,** 698–742.

DONNELLY, P. & TAVARÉ, S. (1995). Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29,** 401–421.

DONNELLY, P. & TAVARÉ, S. (2000). *Ancestral Processes in Population Genetics* Berlin: Springer (in press).

ETHIER, S. N. & KURTZ, T. G. (1986). *Markov Processes, Characterization and Convergence*. New York: Wiley.

ETHIER, S. N. & KURTZ, T. G. (1993). Fleming–Viot processes in population genetics. *SIAM J. Control Optim.* **31,** 345–386.

ETHIER, S. N. & NAGYLAKI, T. (1980). Diffusion approximations of Markov chains with two time scales and applications to population genetics. *Adv. Appl. Probab.* **12,** 14–49.

FISHER, R. A. (1958). *The Genetical Theory of Natural Selections*. Oxford: Oxford Univ. Press; New York: Dover.

FLEMING, W. H. & VIOT, M. (1979). Some measure-valued Markov processes in population genetics theory. *Indiana Univ. Math. J.* **28,** 817–843.

FU, Y. X. (1997). Coalescent theory for a partially selfing population. *Genetics* **146,** 1489–1499.

GRIFFITHS, R. C. & MARJORAM, P. (2000). An ancestral recombination graph. *Population Genetics and Human Evolution.* Berlin: Springer (in press).

GRIFFITHS, R. C. & TAVARÉ, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philos. Trans. Roy. Soc. London B* **344,** 403–410.

GRIFFITHS, R. C. & TAVARÉ, S. (1996). Monte Carlo inference methods in population genetics. *Math. Comput. Modelling* **23,** 141–158.

HARRIS, T. E. (1989). *The Theory of Branching Processes.* New York: Dover Publications, Inc.

HERBOTS, H. M. (1994). Stochastic models in population genetics: genealogy and genetic differentiation in structured populations. Ph.D. Thesis, Queen Mary and Westfield College (QMW), University of London.

HEY, J. & WAKELEY, J. (1997). A coalescent estimator of the population recombination rate. *Genetics* **145,** 833–846.

HUDSON, R. R. (1991). Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7,** 1–44.

HUDSON, R. R. & KAPLAN, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120,** 831–840.

JAGERS, P. (1975). *Branching Processes with Biological Applications.* New York: Wiley.

KAPLAN, N. L., DARDEN, T. & HUDSON, R. R. (1988). The coalescent process in models with selection. *Genetics* **120,** 819–829.

KINGMAN, J. F. C. (1982). On the genealogy of large populations. *J. Appl. Probab.* **19A,** 27–43.

KINGMAN, J. F. C. (1982b). Exchangeability and the evolution of large populations. In: *Exchangeability in Probability and Statistics* (Koch, G. & Spizzichino, F., eds), pp. 97–112. Amsterdam: North-Holland Publishing Company.

KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* **13,** 235–248.

KRONE, S. M. & NEUHAUSER, C. (1997a). The genealogy of samples in models with selection. *Genetics* **145,** 519–534.

KRONE, S. M. & NEUHAUSER, C. (1997b). Ancestral processes with selection. *Theor. Popul. Biol.* **51,** 210–237.

LI, W. H. & FU, Y. X. (1999). Coalescent theory and its application in population genetics. In: *Statistics in Genetics* (Halloran, M. E. & Geisser, S., eds), IMA Volumes in Mathematics and Its Applications, Vol. 112.

LIGGETT, T. M. (1985). *Interacting Particle Systems.* Berlin: Springer.

MARJORAM, P. (1992). Correlation structures in applied probability. Ph.D. Thesis, University College London.

MÖHLE, M. (1998a). A classification of coalescent process for haploid exchangeable population models, Preprint No. 1999:10, Department of Mathematical Statistics, Chalmers University of Technology and Göteborg University. Sweden. Available via http://www.math.chalmers.se/Math/Research/Preprints (submitted to Ann. Probab.)

MÖHLE, M. (1998b). Coalescent patterns in exchangeable deploid population models, Preprint No. 1999:1, Department of Mathematics, University of Mainz, Germany. Available via http://www.mathematik.uni-mainz.de/Stochastik/Arbeitsgruppe/staff/moehle/moehle.html (submitted to J. Math. Biol.)

MÖHLE, M. (1998c). Coalescent results for two-sex population models. *Adv. Appl. Probab.* **30,** 513–520.

MÖHLE, M. (1999). Weak convergence to the coalescent in neutral population models. *J. Appl. Probab.* **36,** 446–460.

MÖHLE, M. & SAGITOV, S. (1999a). A classification of coalescent processes for haploid exchangeable population models, Preprint No. 1999:10, Department of Mathematical Statistics, Chalmers University of Technology and Göteborg University, Sweden. Available via http://www.math.chalmers.se/Math/Research/Preprints (submitted to *Ann. Probab.*)

MÖHLE, M. & SAGITOV, S. (1999b). Coalescent patterns in exchangeable diploid population models. Preprint No. 99-1, Department of Mathematics, University of Mainz, Germany. Available via http.//www.mathematik.unimainz.de/Stochastik/Arbeitsgruppe/staff/moehle/moehle.html (submitted to *J. Math. Biol*)

NAGYLAKI, T. (1990). Models and approximations for random genetic drift. *Theor. Popul. Biol.* **37,** 192–212.

NORDBORG, M. & DONNELLY, P. (1997). The coalescent process with selfing. *Genetics* **146,** 1185–1195.

NOTOHARA, M. (1990). The coalescent and the genealogical process in geographically structured population. *J. Math. Biol.* **29,** 59–75.

PITMAN, J. (1999). Coalescents with multiple collisions. Technical Report No. 495. Department of Statistics, University of California, Berkeley. Available via http://www.stat.berkeley.edu/tech-reports/index.html.

SCHWEINSBERG, J. (1999). A necessary and sufficient condition for the Λ-coalescent to come down from infinity. Technical Report No. 568, Department of Statistics, University of California, Berkeley. Available via http://www.stat.berkeley.edu/tech-reports/index.html.

SCHWEINSBERG, J. (2000). Coalescents with simultaneous multiple collisions. Technical Report No. 571, Department of Statistics, University of California, Berkeley. Available via http://www.stat.berkeley.edu/tech-reports/index.html.

TAJIMA, F. (1989). The effect of change in population size on DNA polymorphism. *Genetics* **123,** 597–601.

TAVARÉ, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* **26,** 119–164.

WILKINSON-HERBOTS, H. M. (1998). Genealogy and subpopulation differentiation under various models of population structure. *J. Math. Biol.* **37,** 535–585.

WRIGHT, S. (1969). *Evolution and the Genetics of Populations*, Vol. 1. Chicago: University of Chicago Press.