Markov Processes Relat. Fields **15**, 417–440 (2009)

Markov Processes and **Related Fields**
© Polymat, Moscow 2009

Family Trees: Languages and Genetics

F. Petroni , L. Prignano and M. Serva

Dipartimento di Matematica, Università dell'Aquila, I-67010 L'Aquila, Italy

Abstract. We consider a large size population which evolves according to neutral haploid reproduction. The genealogical tree is very complex and genealogical distances are distributed according to a probability density which remains random in the limit of a large population. This density which varies for different populations, and varies for the same population at different times, has a distribution that we find out.

The evolution of languages closely resembles the evolution of haploid organisms or mtDNA. This similarity allows for the construction of languages trees. The key point is the definition of a distance between pairs of languages. Here we use a renormalized Levenshtein distance among words with the same meaning and we average on all the words contained in a list. Assuming a constant rate of mutation, these lexical distances are logarithmically proportional, in average, to genealogical distances.

The relation between lexical and genealogical distances is then further investigated in order to take into account the intrinsic randomness associated with the lexical evolution. We test our method by constructing the trees of the Indo-European and Austronesian groups.

Keywords: random processes, fluctuation phenomena, dynamics of social systems, dynamics of evolution, networks and genealogical trees

AMS Subject Classification: 92D25, 60K35, 60F10, 60J70

1. Introduction

Haploid reproduction implies that any individual has a single parent in the previous generation. Since some of the individuals may have the same parent, the number of ancestors of the present population decreases going backwards in time until a complete coalescence to a single ancestor. The genealogical distance between two individuals is simply the time from the last common ancestor. One would expect that in the limit of infinite population size, some quantities would reach a thermodynamic deterministic value. For example this could be the case for the probability density of genealogical distances in a single population. On the contrary, this quantity is random even in the thermodynamic limit and, therefore, it varies for different populations or, at different times, for the same population. The discovery of this non self-averaging behavior is due to the pioneering work of Derrida, Bessis and Peliti [4, 7].

We consider a very general model of a population of constant size N whose generations are not overlapping in time. Any generation is replaced by a new one and any individual has a single parent. Stochastic rules assigning the number of offsprings to any individual can be chosen in many ways. In fact, for a large size population, results do not depend on the details of those rules, the only requirement is that the probability of having the same parent for two individuals is of order $1/N$ for large N.

To be clearer we make two examples of stochastic dynamics that satisfy this assumption. First rule: at any generation one half of the individuals (chosen randomly) has no offsprings while the remaining part has two (see [34]). With this rule the probability of having the same parent is $1/(N-1)$. The second rule (Wright−Fisher) is that any individual in the new generation chooses one parent at random from the previous one, independently from the choice of the others. With this rule, that we will choose for simulations, the probability of having the same parent for two individuals is exactly $1/N$.

The relevant quantity that we compute is the distribution of the random probability density of pair distances in a single large population. We also show how to reconstruct the genealogical tree of a group of individuals. Sections 2, 3, 4 and 5 are devoted to this part of our research.

Languages evolve in time according to a process in which reproduction and extinction are both possible. Reproduction, because a single language may have more then one offspring, for example Latin, and extinction when the number of speakers becomes too small, for example many Australian languages. This is very similar to haploid evolution for asexual organisms. This similarity allows us to use some of the ideas developed in the first part of the paper in order to verify hypotheses concerning the relationship among languages and to reconstruct their family tree. The key point is the definition of a distance among pairs of languages in analogy with the genetic distance among pairs of organisms. Distances can be evaluated comparing grammar and/or vocabulary but while it is difficult, if not impossible, to quantify grammar distance, it is possible to measure a distance from vocabulary differences. The method used by glottochronology computes distances from the percentage of shared “cognates” which are words with a common historical origin. Recent examples are the studies of Gray and Atkinson [12] and Gray and Jordan [13]. The weak point of this method is that subjective judgment plays a relevant role. Here we define the distance of two languages by considering a renormalized edit distance among words with the same meaning and averaging on the two hundred words contained in a Swadesh list [31]. In our approach the vocabulary of a language is the analogue of DNA for organisms. The advantage is that we avoid subjectivity and our results can be replicated by other scholars assuming that the database is the same. We apply our method to the Indo-European and the Austronesian group considering, in both cases, fifty different languages. The two trees we obtain are for many aspects similar to those found by glottocronologists with some important differences concerning the position of few languages. Sections 6, 7 and 8 are devoted to this part of our research.

A last problem that we have to face is that the genetic distances that we measure for languages are not exactly the genealogical distances defined for the coalescent process. In fact, we compute a distance from the lexical differences and not from the historical time separation which is unknown in most of the cases. Genetic distances are proportional to genealogical ones only on average, due to the inner randomness of lexical mutations. This is the same problem that scholars face in biology when they measure the time distance from the last common ancestor measuring genetic differences between two species. We try to fill this gap introducing randomness in the coalescent process which allows comparison between the statistics of genealogical and genetic distances. In particular, we are able to quantify the error that we may make when we reconstruct family trees by genetic distances in spite of the unknown historical records of genealogical ones. Sections 9 and 10 are devoted to this part of our research while Section 11 contains our conclusions.

2. Frequency of pair distances

The distance between two given individuals is the number of generations from the common ancestor. For large N distances are proportional to N, it is then useful to re-scale them dividing by N. In a population of size N we have to specify $N(N-1)/2$ distances which correspond to the entries of an upper triangular matrix.

Let us define $d(\alpha,\beta)$ as the rescaled genealogical distance between individuals α and β in a population of size N. For two distinct individuals α and β in the same generation one has

$$d(\alpha,\beta) = d(g(\alpha), g(\beta)) + \frac{1}{N} \quad , \tag{2.1}$$

where $g(\alpha)$ and $g(\beta)$ are the parents of α and β respectively. Parents are chosen among all possible ones with equal probability $1/N$ and, therefore, $g(\alpha)$ and $g(\beta)$ coincide with probability $1/N$. In this case the distance $d(g(\alpha), g(\beta))$ vanishes. On the contrary, the parents of α and β are distinct individuals α' and β' 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 parents distance.

Briefly, $d(\alpha, \beta) = 1/N$ with probability $1/N$ and $d(\alpha, \beta) = d(\alpha', \beta') + 1/N$ with probability $(N-1)/N$.

This equation can be iterated for any of the possible $N(N-1)/2$ initial pairs α and β and the iteration stops when there is a coincidence of parents. In this way, all the entries $d(\alpha, \beta)$ of the upper triangular matrix of distances can be calculated.

Distances assume values which are multiple of $1/N$, and we may define the frequency $f(x)$ as the number of pairs whose distance is x divided by the total number $N(N-1)/2$ of pairs, i.e., the fraction of pairs whose distance is x. Alternatively, we can introduce the density $q(x)$ defined as

$$q(x) = \frac{2}{N(N-1)} \sum_{\alpha > \beta} \delta(x - d(\alpha, \beta)) \ , \tag{2.2}$$

where δ indicates the Dirac delta function. The two quantities are simply related since the integral of $q(x)$ on the interval $[\,x - 1/(2N),\ x + 1/(2N)\,]$ gives $f(x)$. The advantage in considering $q(x)$ is that in the limit of large N the distance can assume any real positive value but the density $q(x)$ remains well defined.

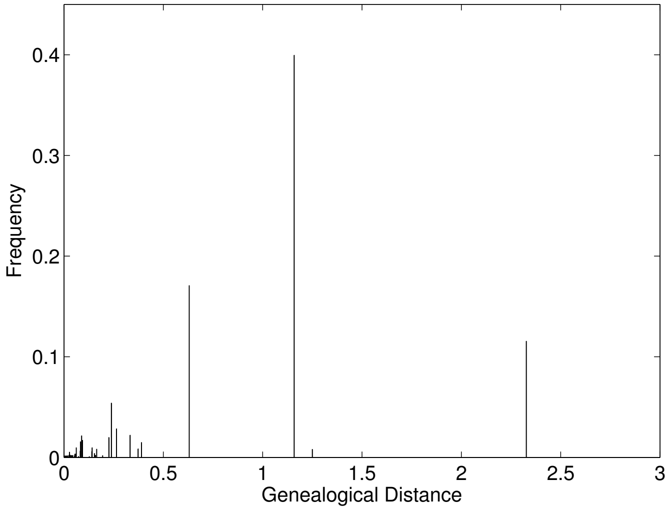

Figure 1. Frequency $f(x)$ of distances in a single population computed for a population of 700 individuals. Most of the distances assume few values corresponding to the distances between major subpopulations.

The dynamics can be easily simulated. One may assign initial arbitrary conditions for all distances. Then, distances of all the following generations are obtained from (2.1). The simulation has to last for a time much larger than N in order to be independent from the initial conditions and, finally, $f(x)$ can be calculated. The frequency inside a single population of 700 individuals can be seen in Fig. 1. It is immediately obvious that this frequency is quite wild, due to the fact that individuals naturally cluster in subpopulations. In fact, most of the distances assume few values corresponding to the distances between the major subpopulations.

One could think that this singular behavior would disappear in the thermodynamic limit of large N. On the contrary, not only this singular behavior remains, but one easily realizes that the density $q(x)$ remains random, varying for different populations and varying for the same population at different times.

The random and singular nature of the density in the $N \to \infty$ limit reminds that of the overlap function in mean field spin glasses. In fact, both show similar non self-averaging properties.

In Section 4 we will be able to give the probability distribution of the random probability density $q(x)$, so that the complete specification of the static properties of the model will be achieved.

Hereafter we use 'average' to intend average on many realizations of the population process or, equivalently, by ergodicity, average on the same population at different times. Average will be indicated by $\langle \cdot \rangle$. The average density $\langle q(x) \rangle$ turns out to be simply $\exp(-x)$ [25, 26]. This smooth average density is completely different from a typical sample. To appreciate this fact, see again Fig. 1 where the frequency $f(x)$ is plotted.

3. The coalescent

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 J.F.C. Kingman [14–17] 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 were all different in the 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 the time lag for a first coalescent event is $p_n(t) = c_n \exp(-c_n t)$. Then, it is easy to find the joint probability density

$$\prod_{k=n}^{m+1} p_k(t_k) \tag{3.1}$$

which gives the statistics for successive coalescence time lags $t_n\,, t_{n-1}\,, \ldots, t_{m+1}$ until the number of ancestors reduces to m. This is the core of the celebrated coalescent, which is mostly associated to the name of the probabilist J.F.G. Kingman.

The coalescence random time lag from n individuals to m ancestors, is simply the sum $\sum_{k=m+1}^{n} t_k$. Its distribution is given by the convolution of $n-m$ successive exponentials. In particular, the time density distribution of the time lag T for complete coalescence of n individuals to a single ancestor is

$$\rho_n(T) = \sum_{l=2}^{n} \, (-1)^l \, (2l-1) \, c_l \left(\prod_{s=1}^{l-1} \frac{n-s}{n+s} \right) \exp\{- \, c_l \, T\} \; . \tag{3.2}$$

Then, the density distribution for the time lag T for the coalescence of all the individuals of a large population to a single ancestor is $\rho_{max}(T) = \lim_{n\to\infty} \rho_n(T)$. We can look at this last distribution in another way: ρ_{max} is the distribution of the maximum d_{max} of all possible distances in a large population, i.e.

$$d_{max} \;=\; \max_{\{\alpha,\beta\}} \, d(\alpha,\beta) \;\; , \tag{3.3}$$

In fact, by definition T is equal to d_{max} which, in turn, coincides with the maximum of the support of the $q(x)$. The randomness of $q(x)$ implies that d_{max} is also random. The statistical properties of this quantity, which measures the time from the common ancestor of the whole population, have been widely investigated in a number of papers in the last two decades [2,8,14–17,20,21,32,33], and more recently [3,10,11,25,26,29]. The dynamics of d_{max} has also been recently studied numerically and analytically in [26,29].

4. Statistics of the random density

The result in the previous section is far from being complete, since it gives the distribution of the maximum distance d_{max} corresponding to the maximum of the support of $q(x)$ but it does not give $q(x)$ itself. Nevertheless, the above line of reasoning allows to compute the probability density which has the form [26]:

$$q(x) = \sum_{l=2}^{\infty} p_l \, \delta(x - q_l) \; . \tag{4.1}$$

Now, what we need is to give the statistics of the random numbers $q_2, q_3, \ldots$ and $p_2, p_3, \ldots$ The first part is simple. In fact, since the probability for the sequence $t_2, t_3, \ldots$ is $\prod_{k=2}^{\infty} c_k \, \exp(- \, c_k \, t_k)$ and since $q_{l+1} = q_l - t_l$ we have that the joint probability for the sequence $q_2, q_3, \ldots$ is

$$\prod_{k=2}^{\infty} c_k \, \exp[- \, (k-1) \, q_k] \;, \tag{4.2}$$

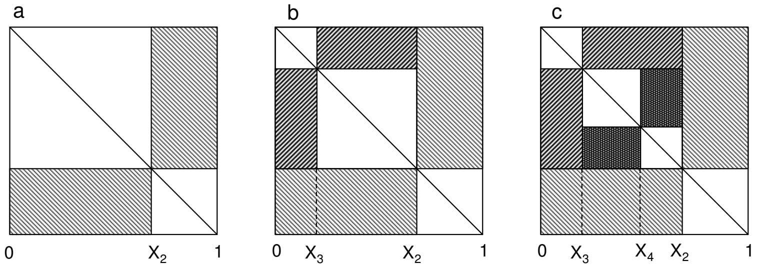

Figure 2. The point x_2 is chosen with uniform distribution on $[0, 1]$ than the shaded area in (a) is p_2. The point x_3 is also chosen with uniform distribution on $[0, 1]$, then, p_3 is the darkest shaded area in (b). The point x_4 is also chosen with uniform distribution on $[0, 1]$, then, p_4 will be the darkest shaded area in (c). The whole square will be shaded when the operation is repeated infinite times corresponding to the fact that $\sum_{i=2}^{\infty} p_i = 1$.

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

The second part concerning the random numbers $p_2, p_3, \ldots$ is a little more difficult and can be found in [26]. There is a useful picture that briefly describes the rule for the distribution of the $p_2, p_3, \ldots$ Consider a square with unitary surface. Choose a point x_2 with uniform distribution between 0 and 1. Put it on the base of the square, then it will cut the unitary segments in two parts, the shaded area in Fig. 2a, then, is p_2. Choose a second point x_3 with uniform distribution between 0 and 1. Put it on the base and it will be in one of the two previously created segments with a probability proportional to their size. Furthermore, the cut in the chosen segment will be uniformly distributed. Then, p_3 will be the darkest shaded area of Fig. 2b. Then, choose a third point x_4 with uniform distribution between 0 and 1. Put it on the base of the square and it will be in one of the three previously created segments with a probability proportional to their size. Furthermore, the cut in the chosen segment will be uniformly distributed. Then p_4 will be the darkest shaded area of Fig. 2c. Then you can continue and the whole square will be shaded when the operation is repeated infinite 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 Fig. 2 for the joint probability for $p_2,\ p_3,\ \ldots$

This also means that for any realization of the process, the density $q(x)$ has an isolated Dirac delta function corresponding to the maximum distance d_{max}

while all the remaining support is concentrated in a segment whose size is, on average, one half of the maximum distance.

5. Trees reconstruction

Iteration of equation (2.1) gives as output the realization of the random $N\,(N-1)/2$ distances $d(\alpha,\beta)$ which are the entries of an upper triangular matrix. This matrix contains all the necessary information for the reconstruction of the family tree of the population. The tree is completely identified by its topology and by the separation time of all branching events. There exist many methods that can be used for this reconstruction, a simple one is the Unweighted Pair Group Method Average (UPGMA). This algorithm works as follows: it first identifies the two individuals with the shortest distance, and put their branching at their separation time. Then, it treats this pair as a new single object whose distance from the other individuals is the average of the distance of its two components. Subsequently, among the new group of objects it identifies the pair with the shortest distance, and so on. In the end, one is left with only two objects which represent the two main branches, whose distance gives the time position of the root of the tree. Then, the time from the last common ancestor of all individuals in the population results fixed.

This method works for any kind of upper triangular matrix representing distances among pairs of individuals, not necessarily originated by the coalescent process. In the coalescent case, nevertheless, the method gives the correct tree reproducing the historical branching events and the correct time separations among them. Notice that, at any time, it chooses two individuals with the shortest distance. Then, it is easy to realize that for the coalescent, the distance of the two individuals from any third one is the same. Therefore, in this case, all UPGMA averages are between pairs with identical distances so that also the resulting new common distances are the same.

6. Method and database for languages

A Swadesh list contains 200 terms which are common to all cultures, concern the basic activities of humans and are more resistant to lexical modifications. The use of Swadesh lists in glottochronology has been popular for half a century. The point of glottochronology is to find the percentage of shared cognates in order to compute the lexical distance between any pair of languages. Cognates are words inferred to have a common historical origin, their identification is often a matter of sensibility and personal knowledge. In fact, the task of counting the number of cognate words in the list is far from being trivial because cognates do not necessarily look similar. Furthermore, results are often biased since it is easier for European or American scholars to find out those cognates belonging to western languages. For instance, the Spanish word *leche* and the Greek word *gala* are cognates. In fact, *leche* comes from the Latin *lac* with genitive form *lactis*, while the genitive form of *gala* is *galactos*. An analogous identification hardly would have been possible for languages, let's say, of Central Africa or Australia.

With our approach we try to avoid this subjectivity. For any language we write down a list of the same 200 words according to the original choice of Swadesh [31], than we compare words with the same meaning belonging to different languages considering only orthographical differences. This may appear reductive since words may look similar by chance, while cognate words may have a completely different orthography, but we will try to convince the reader that indeed this is a simpler, more objective and more efficient choice with respect to the traditional glottochronological approach.

In order to find the lexical distance between pairs of words in different languages we use a modification of the Levenshtein distance (or edit distance) which is defined as the minimum number of operations needed to transform one word into another. An operation is an insertion, deletion, or substitution of a single character. Our definition of distance between two words is taken as the edit distance divided by the number of characters of the longer of the two. With this definition, the distance can take any value between 0 and 1. To understand why we renormalize, let us consider the following case of one substitution between two words: if the compared words are long and the difference between them is given by one substitution they remain very similar; while, if these words are short, let's say two characters, one substitution is enough to make them completely different. Without renormalization, the distance between the words compared in the two examples would be the same, no matter their length. Instead, introducing the normalization factor, in the first case the genetic distance is much smaller than in the second one.

We use distance between pairs of words, as defined above, to construct the matrix of lexical distances. For any pair of languages, the first step is to compute the distance between words with same meaning contained in the Swadesh list. Then, the lexical distance between each languages pair is defined as the average of the distance between all words. As a result we have a number between 0 and 1 which we claim to be the lexical distance between two languages.

The database used here [35] to construct the phylogenetic tree is composed by 50 languages of the Indo-European group and 50 languages of the Austronesian group. The main source for the database for the Indo-European group is the file prepared by Dyen et al. in [9] which contains the Swadesh list of 200 words for 96 languages. Many words are missing in [9] but for our choice of 50 languages we have filled most of the gaps and corrected some errors by finding the words on Swadesh lists and on dictionaries freely available on the web. For the Austronesian group we used as the main source the lists contained in the huge database in [22]. The lists in [22] contain more than 200 words which do not coincide completely with the words in the original Swadesh list [31]. For our choice of 50 Austronesian languages we have retained only the words which are in [9] and are also in the original Swadesh list. The resulting list has many gaps due to missing words in [22] and because of the incomplete overlap of [22] with the original Swadesh list. Also in this case we have filled some of the gaps by finding the words on Swadesh lists available on the web and, in one case (Malagasy), by direct knowledge of the language. For some of the languages in our lists [35] there are still few missing words. When a language has one or more missing words, these are simply not considered in the average that brings to the definition of lexical distance between two languages. This implies that for some pairs of languages, the number of compared words is not 200 but smaller. There is no bias in this procedure, the only effect is that the statistic is slightly reduced.

In the database only the English alphabet is used (26 characters plus space); those languages written in a different alphabet (i.e. Greek etc.) were already transliterated into the English one in [9]. Furthermore, in [22] many additional characters are used which we have eliminated so that also in this case we reduce to the English alphabet plus space. Our database is available at [35].

The result of the analysis described above are two 50×50 upper triangular matrices with the lexical distances of the languages of the two groups. Each matrix contains the 1225 distances among all pairs in a group. Indeed, our method for computing distances is a very simple operation, that does not need any specific linguistic knowledge and requires a minimum of computing time.

7. Time distance between languages

A phylogenetic tree can already be built from one of these matrices whose entries are the lexical distances, but this would only give the topology of the tree whereas the absolute time scale would be missing. In fact, we would like to find quantities which are directly comparable with genealogical distances $d(\alpha, \beta)$ corresponding to time separations.

In order to have this quantitative information, some hypotheses on the time evolution of lexical distances are necessary. We assume that the lexical distance, on one side tends to grow due to random mutations in the vocabulary and, on the other side, may reduce since different words may become more similar by accident or, more likely, by language borrowings.

Therefore, the lexical distance $D(\alpha, \beta)$ between two given languages α and β can be thought to evolve according to the simple differential equation

$$\dot{D}(\alpha, \beta) = a\,(\,1 - D(\alpha, \beta)\,) - b\,D(\alpha, \beta) \tag{7.1}$$

where $\dot{D}$ is the time derivative of D. The parameter a is related to the increase of D due to random permutations, deletions or substitutions of characters (random mutations) while the parameter b considers the possibility that two words become more similar by a "lucky" random mutation or by words borrowing from one language to the other or both from a third. Since a and b are constants, it is implicitly assumed that mutations and borrowings occur at a constant rate.

At time $t = 0$ two languages begin to separate and the lexical distance D is zero. With this initial condition the above equation can be solved and the solution can be inverted. The result is a relation which gives the time separation $t(\alpha, \beta)$ between two languages in terms of their lexical distance $D(\alpha, \beta)$

$$t(\alpha, \beta) = -\varepsilon \ln(1 - \gamma D(\alpha, \beta)) \tag{7.2}$$

The time separation $t(\alpha, \beta)$ is the quantity which we would like to compare to the genealogical distance $d(\alpha, \beta)$. Indeed, the two quantities do not coincide since the process of lexical mutation is not constant in time but random and, therefore, $t(\alpha, \beta)$ equals $d(\alpha, \beta)$ only on average. The problem, obviously, is that it is not possible to perform the average since there is only one single realization of the process of languages differentiation. Hereafter, we will call $t(\alpha, \beta)$ genetic distance.

The values for the parameters $\varepsilon = 1/(a + b)$ and $\gamma = (a + b)/a$ can be fixed experimentally by considering two pairs of languages whose separation time is known. We have chosen a distance of 1600 years between Italian and French and a distance of 1100 years between Icelandic and Norwegian. The resulting values of the parameters are $\varepsilon = 1750$ and $\gamma = 1.09$ corresponding to $\alpha \cong 5 * 10^{-4}$ and $\beta \cong 6 * 10^{-5}$. This means that similar words may become more different at a rate that is about ten times the rate at which different words may become more similar. We use this choice of the parameters both for the Indo-European and Austronesian groups.

A genetic distance $t(\alpha, \beta)$ is then computed for all pairs of languages in the database, obtaining two 50×50 upper triangular matrices with 1225 entries. These matrices preserve the topology of the lexical distance matrices but they also contain the information concerning absolute time scales.

8. Trees and distances distributions

Phylogenetic trees in Fig. 3 and in Fig. 4 are constructed from the matrices using the Unweighted Pair Group Method Average (UPGMA) which works as previously described.

The tree in Fig. 3 is similar to the one in [12] but there are some important differences. First of all, the first separation concerns Armenian, which forms a separate branch close to the root, while the other branch contains all the remaining Indo-European languages. Then, the second one is that of Greek, and only after there is a separation between the European branch and the Indoiranian one. In [12] the separation at the root gives origin to two branches, one with Indoiranian languages plus Armenian and Greek, the other with European

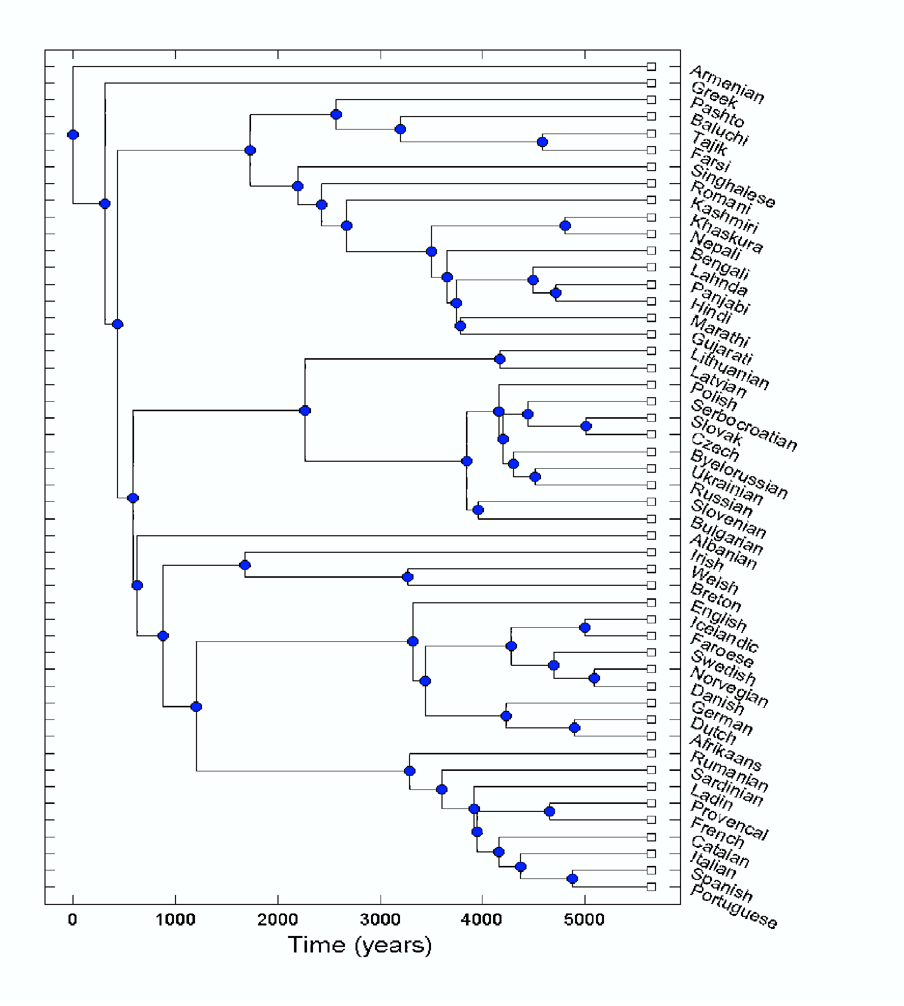

Figure 3. Indo-European phylogenetic tree constructed from the matrix of distances using the UPGMA.

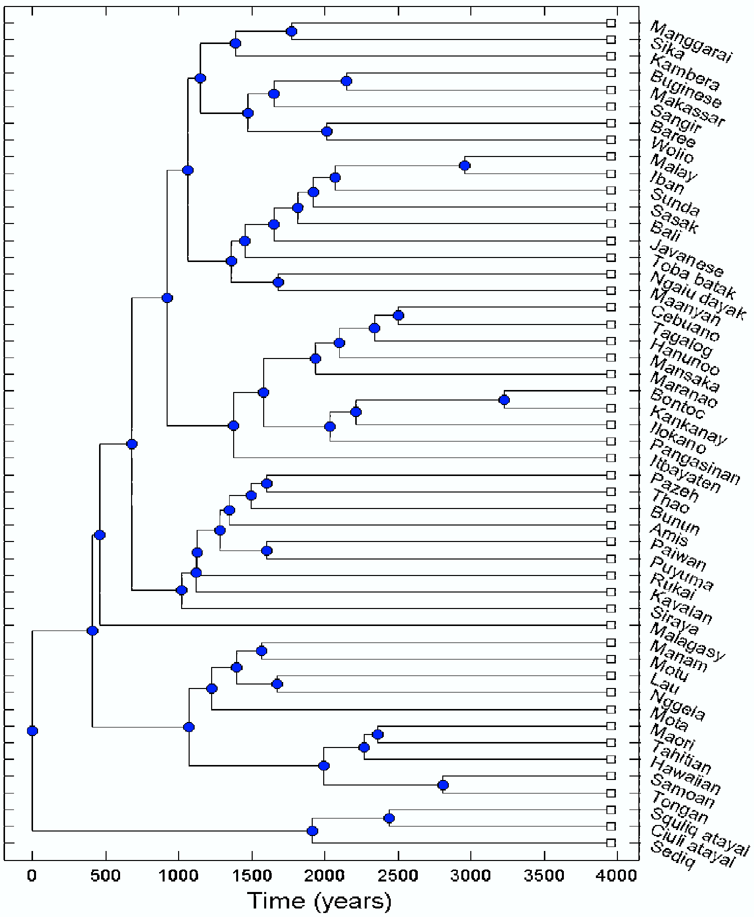

Figure 4. Austronesian phylogenetic tree constructed from the matrix of distances using the UPGMA.

languages. The position of Albanian is also different: in our case it is linked to European languages while in [12] it goes with Indoiranian ones. Finally, the Romani language is correctly located together with Indian languages but it is not as close to Singhalese as reported in [12].

In spite of these differences, our tree seems to confirm the conclusions reported in [12] about the Anatolian origin of the Indo-European languages, in fact, in our research, the first separation concerns those languages geographically closer to Anatolia, that is to say Armenian and Greek.

Also the tree in Fig. 4 is similar to the one in [13] but differences are more important here. The first separation concerns Formosan languages (Atayal group) which are in the first main branch, while all the other Formosan languages (Paiwan group) are in the second main branch together with all the other languages of the group. In this second main branch the first separation concerns the Oceanian languages, the second separation the Malagasy, the third all the remaining Formosans languages (Paiwan), the fourth the Philippino languages, and, finally, the fifth the Indonesian/Sulawesi languages with two sub groups: Sulawesi and Indonesian.

The fact that the first separation concerns Formosan languages of the Atayal group seems to confirm that the Austronesian group originated indeed in Formosa as it is widely accepted by researchers. Nevertheless, the second group of Formosan languages (Paiwan) is located differently and closer to the Philippino and Indonesian languages. This result, if confirmed, would suggest two different waves of migration from Formosa with different origin into the Island. Furthermore, the early separation of the Oceanian languages would be linked with the first wave and the Philippino/Indonesian with the second. Finally the Malagasy language is not grouped close to Kalimantan languages as often claimed in the literature even if the closest language (Maanyan) is in that group, a fact that suggests a multiple origin.

From the entries $t(\alpha, \beta)$ of the two matrices we may compute also the frequencies for the two groups. We plot in Fig. 5 the frequency of genetic distances of the Indo-European group of languages. If we compare this frequency with frequency in Fig. 1 we see that they are qualitatively very different. It should be considered that in the model of Kingman, genetic distances have the objective meaning of measuring time from separation while in our realistic case genetic distances are reconstructed from lexical distances. In this reconstruction we assume that lexical mutations and borrowings happen at a constant rate. This is true only on average, since there is an inherent randomness in this process which is not taken into account by the deterministic differential equation (7.1). Furthermore, the parameters a and b may vary from a pair of languages to another and also they may vary in time according to historical conditions.

There are two major consequences, the first, as already mentioned, is that the distributions of Fig. 5 and of Fig. 1 are quite different; the second, is that the trees 3 and 4 may be also quite different from the genealogical trees associated

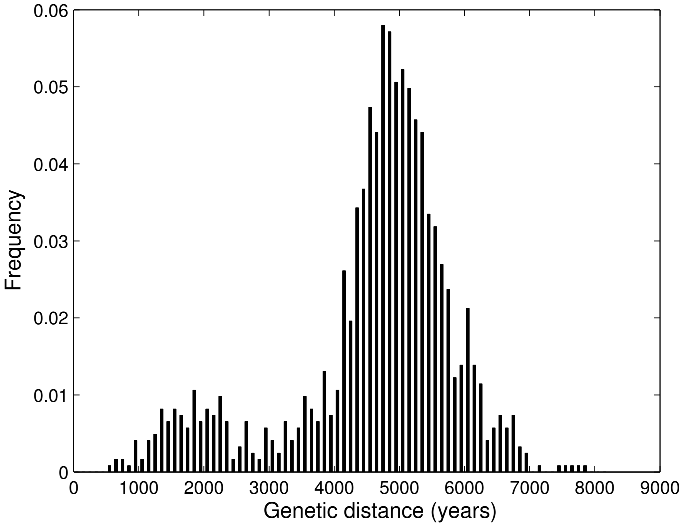

Figure 5. Distribution of distances obtained from all the 1225 pairs of languages from the Indo-European database.

to the history (which is hidden to us) of the languages of the two groups. In order to take into account the inner randomness and measure the probability of wrong tree reconstruction we have to modify equation (2.1) in order to include randomness in the distance. We will try to do this in the next two sections.

9. Random coalescent process

As already mentioned, in the coalescent model genealogical distances measure the time from the last common ancestor of two individuals while, in the case of languages, genetic distances are reconstructed from lexical distances. In turn, lexical distances are measured from differences between words and can take values between 0 and 1. If words are already partially different, new mutations may increase further their lexical distances, may be neutral, or even, they may decrease their distances. This is taken into account in equation (7.1), for this reason we transformed lexical distances into genetic ones. These last should be roughly proportional to the number of mutations which occurred in both the compared languages, independently on the fact that these mutations were able to increase or not the lexical distance.

If we assume that an individual randomly accumulates mutations at a constant rate, the genetic distance of a pair of individuals is then defined as the sum of the mutations that they accumulated since their last common ancestor, eventually rescaled by a factor. As a consequence, genetic distances are proportional only on average to genealogical ones. Therefore, we have to modify the deterministic equation (2.1) in order to take into account this randomness. We may assume that increments in the genetic distance have the simple form

$$t(\alpha,\beta) = t(g(\alpha), g(\beta)) + \gamma_\alpha + \gamma_\beta \tag{9.1}$$

where $g(\alpha)$ and $g(\beta)$ are the parents of α and β respectively, while γ_α and γ_β are random variables associated to the mutations of α and β . They are zero if the genome of the parent is transmitted unaltered and a positive constant if a mutation occurs. We assume that the probability for zero is $1 - \mu/(2N)$ and $\mu/(2N)$ for the positive constant $1/\mu$. In a compact form:

$$\gamma_\alpha = \begin{cases} 0, & p = 1 - \dfrac{\mu}{2N}, \\ \dfrac{1}{\mu}, & p = \dfrac{\mu}{2N}. \end{cases} \tag{9.2}$$

This rule grants that genetic distances are equal to genealogical distances on average. In fact, the expected value of the sum $\gamma_\alpha + \gamma_\beta$ is $1/N$.

Notice that we compare genealogical distances generated by (2.1) with genetic ones generated by (9.1). Since they describe two aspects of the same population, the family history must be the same. This means that the realization of the part of the process which assigns parents in (2.1) and in (9.1) must also be the same.

In Fig. 6 we have plotted the frequencies in a single population of 700 individuals of the genetic distances generated by (9.1) with $\mu = 20$. The realization of the parent attribution process $\alpha \to g(\alpha)$ used for the generation of the genetic distances is the same as the one used for the generation of genealogical distances. Therefore, Fig. 1 and Fig. 6 refer to the same historical event. The first one measures the frequency of unknown historical genealogical time separations the second one measures the frequency of genetic distances used as estimators of genealogical ones. The two distributions are very different since the noise smoothed the spikes in Fig. 1.

On the contrary, the distribution in Fig. 6 is qualitatively much more similar to Fig. 5. The main difference being that the part of distribution corresponding to small distances in Fig. 6 is absent in Fig. 5. This is due to our choice, in fact, we did not use all the languages in the original database but we avoided repetition of various close versions of the same languages. For example, we did not consider both Portuguese an Brazilian and we considered only a version of Irish and a version of Armenian. We made this choice because we were mostly interested in historical remote branchings, and not in trivial recent ones, for the reconstruction of the Indo-European tree. But, in this way we made a cut on the low distances in the frequency.

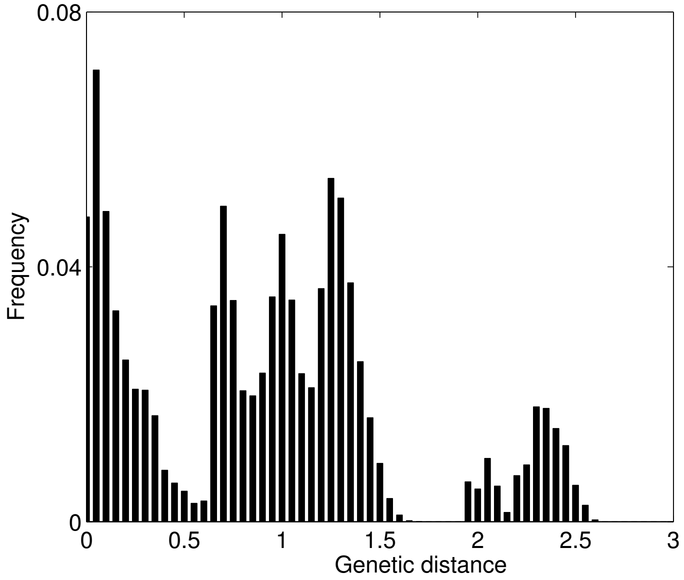

Figure 6. Frequency $f(t)$ of genetic distances in a single population computed for a population of 700 individuals and for $\mu = 20$. The population is the same as in Fig. 1 where the genealogical distances assume only few values, while here the genetic distances spread around the peaks of Fig. 1.

Finally, we would like to mention that at this stage we are still not able to make any quantitative comparison in order to give an estimate of the value of μ which better fits the case of languages.

10. Wrong tree reconstruction

When $\mu = 2N$ equation (2.1) and (9.1) coincide and randomness in mutations is lost. In this limiting case genetic and genealogical distances are equal and, not only the frequency distributions are identical, but also the family trees reconstructed by UPMGA will be exactly the same. For smaller values of μ, we expect that the fidelity level of reconstruction of a tree decreases. Then, we would like to have a quantitative information on the difference between the trees reconstructed from the matrices of genealogical and genetic distances.

We start considering the simplest situation of tree with three leaves. The topology of a three leaves tree is completely determined by the pair of individuals that match together first because the distance is smaller. Consequently, the genealogical and the genetic trees reconstructed by the UPGMA will have the same topology if the same pair of individuals has both the smallest genetic and genealogical distance. Let us call α, β and γ the three individuals, and assume that α and β are the pair with the smallest genealogical distance $d(\alpha,\beta)$. By the argument in Section 3 we know that $d(\alpha,\beta)=t_3$ and $d(\alpha,\gamma) \,= d(\beta,\gamma)=t_2+t_3$ where t_2 and t_3 are independent exponentially distributed variables with average 1 and 1/3 respectively. Then let us consider the two following events concerning genetic distances, the first that we call A is

$$t(\alpha,\beta) < \min\{t(\alpha,\gamma);\, t(\beta,\gamma)\} \tag{10.1}$$

If A is satisfied, the topology of the genetic tree reconstructed by UPGMA is the correct one since it is the same of that of the genealogical tree. The second that we call B is

$$\begin{gathered} t(\alpha,\beta) = \min\{t(\alpha,\gamma);\, t(\beta,\gamma)\}, \\ t(\alpha,\gamma) \,\neq\, t(\beta,\gamma) \end{gathered} \tag{10.2}$$

which corresponds to an ambiguous (but unlikely) situation for UPGMA which will be able to reconstruct correctly the tree with probability 1/2. The third that we call C will be

$$t(\alpha,\beta) = t(\alpha,\gamma) = \, t(\beta,\gamma) \tag{10.3}$$

which is also ambiguous (and even more unlikely). In this case, UPGMA will be able to reconstruct correctly the tree with probability 1/3.

Let us now call $P(A\,|\,t_2\,,t_3)$ the probability of the event A given the realized values t_2 and t_3, and $P(B\,|\,t_2\,,t_3)$ and $P(C\,|\,t_2\,,t_3)$ the equivalent conditional probabilities for the events B and C respectively. Let us also call $P(W\,|\,t_2\,,t_3)$ the probability of a wrong reconstruction of the tree correspondingly to t_2 and t_3. We have

$$P(W\,|\,t_2\,,t_3) \;=\; 1 \;-\; P(A\,|\,t_2\,,t_3) \;-\; \frac{1}{2}P(B\,|\,t_2\,,t_3) \;-\; \frac{1}{3}P(C\,|\,t_2\,,t_3). \tag{10.4}$$

Now we call $n(\alpha)$ the number of mutations along the branch α divided by μ, as shown in Fig. 7, analogously we define $n(\beta)$, $n(\gamma)$ and $n(\alpha\beta)$ as the numbers of mutations divided by μ along the branches indicated in Fig. 7. We will have $t(\alpha,\beta) \,=\, n(\alpha)+n(\beta)$, $t(\alpha,\gamma) \,=\, n(\alpha)+n(\alpha\beta)+n(\gamma)$ and $t(\beta,\gamma) \,=\, n(\beta)+n(\alpha\beta)+n(\gamma)$. The advantage is that the four new variables are independent and can be obtained as the sum of variables of type (9.2) where the sum goes on a number which is N times the time lag of the associated branch. Namely, t_3 for $n(\alpha)$ and $n(\beta)$, t_2 for $n(\alpha\beta)$ and t_2+t_3 for $n(\gamma)$.

Given this construction we can trivially but painfully compute the conditional probability $P(W\,|\,t_2\,,\,t_3)$ and, then, the absolute probability of a wrong tree $P(W)$ as the marginal of the joint probability $P(W\,|\,t_2\,,t_3)p(t_2)p(t_3)$ where

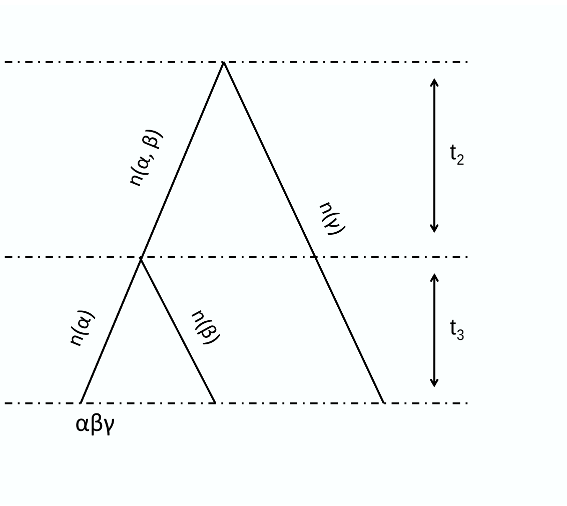

Figure 7. Outline of a three leaves tree. $n(\alpha)$, $n(\beta)$, $n(\gamma)$ and $n(\alpha\beta)$ are the number of mutations divided by μ (as explained in the text).

$p(t_2)$ and $p(t_3)$ are the exponential densities previously described. The probability of a wrong tree $P(W)$ is plotted in Fig. 8 with respect to the parameter μ.

If we take into account more than three individuals the situation immediately becomes more complicated since the possible tree topologies increase exponentially with the number of leaves. So we need to introduce a measure of difference between the genealogical tree and an associated genetic one. The simplest tree distance measure is the Robinson−Foulds Symmetric Difference [23], which only depends on the topology of the two trees and not on the differences in branch lengths.

The Symmetric Difference (SD) is computed by considering all possible branches that could exist on the two trees. Each inner branch, i.e. a branch connecting two nodes or one node to the root, identifies a *clade* in the set of leaves. The resulting distance is simply the number of clades present in one of the considered trees but not in the other. Therefore, two identical trees have zero SD, but it is enough to exchange two leaves on one of them to have a non zero SD.

In general SD has not an immediate statistical interpretation, i.e. we cannot say whether a larger distance is significantly larger than a smaller one. In the particular case of trees with only three leaves the symmetric distance is twice the wrong topology probability $P(W)$. In fact, in a three leaves tree there is only one clade and the Symmetric Distance is equal to 0 in the case of correct

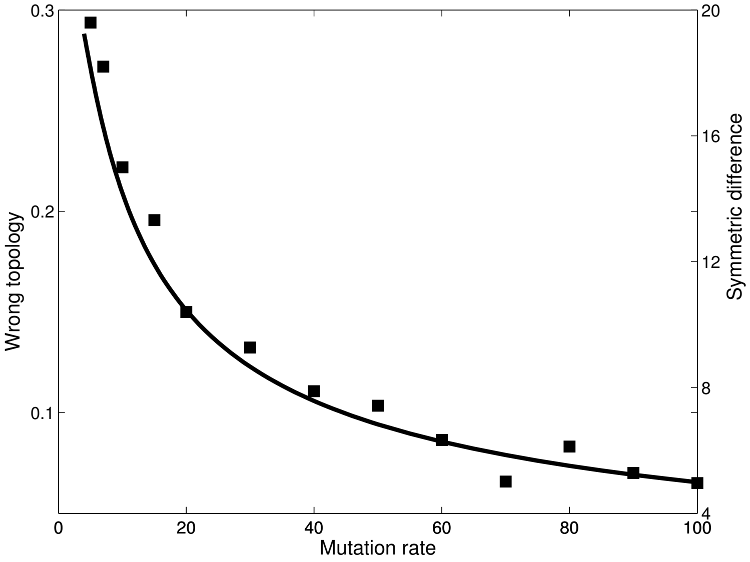

Figure 8. ESD between a genealogical tree and the associated genetic tree plotted with respect to μ. The bold line on the left y axes correspond to a three leaves tree while squares on the right y axes correspond to a 20 leaves tree. In the first case it is computed exactly and in this case the ESD is twice the probability of wrong topology.

topology (if both trees have the same clade) and is equal to 2 in the case of wrong one (if clades are different).

In order to compute numerically the expected SD (that we call ESD) between a genealogical tree and the associated genetic one with parameter μ we use the following procedure: we take 20 individuals in a population of 500 (a large one) and we use UPGMA to reconstruct their genealogical tree from a realization of the genealogical distances matrix. Then, we construct several associated genetic trees (5 for $\mu < 15$, 10 for greater values) and we compute their averaged SD with respect to the associated genealogical tree. We start again with a new realization of the matrix of genealogical distances and we repeat the procedure, ending with a new averaged SD. We do it many times (from 6 for $\mu = 5$ to 30 for $\mu = 100$) and, finally, we further average on all averaged SD ending with a quantity that should be very close to ESD. The number of genealogical trees and that of the associated genetic trees increases with μ since we observe an increasing fluctuation in the SD value, so we need to compute the average of more realizations of the process. In Fig. 8 we plot the estimated ESD of the 20 leaves tree. We also plot the ESD of a three leaves tree which is twice the exactly computed $P(W)$. We find out that they only differ for a factor due to the total number of clades, which depends on the number of leaves.

11. Discussion and conclusions

Before discussing our results concerning languages, we would like to comment on the relevance of the coalescent process phenomenology in biological applications. Our example concerns the use of mtDNA in recent paleoanthropological studies. Mithocondrial DNA (mtDNA) is inherited only from the mother, for this reason, mtDNA of a given species should be considered as an haploid population and results in this paper should apply to it. Let us discuss the first example. In the years from 1997 to 2000 some mtDNA from three different specimen of neandertal was extracted [18, 19] and short strands of the hyper-variable region (HVR1 and HVR2) were amplified using polymerase chain reaction (PRC). After comparison with mtDNA of modern humans they found that the distance neandertal/modern is about three times the distance modern/modern and about twice the distance neandertal/neandertal. The conclusion was that, given the above ranges in differences, neandertals mtDNA is statistically different from modern humans mtDNA

But our point of view is different. Consider the situation as it was 40 thousands years ago, when moderns and neandertals 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 Fig. 1. In this case, distances inside some subpopulations could be easily one third or one half of the distance between different subpopulations. To be clear, we do not conclude here that humanity was a single large interbreeding population (multiregional hypothesis) but we only claim that the mtDNA argument cannot be used to support the opposite conclusion.

For what concerns languages, we would like to stress that our method to find out distances is very simple and does not require any previous knowledge of languages origin. Also, it can be applied directly to all those language pairs for which a translation of a small group of words exists. Furthermore, our method correctly finds out all clusters, but while for the Indo-European group, the hierarchical organization of clusters is similar to that found by [12], for the Austronesian group it is quite different from that obtained by [13]. We find out, according to [12], that in the Indo-European tree the first separation concerns languages geographically close to Anatolia, so that the Anatolian origin of the group seems to be confirmed. The only relevant difference with [12] is that Romani seems to be closer to Northern India subgroup than to Singhalese. In the Austronesian tree, differences with [13] are more relevant since Formosan languages split in two clusters with a different hierarchical position and Oceanian languages separate earlier. This suggests two different waves of migration from the original location.

Finally, we would like to mention that our results must be seen at the light of the last two sections of the paper. A partially wrong reconstruction of the trees remains possible. The probability of errors could be smaller if translations and transliterations are more accurate, nevertheless, the inner randomness of lexical mutation is an insuperable obstacle to a safe reconstruction of the trees.

Acknowledgements

We warmly thank S.J. Greenhill, R. Blust and R.D. Gray, for the authorization to use their: *The Austronesian Basic Vocabulary Database*,

`http://language.psy.auckland.ac.nz/austronesian`

which we consulted in January 2008. We also thank Lydie Irene Andriamiseza for improvements in the Malagasy database and related Austronesian Swadesh lists. We acknowledge the financial support of MIUR Università di L'Aquila, Cofin 2004 n. 2004028108_005. The work by FP has been supported by European Commission Project E2C2 FP6-2003-NEST-Path-012975 Extreme Events: Causes and Consequences.

References

[1] G. Adcock, E. Dennis, S. Easteal, G. Huttley, L. Jermin, W. Peacock and A. Thorne (2001) Mitochondrial DNA sequences in ancient Australians: Implications for modern human origins. *Proceedings of the National Academy of Science* **98**, 537–542.

[2] D. Aldous (1999) Deterministic and stochastic models for coalescence (aggregation, coagulation): a review of mean field theory for probabilists. *Bernoulli* **5**, 3–48.

[3] A. Dalal and E. Schmutz (2002) Compositions of random functions on a finite set. *Electronic Journal of Combinatorics* **9**, R26.

[4] B. Derrida and D. Bessis (1999) Statistical properties of valleys in the annealed random map model. *J. Phys. A: Math. and Gen.* **21**, L509–L515.

[5] B. Derrida and B. Jung-Muller (1999) The genealogical tree of a chromosome. *J. Stat. Phys.* **94**, 277–298.

[6] B. Derrida, S.C. Manrubia and D.H. Zanette (1999) Statistical properties of genealogical trees. *Phys. Rev. Lett.* **82**, 1987–1990.

[7] B. Derrida and L. Peliti (1991) Evolution in a flat fitness landscape. *Bulletin of Mathematical Biology* **53**, 355–382.

[8] P. Donnelly (1991) Weak convergence to a Markov chain with an entrance boundary: ancestral processes in population genetics. *Ann. Prob.* **19** (3), 1102–1117.

[9] I. Dyen, J.B. Kruskal and P. Black (1997) *FILE IE-DATA1*. Available at `http://www.ntu.edu.au/education/langs/ielex/IE-DATA1`

[10] J. Fill (2002) On compositions of random functions on a finite set. Preprint.

[11] W.M.Y. Goh, P. Hitczenko and E. Schmutz (2002) Iterating random functions on a finite set. Preprint `ArXiv:math.CO/0207276v2`.

[12] R.D. Gray and Q.D. Atkinson (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439.

[13] R.D. Gray and F.M. Jordan (2000) Language trees support the express-train sequence of Austronesian expansion. *Nature* **405**, 1052–1055.

[14] J.F.C. Kingman (1982) The coalescent. *Stoch. Process. and Appl.* **13**, 235–248.

[15] J.F.C. Kingman (1982) On the genealogy of large populations. Essays in statistical science. *J. Appl. Probab.* **19A**, 27–43.

[16] J.F.C. Kingman (1982) Exchangeability and the evolution of large populations. In: *Exchangeability in Probability and Statistics*, North-Holland, Amsterdam-New York, 97–112.

[17] J.F.C. Kingman (1980) *Mathematics of Genetic Diversity.* CBMS-NSF Regional Conference Series in Applied Mathematics, **34**. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa. ISBN: 0-89871-166-5.

[18] M. Krings, C. Capelli, F. Tschentscher, H. Geisert, S. Meyer, A. von Haeseler, K. Grossshmidt, G. Possnert, M. Paunovic and S. Pääbo (2000) A view of Neandertal genetic diversity. *Nature Genetics* **26**, 144–146.

[19] M. Krings, A. Stone, R.W. Schmitz, H. Krainitzki, M. Stoneking and S. Pääbo (1997) Neandertal DNA sequences and the origin of modern humans. *Cell* **90**, 19–30.

[20] M. Möhle (2000) Total variation distances and rates of convergence for ancestral coalescent processes in exchangeable population models. *Adv. Appl. Probab.* **32**, 983–993.

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

[22] S.J. Greenhill, R. Blust and R.D. Gray (2003–2008) *The Austronesian Basic Vocabulary Database.* `http://language.psy.auckland.ac.nz/austronesian`.

[23] D.F. Robinson and L.R. Foulds (1981) Comparison of phylogenetic trees. *Math. Biosci.* **53** (1/2), 131–147.

[24] F. Petroni and M. Serva (2008) Languages distances and trees reconstruction. *J. Stat. Mechanics: Theory and Experiment*, P08012.

[25] M. Serva (2004) Lack of self averaging in family trees. *Physica A* **332**, 387–393.

[26] M. Serva (2005) On the genealogy of populations: trees, branches and offspring. *J. Stat. Mechanics: Theory and Experiment*, P07011.

[27] M. Serva and L. Peliti (1991) A statistical model of an evolving population with sexual reproduction. *J. Phys. A: Math. and Gen.* **24**, L705–L709.

[28] M. Serva and F. Petroni (2008) Indo-European languages tree by Levenshtein distance. *EuroPhysics Letters* **81**, 68005.

[29] D. Simon and B. Derrida (2006) Evolution of the most recent common ancestor of a population with no selection. *J. Stat. Mechanics: Theory and Experiment*, P05002.

[30] P.H.A. Sneath and R.R. Sokal (1973) *Numerical Taxonomy.* Freeman, San Francisco.

[31] M. Swadesh (1952) Lexicostatistic dating of prehistoric ethnic contacts. *Proc. Amer. Phil. Soc.* **96**, 452–463.

[32] S. Tavare (1984) Line-of-descent and genealogical processes and their applications in population genetics models. *Theoretical Population Biology* **26**, 119–164.

[33] G.A. Watterson (1984) Lines of descent and the Coalescent. *Theoretical Population Biology* **26**, 256–276.

[34] Y.-C. Zhang, M. Serva and M. Policarpov (1990) Diffusion reproduction processes. *J. Stat. Phys.* **58**, 849–861.

[35] The database, as modified by the Authors, is available at the following web address: `http://univaq.it/∼serva/languages/languages.html`. Readers are welcome to modify, correct and add words to the database.