On the genealogy of populations: trees, branches and offspring

Maurizio Serva

Dipartimento di Matematica, Università dell'Aquila, I-67010 L'Aquila, Italy
E-mail: serva@univaq.it

Abstract. We consider a neutral haploid population whose generations are not overlapping and whose size is large and constantly of N individuals. Any generation is replaced by a new one and any individual has a single parent. We do not choose the stochastic rule assigning the number of offspring to any individual since results do not depend on the dynamics details, and, as a consequence, the model is parameter free. The genealogical tree is very complex, and distances between individuals (number of generations from the common ancestor) are distributed according to probability density which remains random in the thermodynamic limit (large population). We give a theoretical and numerical description of this distribution and we also consider the dynamical aspects of the problem describing the time evolution of the maximum and mean distances in a single population.

Contents

1. Introduction

In a population with asexual reproduction any individual has a single parent in the previous generation. If the size of population is constant, some of the individuals may have the same parent and, therefore, the number of ancestors of the present population decreases going backward in time. At a finite past time one has complete coalescence and all the population has a single ancestor. The genealogical distance between two individuals is simply the number of generations from the common ancestor. The resulting genealogical tree is very complex and has many branches; nevertheless, one would expect that in the limit of infinite population size, some quantities would reach some thermodynamic deterministic value. For example this could be the case for the frequency of genealogical distances in a single population or, at least, for the mean genealogical distance obtained considering all pairs of individuals. In contrast, however, the frequency of distances in a single population is random even in the thermodynamic limit. This means that this frequency is different for different populations and also the mean distance obtained considering all pairs in a single population is random. This non-self-averaging behaviour is known thanks to pioneering works of Derrida, Bessis and Peliti [4, 7]. In this paper we consider only the genealogical aspects of the problem, since mutation, at this level, is only a measure of genealogical distance through the Hamming distance.

Let us define the model. We consider a population with asexual reproduction and whose generations are not overlapping in time. Any generation is replaced by a new one and any individual has a single parent. The population size is large and constantly of N individuals; therefore, the average number of offspring of any individual is one. The stochastic rules assigning the number of offspring to any individual can be chosen in many ways. In fact, results do not depend on the details of this rule, the only requirement is

that the probability of having the same parent for two individuals must be of order $1/N$ for large N. As a consequence of the freedom in the choice of the rule, the model is parameter free. This is a typical situation if reproduction involves a fraction of order N of the population. To be clearer we give two examples of stochastic dynamics that satisfy this assumption. First rule: at any generation one half of the individuals (chosen at random) has no offspring while the remaining part has two (see [23]). With this rule the probability of having the same parent is $1/(N-1)$. The second rule (Wright and Fisher) is that any individual in the new generation chooses one parent at random in the previous one, independently on the choice of the others. In this case the probability of having the same parent for two individuals is exactly $1/N$.

In this paper we continue the work started in [19] obtaining analytical and numerical results on statics and dynamics of the model. For numerical results we simulate a population of some hundred individuals for 10^7 generations according to the Wright–Fisher rule. The population is large enough to avoid finite size corrections and time is sufficiently long to profit from ergodicity for replacing sample averages with time averages. Notice that we will use the world 'mean' intending mean over different individuals of the same population and we will use 'average' to intend average over many realizations of the population process or, equivalently, by ergodicity, average over the same population at different times.

The relevant quantity is the random probability density (rpd) of pair distances in a single population. This quantity differs for different populations and changes in time for a single population. The aim of this paper is to obtain both the statistics of this density and some information about its dynamics. Sections 2–5 are devoted to the first part of this programme while section 6 is devoted to its dynamical aspects.

In the final section we point out the open problems and, following the line of discussion started in [19], we show the relevance of results for the genealogy of mitochondrial DNA (mtDNA) populations.

2. Distribution of the pair distance

The genealogical tree of a population of N individuals is determined by considering the set of all genetical distances between them. The distance between two given individuals is the number of generations from the common ancestor and since there are $N(N-1)/2$ possible pairs we have to specify $N(N-1)/2$ distances.

For large N, distances are proportional to N so it is useful to rescale them by dividing by N. Equivalently we can say that distances are defined as the time from the common ancestor and contemporaneously define time as the number of generations divided by N.

Let us call the rescaled distance between individuals α and β in the population $d(\alpha, \beta)$. By definition, if α and β coincide, the distance vanishes ($d(\alpha, \alpha) = 0$). In contrast, for two distinct individuals α and β in the same generation one has

$$d(\alpha, \beta) = d(g(\alpha), g(\beta)) + \frac{1}{N}, \tag{1}$$

where $g(\alpha)$ and $g(\beta)$ are the two parent individuals which coincide with probability $1/N$ and are distinct individuals α' and β' with probability $(N-1)/N$. In other words, $d(\alpha, \beta) = 1/N$ with probability $1/N$ and $d(\alpha, \beta) = d(\alpha', \beta') + 1/N$ with probability $(N-1)/N$.

The above equation entirely defines the dynamics of the population, and simply states that the rescaled distance in the new generation increases by $1/N$ with respect to the parent distance. This dynamics can be easily simulated and it can be stopped at a given time (much larger than N in order to forget initial conditions). The distances obtained are different for different pairs and their frequency can be calculated. For finite N, frequency is simply the number of pairs in a given population with given distance x divided by the total number $N(N-1)/2$ of possible pairs.

This frequency inside a single population of 500 individuals can be seen in figure 1. It is immediately seen that this frequency is quite wild, due to the fact that individuals naturally cluster in subpopulations. In fact, most of the distances assume a few of values corresponding to the distances between the major subpopulations.

One could think that this singular behaviour would disappear in the thermodynamic limit of large N. On the contrary; not only does the singularity remain, but also one readily realizes that this frequency remains random, being different for different populations and different for the same population at different times. Indeed, even the mean distance in a population and the largest distance in a population are random quantities in the thermodynamic limit, as we will see in the next section.

Let us stress again that we use 'mean' hereafter intending the mean over different pairs of the same population and we use 'average' to intend the average over many realizations of the population process or, equivalently by ergodicity, the average over the same population at different times. Averages will be indicated by $\langle \cdot \rangle$.

In place of frequency we can consider the density $q(x)$:

$$q(x) = \frac{2}{N(N-1)} \sum_{\alpha > \beta} \delta(x - d(\alpha, \beta)), \tag{2}$$

were the δ indicates the Dirac delta function.

This quantity is simply related to the frequency since $q(x)\,\mathrm{d}x$ is the number of pairs whose distance lies in the interval $[x - (\mathrm{d}x/2), x + (\mathrm{d}x/2)]$ divided by the total number $N(N-1)/2$.

The random and singular nature of the density remains in the $N \to \infty$ limit and it is much the same as that of the overlap function in mean field spin glasses. In fact, they show similar non-self-averaging properties. Indeed, if one could give the probability distribution of $q(x)$, the complete specification of the static properties of the model would be reached. We postpone this goal to section 5 and we only compute in this section the average of the density $\langle q(x) \rangle$ and, in the next two sections, the distribution of the largest distance (the distribution of the maximum of the support of $q(x)$) and the first two moments of the distribution of the mean distance.

In order to derive analytically the average density $\langle q(x) \rangle$ we recall the following result obtained in [19]. The average probability density for $d(\alpha, \beta)$ (i.e. $\langle \delta(x - d(\alpha, \beta)) \rangle$) is simply $\exp(-x)$. We remark that this is not the density of the distances inside a single large population, but the average distribution of two individuals' distances sampled over many stochastically equivalent populations or, which is the same, sampled over the same population at many different times.

Notice that this result was already implicitly found in [7]. In fact, in [7], the genetic overlap $o(\alpha, \beta)$ of two individuals is deterministically associated with the genealogical distance by $o(\alpha, \beta) = \exp(-d(\alpha, \beta)/\lambda)$ and the probability density for $o(\alpha, \beta)$ is given
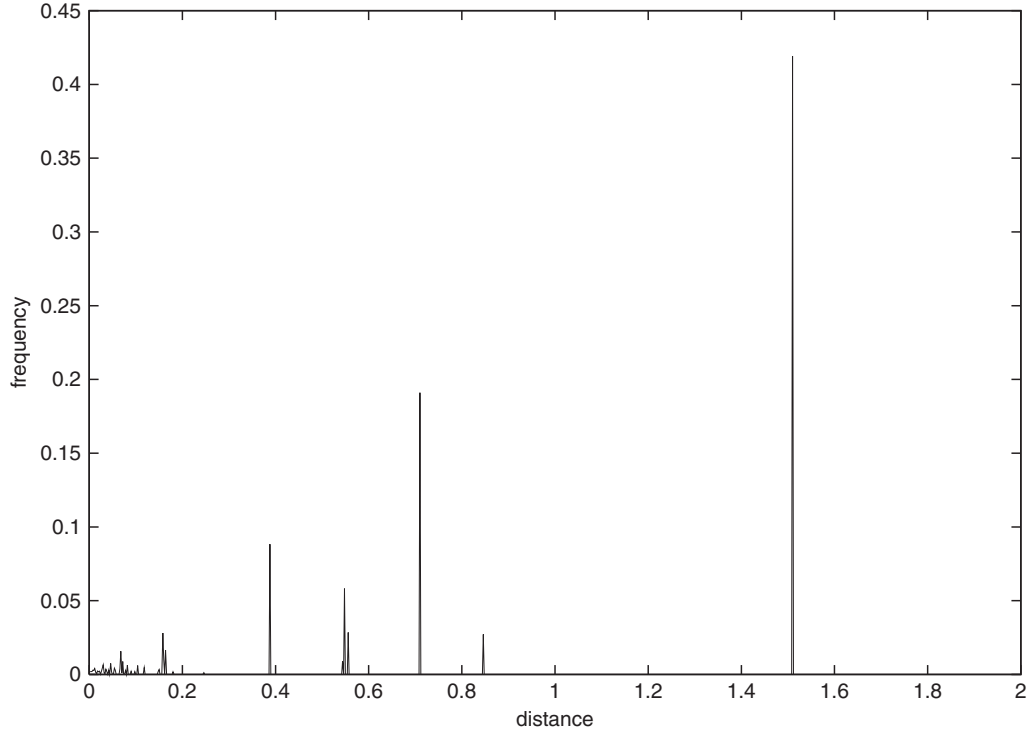
Figure 1. Frequency $q(x)\,dx$ of distances in a single population. Here we compute this quantity for a population of 500 individuals. Most of the distances assume a few of the values corresponding to the distances between major subpopulations. Notice that this frequency is very different from its average $\exp(-x)\,dx$.

as $\lambda x^{\lambda-1}$, which is directly obtainable from the density $\exp(-x)$ for the distance. The deterministic relation between distance and overlap is due to the infinite genome limit and λ is simply the inverse of the mutation rate. Let us mention that the Hamming distance is linearly associated with the overlap via $1 - o(\alpha, \beta)$. In conclusion, one can easily understand that all the complex behaviour of the genetics of the populations is due to the complexity of the structure of the genealogical tree, the role of mutation being simply accounted for by the relations $o(\alpha, \beta) = \exp(-d(\alpha, \beta)/\lambda)$.

We can now easily compute the average density using $\langle \delta(x - d(\alpha, \beta)) \rangle = \exp(-x)$. In fact, it is immediate that

$$\langle q(x) \rangle = \frac{2}{N(N-1)} \sum_{\alpha > \beta} \langle \delta(x - d(\alpha, \beta)) \rangle = \exp(-x). \tag{3}$$

This smooth average density is completely different from that for a typical sample. To appreciate this, look again at figure 1 where the frequency $q(x)\,dx$ is plotted. The most important consequences of this randomness will be discussed in the next section.

3. Distribution of mean and maximum distances

Let us introduce now two quantities which describe the 'thermodynamic' state of a population.

The first is the mean distance

$$d = \frac{2}{N(N-1)} \sum_{\alpha > \beta} d(\alpha, \beta), \tag{4}$$

which is simply the mean for a single population (and at a given time) of the internal distances considering all the $N(N-1)/2$ possible pairs. The above equation can be simply rewritten as $d = \int y \, q(y) \, \mathrm{d}y$. Since the probability density $q(y)$ is random, we expect d to also be random.

The second quantity is the maximum distance

$$d_{\max} = \max_{\{\alpha, \beta\}} d(\alpha, \beta), \tag{5}$$

which is the largest distance in a single population, i.e. the maximum of the support of $q(y)$. Again, as a consequence of the randomness of the density $q(y)$ we expect that $d_{\max}$ is also random. This quantity can be interpreted as the time from the common ancestor of the whole population and it has an evident relevance in palaeontology. In fact, the mtDNA of a single species is only transmitted by females and, therefore, it can be considered as a haploid population. As far as modern humans are concerned, $d_{\max}$ is the time from the celebrated mitochondrial Eve.

These two quantities can be studied in the context of the coalescence problem which has been widely investigated in a number of papers in the last two decades [2, 8], [11]–[14], [17, 18, 21, 22], and is still being investigated at present times [3, 9, 10, 19]. We will come back to this approach in the next two sections.

Both the distances d and $d_{\max}$ are random quantities even in the infinite population size limit and our goal is to find their density distributions $\rho(x) = \langle \delta(x - d) \rangle$ and $\rho_{\max}(x) = \langle \delta(x - d_{\max}) \rangle$. We have computed them numerically, iterating the dynamics (1) for 10^7 generations for a population of $N = 100$ individuals. The results are shown in figure 2 where both the numerical densities are plotted.

The theoretical $\rho_{\max}(x)$ will be obtained in the next section for a thermodynamic ($N = \infty$) population and it is also plotted in figure 2. The coincidence of the numerical and theoretical densities proves that $N = 100$ can be already considered large.

In contrast, we have not been able to deduce theoretically the density $\rho(x)$. Nevertheless, we computed in [19] its first two moments and we have shown there how in principle and with a lot of work the higher moments can be obtained.

The theoretical values of the first two moments are

$$\langle d \rangle = 1, \qquad \langle d^2 \rangle = \tfrac{11}{9}, \tag{6}$$

which coincide with the numerical values obtained from the string of 10^7 generations. The above results are related to those in [7] where analogous quantities are computed for the mean overlap of a population.

Let us stress again that higher moments can be computed following the strategy in [19]. The result can always be found by solving a system of linear equations. The problem is that the number of equations in the system grows exponentially with the power of the moment.
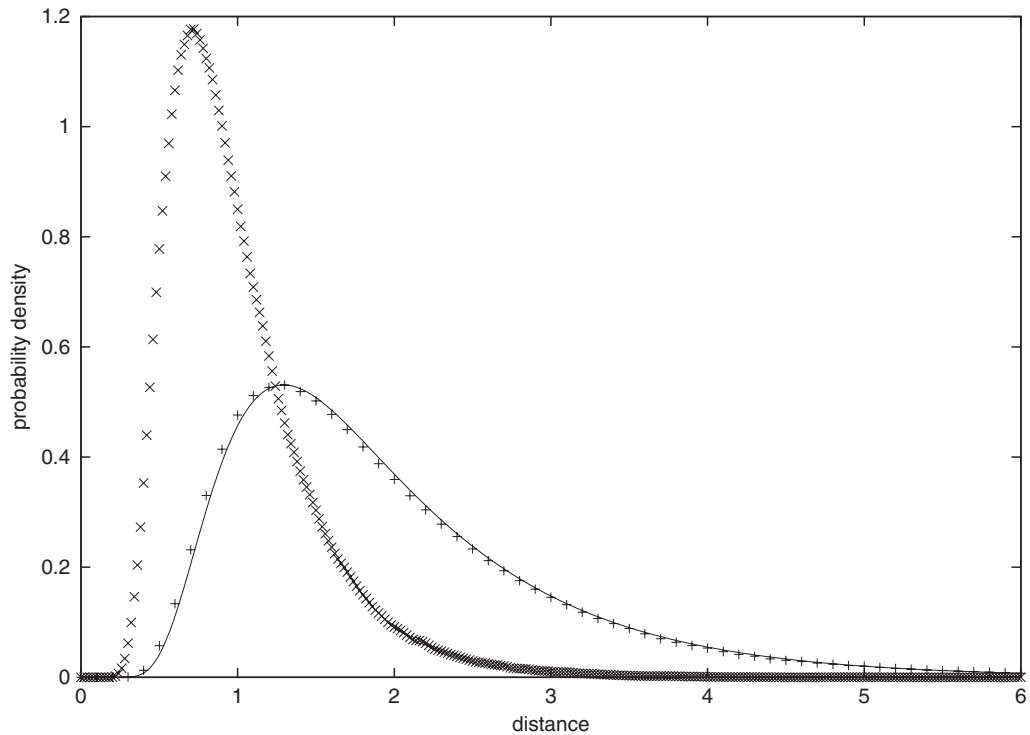
Figure 2. Probability density for the maximum distance ($+$) and for the mean distance ($\times$). The two densities are computed using a sample resulting from the dynamics of a population of 100 individuals for 10^7 generations. The continuous line corresponds to the theoretical probability density for the maximum distance ($N = \infty$). Since the empirical density corresponds to 100 individuals only, we deduce that a population size of 100 is sufficiently large to destroy finite size effects.

4. Coalescence

The content of this section is devoted to the most studied problem for this model: the coalescent. The idea is very simple and goes back to the papers of Kingman [11]–[14] and some results have also been independently discovered in [4, 7].

Consider a sample of n individuals in a population of size N. The probability that they all have different parents in the previous generation is $\prod_{k=0}^{n-1} (1 - k/N)$. Therefore, the probability that their ancestors are still all different in a past time t corresponding to tN generations is $[\prod_{k=0}^{n-1} (1 - k/N)]^{tN}$. If N is large compared to n this quantity is approximately $\exp(-c_n t)$ where $c_n = n(n-1)/2$. Therefore, the average probability density for first coalescence is

$$p_n(t) = c_n \exp(-c_n t). \tag{7}$$

This expression is the probability density for the first past time at which the ancestors of the n individuals reduce to $n - 1$. In particular, for $n = 2$ one also re-obtains the probability density $\exp(-x)$ for the distance of a pair of individuals already found in [19].

At the random time τ_n distributed according to the exponential of the parameter c_n, the number of ancestors is $n - 1$ and one has to go back an exponentially distributed

time τ_{n-1} with parameter c_{n-1} before further coalescence occurs and so on. Therefore, the joint probability density $\prod_{k=m+1}^{n} p_k(t_k)$ gives the statistics for successive coalescence times $\tau_n, \tau_{n-1}, \ldots, \tau_{m+1}$ until the number of ancestor reduces to m. This is the core of the celebrated coalescent, which is mostly associated with the name of the probability theorist Kingman.

If one wants to know the density distribution of the coalescence time from n individuals to m ancestors, one simply has to compute the convolution of the $n - m$ successive exponentials. In other words this random time is simply the sum $\sum_{k=m+1}^{n} \tau_k$.

The computation of the coalescence time from all N individuals to m ancestors needs some care in dealing with limits since, in this case, $n = N$. Nevertheless, for large N, one obtains $\sum_{k=m+1}^{\infty} \tau_k$. In particular, when $m = 1$, one has $d_{\max} = \sum_{k=2}^{\infty} \tau_k$.

In order to compute explicitly the statistics for $d_{\max}$, let us define $\rho_n(x)$ as the density distribution of time for complete coalescence of n individuals to a single ancestor. We have the convolution

$$\rho_n(x) = \int_0^x \mathrm{d}t \, p_n(x) \, \rho_{n-1}(t - x), \tag{8}$$

with the obvious $\rho_2(x) = p_2(x) = \exp(-x)$. Then, the density for $d_{\max}$ is simply

$$\rho_{\max}(x) = \lim_{n \to \infty} \rho_n(x). \tag{9}$$

In appendix A we compute explicitly the convolution (8) and we obtain the simple sum representation for the coalescent probability density $\rho_n(x)$:

$$\rho_n(x) = \sum_{l=2}^{n} (-1)^l (2l - 1) c_l \left(\prod_{s=1}^{l-1} \frac{n - s}{n + s} \right) \exp\{-c_l \, x\}. \tag{10}$$

In the limit $n \to \infty$ one obtains the density for $d_{\max}$:

$$\rho_{\max}(x) = \sum_{l=2}^{\infty} (-1)^l (2l - 1) c_l \, \exp\{-c_l \, x\}. \tag{11}$$

This theoretical density (see also [4, 7] and very recently [9, 10]) is plotted in figure 2, where it is compared with the density obtained by the simulation of a 100-individual population. As already mentioned, the fact that they coincide so precisely can be considered further evidence that $N = 100$ is sufficiently large that finite size effects are negligible.

Notice that this result is far from being complete, since it gives the distribution of the maximum distance $d_{\max}$ of the support of $q(x)$ but it does not give more general information on the distribution of the density $q(x)$ itself. This problem will be faced in the next section.

Before ending this section we would like to complete the description of the coalescent process by considering the number of offspring of any of the ancestors. Suppose that the total number of individuals in the present generation is N; then, any of the n ancestors has a number $\eta_i^n N$ of offspring (with $i = 1, 2, \ldots, n$ and with $\sum_{i=1}^{n} \eta_i^n = 1$) in the present generation. Successive coalescence reduces the number of ancestors to $n - 1$ and one has that any of them has $\eta_i^{n-1} N$ offspring in the present population ($\sum_{i=1}^{n-1} \eta_i^{n-1} = 1$). These latter numbers are such that $n - 2$ of them are the same as the η_i^n and one is the sum of the

two remaining η_i^n, correspondingly to the pair which have matched in a single ancestor. This rule can be iterated until the number of ancestors reduces to a single one.

Therefore, the coalescent picture is completed by considering this random rule which permits one to obtain the $\eta_1^m, \ldots, \eta_m^m$ from the $\eta_1^n, \ldots, \eta_n^n$ for any $m < n$, the rule being simply that at any step two of the numbers are chosen at random and summed, while the others are left unchanged. Notice that this part of the coalescent process is independent from the random times $\tau_m, \ldots, \tau_n$.

5. Statistics of the random density

We have seen that the time one has to go backward in order that the ancestors of all N individuals of a population reduce to m is $q_{m+1} = \sum_{k=m+1}^{\infty} \tau_l$. In this case, any of the m individuals will be the ancestor of a number $\eta_i^m N$ of individuals ($i = 1, 2, \ldots, m$) with $\sum_{i=1}^{m} \eta_i^m = 1$. In other words, any of the m ancestors will be at the basis of a branch with a number $\eta_i^m N$ of final offspring. Successive coalescence reduces the number of ancestors to $m - 1$, which means the branches of two of the m ancestors are now sub-branches of a single one.

We can now easily see how distances are distributed in a population. At a past time q_2 we have that the last two common ancestors, any of them with a number of final offspring η_1^2 and η_2^2, match in a single ancestor. Therefore, there are $(\eta_1^2 N)(\eta_2^2 N)$ pairs whose distance is q_2 ($q_2 = d_{\max}$) which means that the fraction of pairs whose distance is q_2 is $p_2 = 2\eta_1^2 \eta_2^2$ ($\eta_1^2 + \eta_2^2 = 1$) according to the fact that the total number of pairs is $N(N-1)/2$.

At a past time q_3 we have that two of the last three common ancestors match in a single ancestor. Their final offspring before matching are η_1^3, η_2^3 and η_3^3. One of these three numbers equals η_1^2 or η_2^2 and the sum of the other two (say η_i^3 and η_j^3) equals the remaining one of η_1^2 and η_2^2. The fraction of pairs whose distance is q_3 is $p_3 = 2\eta_i^3 \eta_j^3$.

Then we go on and at time q_m we have that two of the last m common ancestors match in a single one. The numbers of their offspring before matching are $\eta_1^m, \eta_2^m, \ldots, \eta_m^m$. There are $m - 2$ of these numbers which equal $m - 2$ of the $\eta_1^{m-1}, \eta_2^{m-1}, \ldots, \eta_{m-1}^{m-1}$ and two (say η_i^m and η_j^m) whose sum equals the remaining one. The fraction of pairs whose distance is q_m is $p_3 = 2\eta_i^m \eta_j^m$.

It is now quite clear what the probability density $q(x)$ looks like. First, its support is only in the random times q_k with $2 \le k < \infty$ and where $q_k = \sum_{l=k}^{\infty} \tau_l$. Second, the fraction of pairs corresponding to distances $q_1, q_2 \ldots$ is $p_1, p_2, \ldots$ which satisfy $\sum_{k=1}^{\infty} p_k = 1$.

Therefore, the probability density $q(x)$ is

$$q(x) = \sum_{l=2}^{\infty} p_l \, \delta(x - q_l). \tag{12}$$

Now, what we need is to give the statistics of the numbers $q_2, q_3, \ldots$ and $p_2, p_3, \ldots$.

The first part of this programme is simple. In fact, since the probability for the sequence $\tau_2, \tau_3, \ldots$ is $\prod_{k=2}^{\infty} c_k \exp(-c_k t_k)$ and since $q_{l+1} = q_l - \tau_l$, we have that the joint probability for the sequence $q_2, q_3, \ldots$ is

$$\prod_{k=2}^{\infty} c_k \exp[-(k-1)\, q_k], \tag{13}$$

where it is assumed that $q_k \ge q_{k+1}$.

The second part of the programme is a little more difficult. First we stress that $\tau_2, \tau_3, \ldots$ are independent from the sizes η_i^j and, therefore, the random sequence $p_2, p_3, \ldots$ is independent from the sequence $q_2, q_3, \ldots$.

To obtain the statistics for $p_2, p_3, \ldots$ we have to consider the coalescence rule for the η_i^j described at the end of the previous section. According to that, one has the conditional probability $p(\eta_1^{m-1}, \ldots, \eta_{m-1}^{m-1} \,|\, \eta_1^m, \ldots, \eta_m^m)$ which is constant whenever the rule is satisfied and vanishes elsewhere.

Assume that the limit $N \to \infty$ holds; in this case the numbers $\eta_1^m, \ldots \eta_m^m$ may assume any real value on $\sum_{i=1}^m \eta_i^m = 1$. Also assume that the probability density $q(\eta_1^m, \ldots, \eta_m^m)$ is constant on $\sum_{i=1}^m \eta_i^m = 1$ and it vanishes elsewhere. Then, the probability density $q(\eta_1^{m-1}, \ldots, \eta_{m-1}^{m-1})$ is also constant on $\sum_{i=1}^{m-1} \eta_i^{m-1} = 1$ and vanishing elsewhere. This property can be easily verified using the above-described conditional probability density (see also [11]).

Therefore, if the probability density $q(\eta_1^n, \ldots, \eta_n^n)$ is constant for a given n, then it is constant for any $m \leq n$ and the process rule can be easily reversed. In other words, the conditional probability $p(\eta_1^m, \ldots, \eta_m^m \,|\, \eta_1^{m-1}, \ldots, \eta_{m-1}^{m-1})$ can be computed from $p(\eta_1^{m-1}, \ldots, \eta_{m-1}^{m-1} \,|\, \eta_1^m, \ldots, \eta_m^m)$ and $q(\eta_1^m, \ldots, \eta_m^m)$. The only point which needs some care is showing that for infinite N the density $q(\eta_1^n, \ldots, \eta_l^n)$ is, indeed, constant for a given n. This task is accomplished in appendix B.

Using the above results we obtain with a simple calculation that the conditional density $p(\eta_1^m, \ldots, \eta_m^m \,|\, \eta_1^{m-1}, \ldots, \eta_{m-1}^{m-1})$ corresponds to the following reversed rule: one chooses at random $1 \leq i \leq m-1$ with probability η_i^{m-1} and cuts η_i^{m-1} into two segments $\chi \eta_i^{m-1}$ and $(1-\chi)\eta_i^{m-1}$ with χ uniformly distributed between 0 and 1. Then one has that two of the $\eta_1^m, \ldots, \eta_m^m$ are $\chi \eta_i^{m-1}$ and $(1-\chi)\eta_i^{m-1}$ while the other $m-2$ equal the remaining $m-2$ of the $\eta_1^{m-1}, \ldots, \eta_{m-1}^{m-1}$.

There is a useful picture that briefly describes the rule. Consider a square with a unitary surface. Choose a point x_2 with uniform distribution between 0 and 1. Put it on the basis of the square; then it will cut the unitary segments into two parts which can be identified with η_1^2 and η_2^2. Therefore, the shaded area in figure 3(a) is p_2. Then choose a second point x_3 with uniform distribution between 0 and 1. Put it on the basis and it will be in one of the two previously created segments with probability proportional to their size. Furthermore, the cut in the chosen segment will be uniformly distributed on it. Then, p_3 will be the darker shaded area of figure 3(b). Then choose a third point x_4 with uniform distribution between 0 and 1. Put it on the basis of the square and it will be in one of the three previously created segments with probability proportional to the their size. Furthermore, the cut in the chosen segment will be uniformly distributed on it. Then p_4 will be the darkest shaded area of figure 3(c). Then you can go on and the whole square will be shaded when the operation is repeated an infinite number of times.

In conclusion, we have the complete rule for constructing $q(x)$ since we have the joint probability for $q_2, q_3, \ldots$ and we have the simple rule exemplified in figure 3 for the joint probability for $p_2, p_3, \ldots$.

Indeed, we are not able to find this second joint probability density explicitly and, at this stage, the result is little more than transforming a complicated random dynamics (1) into a simpler random rule of repeated fractioning.

Before ending this section we would like to make some comments. Notice that the average value of $d_{\max}$ ($d_{\max}$ is q_2) is $\langle d_{\max} \rangle = 2$, which means that a population has
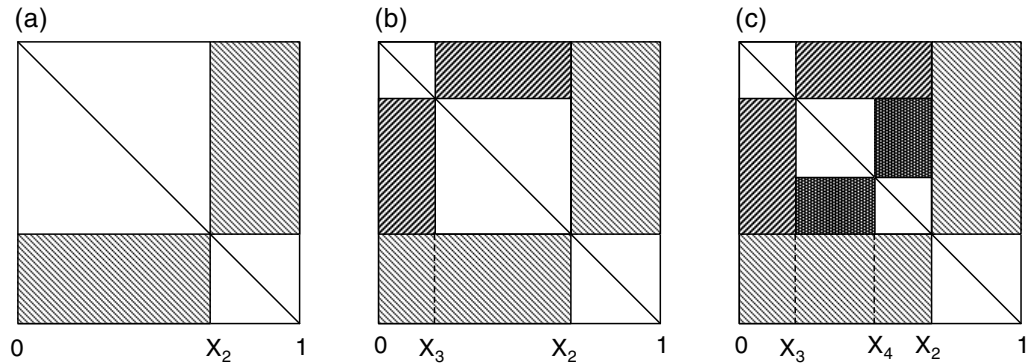
Figure 3. The point x_2 is chosen with uniform distribution on $[0, 1]$; then the shaded area in (a) is p_2. The point x_3 is also chosen with uniform distribution on $[0, 1]$; then p_3 is the darker shaded area in (b). The point x_4 is also chosen with uniform distribution on $[0, 1]$; then p_4 will be the darkest shaded area in (c). The whole square will be shaded when the operation is repeated an infinite number of times, corresponding to the fact that $\sum_{i=2}^{\infty} p_i = 1$.

a common ancestor at a past time which corresponds on average to $2N$ generations. On the other hand, the time for the number of ancestors to reduce to two is q_3 with $\langle q_3 \rangle = 1$. Therefore, the number of generations one has to step backward in order that the ancestors reduce to a pair is N on average and, then, it is necessary to step backward N more generations on average before the ancestors reduce to a single one. This also means that for any realization of the process, the density $q(x)$ has an isolated Dirac delta corresponding to the maximum distance while all the remaining support is concentrated in a segment whose size is, on average, one half of the maximum distance.

In other terms, any population naturally splits into two subpopulations which are the descendants of two different ancestors. All the distances between pairs of individuals from the two different subpopulations coincide with the maximum distance d_{max} whose average is 2; in contrast, the distances inside the two subpopulations are on average $1/2$. In fact, the maximum distance and the mean distance can be related by $d = p_2 d_{max} + (1 - p_2) d_r$ where p_2 is the fraction of pairs with maximum distance and d_r is the average distance of pairs whose distance is not the maximum (pairs inside the two subpopulations). Since $p_2 = 2 x_2 (1 - x_2)$ and x_2 is uniformly distributed between 0 and 1, we take the average and obtain $\langle d_r \rangle = 1/2 = \langle d_{max} \rangle/4$. Therefore, d_r can be several times smaller than d_{max}. These considerations will find motivation in the final discussion.

6. Dynamics

The dynamics of the model is in principle very complicated, since one should be able to describe the time evolution of the density $q(x)$. As a more reachable goal one could try to describe the time evolution of the maximum distance and of the mean distance in a population. The behaviour of these quantities is shown in figure 4 where we plot the maximum distance and the mean distance of the individuals of a single population as a function of time. The two distances result from the dynamics of a population of $N = 500$ individuals generated for 5000 generations which correspond to a time lag of 10. Notice
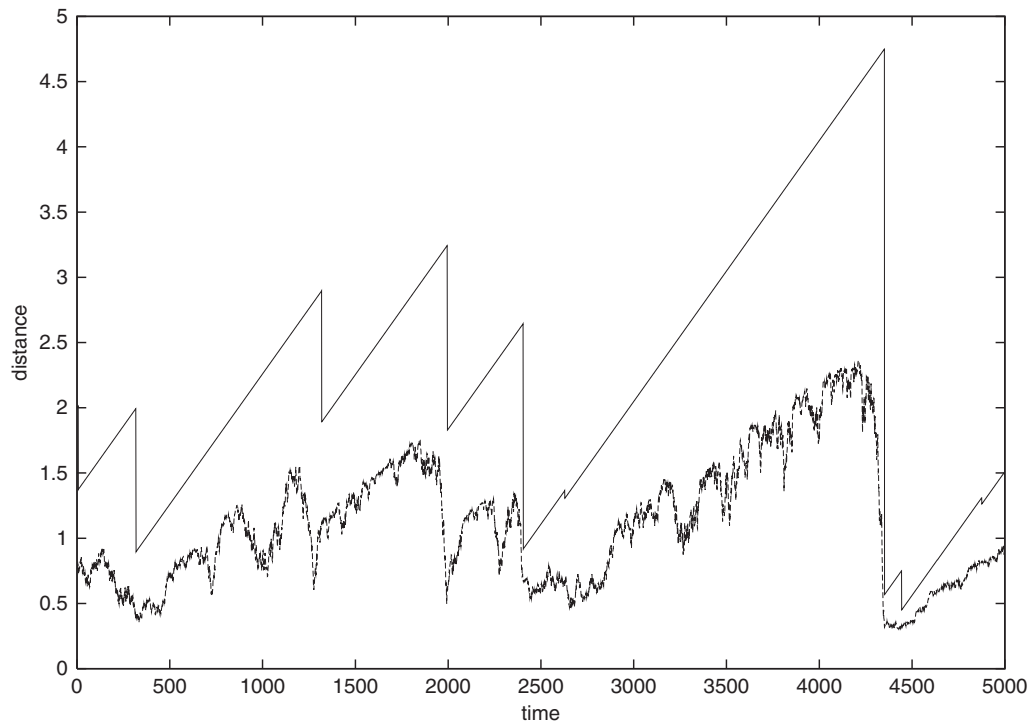
Figure 4. Maximum distance (full line) and mean distance (slashed line) of the individuals of a single population as a function of the progressive generation number. The two distances result from the dynamics of a population of $N = 500$ individuals generated for 5000 generations (time is number of generations divided by N). The size of the population is sufficiently large to destroy finite size effects. Notice that both distances are subject to abrupt negative variations due to the extinction of large subpopulations.

that both distances are subject to abrupt negative variations due to the extinction of large subpopulations. In particular the maximum distance increases constantly until it has a large negative jump due to the extinction of one of the two subpopulations composed of the offspring of the oldest two ancestors. At that point, the ancestor of the surviving subpopulation becomes the last common ancestor of all individuals and the maximum distance is reduced consequently.

The full line in figure 4 gives the maximum distances at all times, while the maximum distances at the times of jumps correspond to its relative maxima. Furthermore, the jump sizes are the differences between relative maxima and subsequent relative minima of the same full line.

How are jumps and relative maxima distributed? In order to compute the densities of these two quantities we have generated a dynamics for a population of 100 individuals for 10^7 generations corresponding to about of 10^5 relative maxima. Both densities are plotted in figure 5.

The probability density for the sizes of jumps, shown in figure 5, is compatible with $\exp(-x)$ which is quite surprising. In fact, it is true that the density of distances between the last two ancestors is $\exp(-x)$, but this is true on average with respect to a generic time, and not necessarily at the times of jumps. Even more surprising is that the empirical
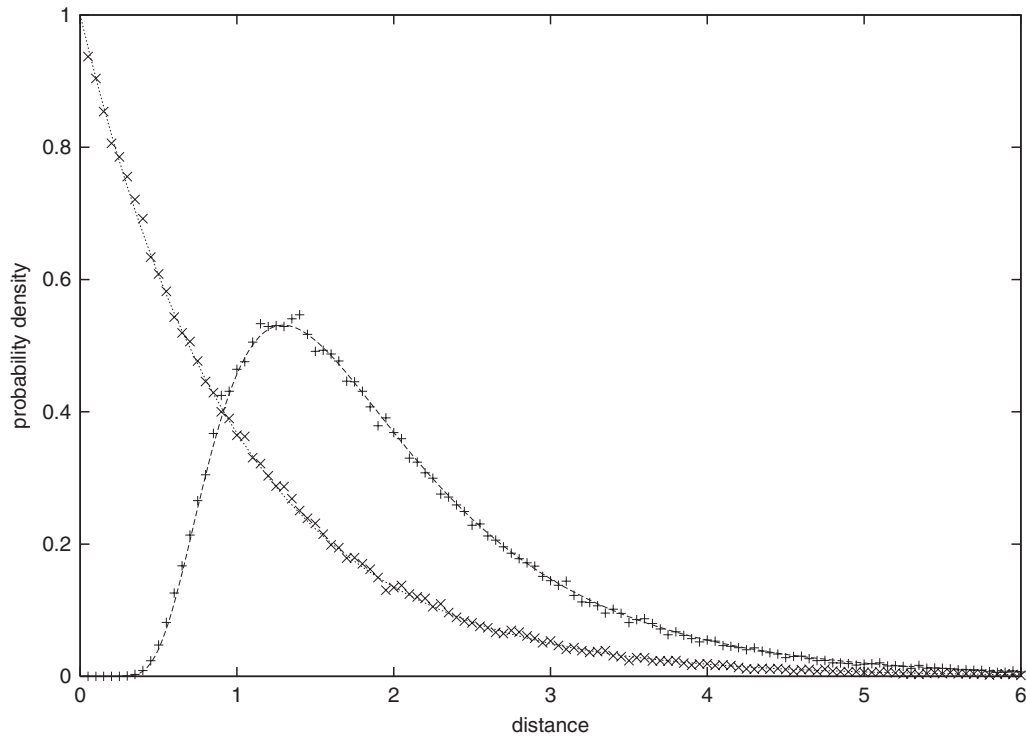
Figure 5. In this picture we have the probability density of the maximum distance (+) at the times of jumps and the probability density for the sizes of jumps (×). The maximum distances at the times of jumps correspond to the relative maxima of the maximum distance process (see figure 4). These densities (100 individuals) are computed from a sample of 10^5 maxima. Comparison with the full line (the theoretical probability density for the maximum distance) shows that the statistical properties of the maximum distance and of the maximum distance at the times of jumps are the same. The probability density for the sizes of jumps (×) is compared with $\exp(-x)$.

density of the maximum distance at the times of jumps coincides (figure 5) with the theoretical density (11). The latter, in fact, gives the statistics of the maximum density at a generic time. In other words, the former is the density of the relative maxima of the full line in figure 4, while the latter is the density of all the points of the same full line.

Finally, we will find the statistics for the lags between jumps. Again, in order to compute this density we have generated a dynamics (1) for a population of 100 individuals for 10^7 generations corresponding to about 10^5 lags. The result, as shown in figure 6, is that lags between jumps are exponentially distributed according to $\exp(-x)$.

In order to understand this behaviour it is sufficient to consider that the times of jumps are when one of the two subpopulations corresponding to the two more recent ancestors of all individuals are extinguished. Assume that at a given time t the number of individuals belonging to the two subpopulations is yN and $(1-y)N$; then at the next generation (at time $t + 1/N$) these numbers are zN and $(1-z)N$. Assuming the Wright–Fisher rule, we have that the probability density for z given y is

$$\rho(z \mid y) = \binom{N}{Nz} y^{Nz} (1-y)^{N(1-z)}, \tag{14}$$
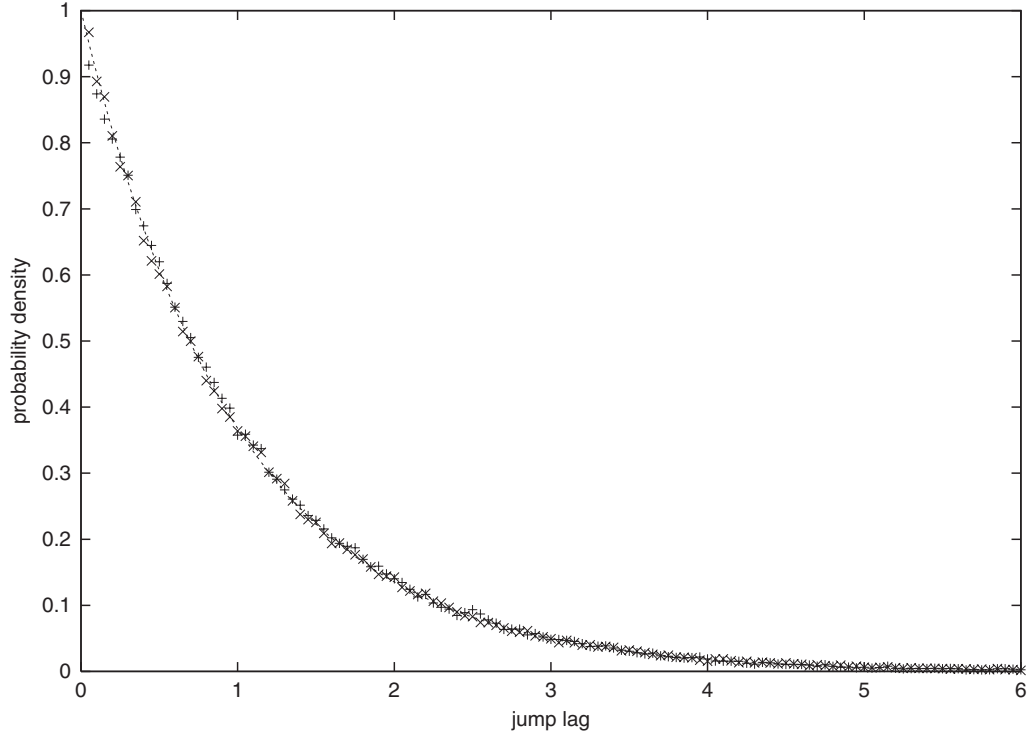
Figure 6. Probability density for the lag between jumps computed from the generated dynamics of 100 individuals ($\times$) and the probability density of lags between jumps computed from the simulation of the exit time for the Wiener process ($+$). The two densities are both obtained from a sample of 10^5 lags and both coincide with the exponential density $\exp(-x)$.

which in particular implies the following two conditional expectations for z and z^2 given y:

$$\langle (z \mid y) \rangle = y \qquad \langle (z^2 \mid y) \rangle = y^2 + \frac{y(1-y)}{N}. \tag{15}$$

It is now simple to construct the diffusion limit of (14). In fact, if we write $x(t + (1/N)) = z$ and $x(t) = y$ we have $\langle x(t + (1/N)) - x(t) \rangle = 0$ and $\langle (x(t + (1/N)) - x(t))^2 \rangle = x(t)(1 - x(t))/N$ which can be written in the continuous time limit as

$$\mathrm{d}x(t) = \sqrt{x(t)(1 - x(t))}\,\mathrm{d}w(t), \tag{16}$$

where $w(t)$ is the Brownian motion (see also [8]).

All we need to do now for computing the statistics of the lags between extinctions (which are the lags between jumps) is to compute the statistics of the hitting times for this process at the frontier $z = 0, 1$. After the process reaches the frontier a new process starts at a point which is uniformly distributed between 0 and 1. This choice depends on the known fact that the two main branches of the surviving subpopulation have a size that is uniformly distributed.

Indeed, the statistics is simply exponential. In order to show this, we have simulated the above equation for a time sufficient to have 10^5 extinctions (hitting times). The

resulting probability density is shown in figure 6 where we have also plotted the density resulting from (1).

7. Discussion

Before discussing open problems concerning the model, we would like to comment on the eventual relevance of its phenomenology for biological applications. Our examples concern the use of mtDNA in recent palaeoanthropological studies. What makes mtDNA interesting is that it is inherited only from the mother and it reproduces asexually, unlike nuclear DNA. In this sense, the mtDNA for a given species should be considered as a haploid population and the results in this paper should apply to it. Furthermore, assuming that mtDNA mutates at a constant rate, the number of differences in mtDNA between two individuals is a measure of their genealogical distance in maternal lineage.

Let us discuss the first example. In the years from 1997 to 2000 some mtDNA from three different specimens of Neanderthals was extracted [15, 16] and short strands of the hypervariable region (HVR1 and HVR2) were amplified using a polymerase chain reaction (PCR).

Two different mtDNA sequences were extracted from the first specimen. For the first sequence, modern humans differed from each other in 8.0 ± 3.1 positions, while Neanderthals differed in 27.0 ± 2.2 positions from modern humans. For the second mtDNA sequence, modern humans differed from each other in 10.9 ± 5.1 positions and the Neanderthals differed in 35.3 ± 2.3 positions from modern humans. The mtDNA sequence of the second Neanderthal was compared with a particular modern human sequence, known as the reference sequence. The difference from the reference modern human sequence was in 22 positions (27 for the first Neanderthal) while the two Neanderthals differed from each other in 12 positions. Sequencing of a third Neanderthal mtDNA confirmed the previous result, since the difference from modern humans was in 34.9 ± 2.4 positions.

The conclusion was that, given the above ranges of differences, Neanderthal mtDNA is statistically different from modern human mtDNA.

But our point of view is different. Let us summarize the results as follows: the Neanderthal/modern human distance is about three times the modern human/modern human distance and about two times the Neanderthal/Neanderthal distance. Consider the situation as it was 40 thousands years ago, when modern humans and Neanderthals coexisted (as well as Erectus and Florensis). If mankind was a single large interbreeding population, we would have a distribution of mtDNA distances similar to that in figure 1. In this case, distances inside some subpopulations could easily have one third or one half of the distance between different subpopulations. This is confirmed by the discussion at the end of section 5, where we find that distances between individuals from the two main subpopulations are, on average, four times larger than distances between individuals inside the two subpopulations. Therefore, measured distances are as expected. To be clear, we do not conclude here that humanity was a single large interbreeding population (multiregional hypothesis); we only claim that the mtDNA argument cannot be used to support the opposite conclusion.

Let us continue with a second example. A fossil of a modern human 60 000 years old (older then the three Neanderthal fossils) was discovered in 1974 in the dry bed of

lake Mungo in Australia. Recently, some sequences of his mtDNA were extracted from fragments of his skeleton [1] and it was observed that Mungo mtDNA is of a type which cannot be found in modern humans including modern aborigines. The conclusion was that Mungo man belongs to a lineage diverging before the most recent common ancestor of contemporary humans, disproving the 'Out of Africa' hypothesis. In this case also, the argument is doubtful; in fact, rapid extinctions of mtDNA subpopulations at all scales are clearly evident in figure 4. Therefore, it is absolutely possible that some mdDNA subpopulations disappeared after modern humans came out of Africa.

In the first example the mtDNA argument was used to disprove the multiregional hypothesis and in the second it was to disprove the 'Out of Africa' hypothesis. We think, however, that in both case a conclusion is impossible. Perhaps a larger sample would improve the situation. In fact, the Neanderthal/modern human distance is the time from their separation which, in the case of impossible interbreeding (different species), should be the same for all pairs. Nevertheless, if interbreeding was possible, some Neanderthals should share the mdDNA of modern humans. In this case, mtDNA subpopulations and geographical subpopulations would not coincide exactly and in larger samples of Neanderthals one could by lucky enough to find evidence for this (sample, up to now, is just of three individuals).

In contrast, as regards modern human/modern human distances, larger samples should not help. In fact, the maximum distance $d_{\max}$ in the modern human population (time from mitochondrial Eve) should be the distance of a fraction $p_2 = 2x_2(1 - x_2)$ of all possible pairs of modern humans. A sample of B individuals can be used to find the correct value of $d_{\max}$ with probability $1 - (x_2)^B - (1 - x_2)^B$. Since we deal with modern humans, B is always sufficiently large to make the probability very close to 1. Only if x_2 is extremely close to 0 or to 1 (which is unlikely since x_2 is uniformly distributed between 0 and 1) could the size of B be relevant.

In order to set up the debate, one should consider nuclear DNA. In fact, sexually reproducing nuclear DNA has a completely different statistics [5, 6, 20] and in large populations the distance for almost all pairs of individuals coincides with the average value [20]. Therefore, strong differences in distances would necessarily imply the existence of different species.

We would like to conclude with a list of open problems. First of all, we would like to compute the probability density for the mean distance d. We are in principle able to 'painfully' compute all moments of the random variable d following the calculations in section 3, but we are not able at the moment to give an explicit expression for its probability density. More importantly, we would like to find the explicit joint probability for $p_2, p_3, \ldots$. Notice that we are able to give this probability only indirectly using the procedure described in figure 3. Finally, we would like to characterize the time behaviour of the maximum distance, which means finding the process for $d_{\max}$, for which we have a realization in figure 4.

Acknowledgments

Appendix A

We give here a very simple derivation of an explicit representation for $\rho_n(x)$. We first show that

$$\rho_n(x) = \left(\prod_{s=2}^{n} c_s\right) \sum_{l=2}^{n} \left(\prod_{s=2; s\neq l}^{n} \frac{1}{c_s - c_l}\right) \exp(-c_l x). \tag{A.1}$$

It can be directly verified that the above equation holds for $n = 3$ according to (7) and (8). Furthermore, assuming that it holds for a given $n - 1$, from (7) and (8) we obtain

$$\rho_n(x) = \left(\prod_{s=2}^{n} c_s\right) \sum_{l=2}^{n-1} \left(\prod_{s=2; s\neq l}^{n} \frac{1}{c_s - c_l}\right) \exp(-c_l x)$$

$$- \left(\prod_{s=2}^{n} c_s\right) \sum_{l=2}^{n-1} \left(\prod_{s=2; s\neq l}^{n} \frac{1}{c_s - c_l}\right) \exp(-c_n x). \tag{A.2}$$

If we compare (A.1) and (A.2) we see that they coincide provided that

$$- \left(\prod_{s=2}^{n} c_s\right) \left(\prod_{s=2; s\neq n}^{n} \frac{1}{c_s - c_n}\right) \exp(-c_n x) - \left(\prod_{s=2}^{n} c_s\right) \sum_{l=2}^{n-1} \left(\prod_{s=2; s\neq l}^{n} \frac{1}{c_s - c_l}\right) \exp(-c_n x) = 0, \tag{A.3}$$

which holds assuming that

$$\sum_{l=2}^{n} \left(\prod_{s=2; s\neq l}^{n} \frac{1}{c_s - c_l}\right) = 0. \tag{A.4}$$

Therefore, all we need to prove the preliminary representation (A.1) is that (A.4) holds. To reach this goal let us define the Lagrange polynomial

$$Q(x) = \sum_{l=2}^{n} c_l \left(\prod_{s=2; s\neq l}^{n} \frac{c_s - x}{c_s - c_l}\right). \tag{A.5}$$

We can immediately verify that for every l such that $2 \leq l \leq n$, one has $Q(c_l) = c_l$. Since the degree of the polynomial is at most $n - 2$ and since it crosses the above $n - 1$ points, it is necessarily the case that $Q(x) = x$ and in particular $Q(0) = 0$ Then, since by definition

$$Q(0) = \sum_{l=2}^{n} c_l \left(\prod_{s=2; s\neq l}^{n} \frac{c_s}{c_s - c_l}\right) = \left(\prod_{s=2}^{n} c_s\right) \sum_{l=2}^{n} \left(\prod_{s=2; s\neq l}^{n} \frac{1}{c_s - c_l}\right), \tag{A.6}$$

equation (A.4) holds and (A.1) is demonstrated.

Furthermore, by a simple calculation one can show that

$$\left(\prod_{s=2}^{n} c_s\right) \sum_{l=2}^{n} \left(\prod_{s=2; s\neq l}^{n} \frac{1}{c_s - c_l}\right) = (-1)^l (2l - 1) c_l \prod_{s=1}^{l-1} \frac{n - s}{n + s}, \tag{A.7}$$

and finally we have the simple sum representation for the coalescent density distribution:

$$\rho_n(x) = \sum_{l=2}^{n} (-1)^l (2l-1) c_l \left(\prod_{s=1}^{l-1} \frac{n-s}{n+s} \right) \exp\{-c_l\, x\}. \tag{A.8}$$

Appendix B

We show here that the probability density $q(\eta_1^n, \ldots, \eta_l^n)$ is constant for a given n when the limit of large N is taken.

At a given time in the past, the number of ancestors of all N individuals of a population is l. At an intermediate time, always in the past, the number of ancestors is $k \geq l$. This means that any of the l individuals is an ancestor of one or more of the k individuals, i.e., any of the l branches has one or more sub-branches. Let us call this (integer) number of sub-branches for individual $i r_i^l$; then, $\sum_{i=1}^{l} r_i^l = k$ with $r_i^l \geq 1$. Let us call the ensemble of $r_1^l, \ldots r_l^l$ such that $\sum_{i=1}^{l} r_i^l = k$, and $r_i^l \geq 1$, Γ_l.

Let us define $f_l(r_1^l, \ldots r_l^l)$ as the probability of $r_1^l, \ldots r_l^l$. We first show that this probability is constant on Γ_l for any $l \leq k$.

Assume that $f_{l+1}(r_1^{l+1}, \ldots r_{l+1}^{l+1})$ is constant on Γ_{l+1}. Coalescence implies that one of $r_1^l, \ldots r_l^l$ equals the sum of two of the $r_1^{l+1}, \ldots r_{l+1}^{l+1}$ randomly chosen, while the remaining $l-1$ coincide. Then, according to this rule, $f_l(r_1^l, \ldots r_l^l)$ is constant on Γ_l. To have a proof, it is now sufficient to note that $f_k(r_1^k, \ldots r_k^k)$ is constant on Γ_k. In fact, all r_i^k must equal one, i.e. $f_k(1, \ldots 1) = 1$, while it vanishes elsewhere.

Now, let us recall that the numbers of offspring of the k ancestors are $\eta_1^k N, \ldots, \eta_k^k N$ with $\sum_{i=1}^{k} \eta_i^k = 1$. Assume now that at a previous time the number of ancestors is $n < k$ and assume that the number of sub-branches of any of them is $r_1^n, \ldots r_n^n$. Then, the numbers η_i^n will be obtained as the sum of the r_i^n of the $\eta_1^k, \ldots, \eta_k^k$ chosen at random.

Now let us recall that $\sum_{i=1}^{n} r_i^n = k$; therefore, large k implies that for almost all possible choices on Γ_n the n numbers r_i^n must be of order k. We can define $r_i^n = \alpha_i^n k$ with $\sum_{i=1}^{n} \alpha_i^n = 1$ and the numbers $0 \leq \alpha_i^n \leq 1$ of order 1. Furthermore, since $\sum_{i=1}^{k} \eta_i^k = 1$, we assume that almost all of the η_i^k are of order $1/k$.

We can now take the limit of large k after taking the limit of large N. Since the η_i^n are the sums of $r_i^n = \alpha_i^n k$ of the $\eta_1^k, \ldots, \eta_k^k$ and since from the definition $\sum_{i=1}^{k} \eta_i^k = 1$ with η_i^k of order $1/k$, one has $\eta_i^n = \lim_{k \to \infty} r_i^n / k = \alpha_i^n$.

Finally, since $f_l(r_1^n, \ldots r_n^n)$ is constant on Γ_n, one has that $q(\eta_1^n, \ldots, \eta_n^n)$ is constant on $\sum_{i=1}^{n} \eta_i^n = 1$.

References

[1] Adcock G, Dennis E, Easteal S, Huttley G, Jermin L, Peacock W and Thorne A, *Mitochondrial DNA sequences in ancient Australians: implications for modern human origins*, 2001 *Proc. Nat. Acad. Sci.* **98** 537

[2] Aldous D, *Deterministic and stochastic models for coalescence (aggregation, coagulation): a review of mean field theory for probabilists*, 1999 *Bernoulli* **5** 3

[3] Dalal A and Schmutz E, *Compositions of random functions on a finite set*, 2002 *Electron. J. Comb.* **9** R26

[4] Derrida B and Bessis D, *Statistical properties of valleys in the annealed random map model*, 1999 *J. Phys. A: Math. Gen.* **21** L509

[5] Derrida B and Jung-Muller B, *The genealogical tree of a chromosome*, 1999 *J. Stat. Phys.* **94** 277
[6] Derrida B, Manrubia S C and Zanette D H, *Statistical properties of genealogical trees*, 1999 *Phys. Rev. Lett.* **82** 1987
[7] Derrida B and Peliti L, *Evolution in a flat fitness landscape*, 1991 *Bull. Math. Biol.* **53** 355
[8] Donnelly P, *Weak convergence to a Markov chain with an entrance boundary: ancestral processes in population genetics*, 1991 *Ann. Probab.* **19** 1102
[9] Fill J A, *On compositions of random functions on a finite set, Unpublished paper*
[10] Goh W M Y, Hitczenko P and Schmutz E, *Iterating random functions on a finite set*, 2002 *Preprint* math.CO/0207276v2
[11] Kingman J F C, *The coalescent*, 1982 *Stoch. Process. Appl.* **13** 235
[12] Kingman J F C, *On the genealogy of large populations*, 1982 *J. Appl. Probab.* **19A** 27
[13] Kingman J F C, *Exchangeability and the evolution of large populations*, 1982 *Exchangeability in Probability and Statistics* (Amsterdam: North-Holland) pp 97–112
[14] Kingman J F C, *Mathematics of genetic diversity*, 1980 *CBMS-NSF Regional Conf. Series in Applied Mathematics* vol 34 (Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM)) ISBN: 0-89871-166-5
[15] Krings M, Capelli C, Tschentscher F, Geisert H, Meyer S, von Haeseler A, Grossshmidt K, Possnert G, Paunovic M and Pbo S, *A view of Neandertal genetic diversity*, 2000 *Nat. Genet.* **26** 144
[16] Krings M, Stone A, Schmitz R W, Krainitzki H, Stoneking M and Pbo S, *Neandertal DNA sequences and the origin of modern humans*, 1997 *Cell* **90** 19
[17] Möhle M, *Total variation distances and rates of convergence for ancestral coalescent processes in exchangeable population models*, 2000 *Adv. Appl. Probab.* **32** 983
[18] Möhle M, *Weak convergence to the coalescent in neutral population models*, 1999 *J. Appl. Probab.* **36** 446
[19] Serva M, *Lack of self averaging in family trees*, 2004 *Physica A* **332** 387
[20] Serva M and Peliti L, *A statistical model of an evolving population with sexual reproduction*, 1991 *J. Phys. A: Math. Gen.* **24** L705
[21] Tavare S, *Line-of-descent and genealogical processes and their applications in population genetics models*, 1984 *Theor. Population Biol.* **26** 119
[22] Watterson G A, *Lines of descent and the coalescent*, 1984 *Theor. Population Biol.* **26** 256
[23] Zhang Y-C, Serva M and Policarpov M, *Diffusion reproduction processes*, 1990 *J. Stat. Phys.* **58** 849