A LETTERS JOURNAL EXPLORING THE FRONTIERS OF PHYSICS

OFFPRINT

Recovering geography from a matrix of genetic distances

M. SERVA, D. VERGNI, D. VOLCHENKOV and A. VULPIANI

EPL, **118** (2017) 48003

A Letters Journal Exploring the Frontiers of Physics

AN INVITATION TO SUBMIT YOUR WORK

epljournal.org

The Editorial Board invites you to submit your letters to EPL

EPL is a leading international journal publishing original, innovative Letters in all areas of physics, ranging from condensed matter topics and interdisciplinary research to astrophysics, geophysics, plasma and fusion sciences, including those with application potential.

The high profile of the journal combined with the excellent scientific quality of the articles ensures that EPL is an essential resource for its worldwide audience. EPL offers authors global visibility and a great opportunity to share their work with others across the whole of the physics community.

Run by active scientists, for scientists

EPL is reviewed by scientists for scientists, to serve and support the international scientific community. The Editorial Board is a team of active research scientists with an expert understanding of the needs of both authors and researchers.

OVER
568,000
full text downloads in 2015

18 DAYS
average accept to online publication in 2015

20,300
citations in 2015

"We greatly appreciate the efficient, professional and rapid processing of our paper by your team."

Cong Lin
Shanghai University

Six good reasons to publish with EPL

We want to work with you to gain recognition for your research through worldwide visibility and high citations. As an EPL author, you will benefit from:

1. **Quality** – The 60+ Co-editors, who are experts in their field, oversee the entire peer-review process, from selection of the referees to making all final acceptance decisions.
2. **Convenience** – Easy to access compilations of recent articles in specific narrow fields available on the website.
3. **Speed of processing** – We aim to provide you with a quick and efficient service; the median time from submission to online publication is under 100 days.
4. **High visibility** – Strong promotion and visibility through material available at over 300 events annually, distributed via e-mail, and targeted mailshot newsletters.
5. **International reach** – Over 3200 institutions have access to EPL, enabling your work to be read by your peers in 100 countries.
6. **Open access** – Articles are offered open access for a one-off author payment; green open access on all others with a 12-month embargo.

Details on preparing, submitting and tracking the progress of your manuscript from submission to acceptance are available on the EPL submission website **epletters.net**.

If you would like further information about our author service or EPL in general, please visit **epljournal.org** or e-mail us at **info@epljournal.org**.

EPL is published in partnership with:

European Physical Society | Società Italiana di Fisica | EDP Sciences | IOP Publishing

A Letters Journal Exploring the Frontiers of Physics

May 2017

EPL, **118** (2017) 48003 www.epljournal.org
doi: 10.1209/0295-5075/118/48003

Recovering geography from a matrix of genetic distances

M. Serva[1], D. Vergni[2], D. Volchenkov[3,4] and A. Vulpiani[5,6]

[1] *Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica, Università dell'Aquila - L'Aquila, Italy*
[2] *Istituto per le Applicazioni del Calcolo "Mauro Picone" (IAC-CNR) - Roma, Italy*
[3] *Mathematics & Statistics, Texas Tech University - Lubbock, TX, USA*
[4] *Sichuan University of Science and Engineering - Sichuan, Zigong, China*
[5] *Dipartimento di Fisica, Università di Roma "Sapienza" - Roma, Italy*
[6] *Centro Interdisciplinare "B. Segre", Accademia dei Lincei - Roma, Italy*

received 25 May 2017; accepted in final form 5 July 2017
published online 26 July 2017

PACS 87.23.Ge – Dynamics of social systems
PACS 87.23.Kg – Dynamics of evolution

Abstract – Given a population of N elements with their geographical positions and the genetic (or lexical) distances between couples of elements (inferred, for example, from lexical differences between dialects which are spoken in different towns or from genetic differences between animal populations living in different faunal areas) a very interesting problem is to reconstruct the geographical positions of individuals using only genetic/lexical distances. From a technical point of view the program consists in extracting from the genetic/lexical distances a set of reconstructed geographical positions to be compared with the real ones. We show that geographical recovering is successful when the genetic/lexical distances are not a simple consequence of phylogenesis but also of horizontal transfers as, for example, vocabulary borrowings between different languages. Our results go well beyond the simple observation that geographical distances and genetic/lexical distances are correlated. The ascertainment of a correlation, in our perspective, merely is a prerequisite.

Introduction. – It is well known that in genetics and in lexicostatistics distances $D_{i,j}$ between couple of individuals (taken from a population of N individuals, with $i, j = 1, \ldots, N$) can be operatively computed starting from genetic [1–8] and lexical [9–19] data.

Both in biology and linguistics, the matrix of genetic/lexical distances, $D_{i,j}$, is often used for the construction of phylogenetic trees, as, for example, the $UPGMA$ tree [20] and the NJ tree [21]. Over each of these trees one can measure the reconstructed phylogenetic distances between pairs of individuals. The major problem is that $D_{i,j}$ is a symmetric matrix (with vanishing diagonal elements) with $N(N-1)/2$ elements, while the cited trees try to recover the matrix $D_{i,j}$ with a number of free variables which is smaller, typically of order of N. As a consequence, reconstruction of distances is usually approximated; only if $D_{i,j}$ is itself the output of a process whose nature is purely phylogenetic, the reconstruction can be totally correct. For example, $UPGMA$ leads to a totally correct reconstruction only in the case of a process with haploid reproduction and constant mutation rate, while for NJ variable mutation rate is also allowed.

In reality, pure phylogenesis is quite rare and the entries of the matrix of distances and distances on the generated tree are different. Their degree of similarity can be quantified by a proper index as, for example, that one used in [22] or, more simply, by measuring their correlation. It should be stressed that the inaccuracy of tree reconstruction can also affect the topological structure as, for example, pointed out in [23].

The main reason of inadequacy of tree reconstruction lies in horizontal transfer processes such as horizontal gene transfer between neighbors or vocabulary borrowings between different languages. These processes break the purely ultra-metric phylogenetic structure of the matrix of distances. Thus, in the translation of the matrix of genetic distances $D_{i,j}$ in a phylogenetic trees, many information can be totally lost especially those concerning geography which, indeed, is relevant both in biology and linguistics [24–26].

We propose in this paper a different interpretation of a matrix of genetic distances which privileges geography with respect to phylogenetics. The next section is devoted to the presentation of the model of geographical

reconstruction while in the section "A simple model" a simple stochastical model able to directly generate a matrix of genetic distances $D_{i,j}$ is given. In the "Result" section the reader can find the discussion of results while some remarks and conclusions are given in the section "Conclusions".

Reconstruction of geographical positions. – Once we know that $[x_i^g, y_i^g]$ are the geographical locations of individuals, a preliminary requirement for the feasibility of our program is that there is a strong correlation between genetic/lexical distances $D_{i,j}$ and geographical distances $D^g_{i,j} = [(x_i^g - x_j^g)^2 + (y_i^g - y_j^g)^2]^{\frac{1}{2}}$.

In order to focus on a real case we considered the $N(N-1)/2 = 253$ lexical distances $D_{i,j}$ between pairs of $N = 23$ Malagasy dialects that we computed in [27,28] from Swadesh lists of words. We also considered the geographical distances $D^g_{i,j}$ obtained by the geographical coordinates of the corresponding towns where the dialects are spoken. The geographical distances are indifferently computed using great-circle or chord distance, considering Madagascar a flat bi-dimensional object for all purposes of the present article.

We find a correlation coefficient between the geographical distance and the genetic distance such as $C(D, D_g) = 0.675$, which is a quite large value indicating that geography strongly influences the relatedness among dialects. Therefore, in this case, we expect that the construction of phylogenetic trees is not sufficient since the matrix of genetic distances $D_{i,j}$ contains information concerning geography of Madagascar which are neglected by trees.

In what follows the methodology able to extract geographical information from the matrix of genetic distances is presented. Imagine that geography is unknown, *i.e.*, the geographical positions of N individuals, $[x_i^g, y_i^g]$, are unknown and we want to reconstruct them form genetic/lexical data. To each individual, i, we arbitrarily associate a position $[x_i, y_i]$. Then, the Euclidean distance between two individuals is $[(x_i - x_j)^2 + (y_i - y_j)^2]^{\frac{1}{2}}$ so that we can define the cost function

$$R(\mathbf{x}, \mathbf{y}) = \sum_{i<j} \left[D^2_{i,j} - (x_i - x_j)^2 - (y_i - y_j)^2 \right]^2, \qquad (1)$$

where $\mathbf{x}, \mathbf{y}$ indicates the configuration $[x_1, y_1]$, $[x_2, y_2], \ldots, [x_N, y_N]$ and the set of genetic distances, $D_{i,j}$, is given.

Since positions $[x_i, y_i]$ are arbitrary, the quantity $R(\mathbf{x}, \mathbf{y})$ is meaningless unless one finds those positions $[\bar{x}_i, \bar{y}_i]$ whose distances $[(\bar{x}_i - \bar{x}_j)^2 + (\bar{y}_i - \bar{y}_j)^2]^{\frac{1}{2}}$ better coincides with the genetic distances $D_{i,j}$. This optimal configuration $\bar{\mathbf{x}}, \bar{\mathbf{y}} = [\bar{x}_1, \bar{y}_1]$; $[\bar{x}_2, \bar{y}_2]; \ldots; [\bar{x}_N, \bar{y}_N]$ can be simply found minimizing (1) with respect to all variables of the configuration

$$\bar{R} = \min_{\mathbf{x},\mathbf{y}} \left(R(\mathbf{x}, \mathbf{y}) \right) = R(\bar{\mathbf{x}}, \bar{\mathbf{y}}), \qquad (2)$$

which gives the optimal configuration $\bar{\mathbf{x}}, \bar{\mathbf{y}}$. We used a cost function where squared distances are compared instead of distances. Note that the quantities in eq. (1) are homogeneous since the 2d Euclidean coordinates resulting from the minimization of such equation are not the "true" geographical positions, but their Euclidean distance is similar, in a geometric sense, to the real one. In the limit case in which the unique mechanism is the geographical one, the minimum of eq. (1) is zero.

Indeed, it is easy to get convinced that the minimum is for sure not unique unless one preliminary anchors the two-dimensional representation by fixing origin, orientation and specularity for reflection with respect to the two cardinal axes. For example, with the choice $x_1 = y_1 = y_2 = 0$, the first individual is in the origin and the second on the x-axis. The number of variables to optimized is, therefore, $2N - 3$. Moreover the problem of specularity for reflection with respect to the two cardinal axes can be resolved by choosing the signs of x_2 and y_3. For example, choosing both positive, the second individual is on the positive x semi-axis and the third individual is in the upper half-plane. Also doing so the structure of minima of the function (1) could be very complicated and an accurate study of the minimal values found by the numerical algorithm starting with different initial configuration $\mathbf{x}, \mathbf{y}$ is needed in order to assure the reaching of a satisfactory minimum.

As an output we obtain the set of reconstructed optimal positions $[\bar{x}_i, \bar{y}_i]$ (reconstructed geographical positions) and also the set of reconstructed distances $\bar{D}_{ij} = [(\bar{x}_i - \bar{x}_j)^2 + (\bar{y}_i - \bar{y}_j)^2]^{\frac{1}{2}}$. Necessarily some information is lost in this procedure since the original matrix, D, has $N(N-1)/2$ entries, D_{ij}, which we try to reproduce by the $\bar{D}_{ij}$ which depends only on $2N - 3$ coordinates. A measure of the loss of information is the correlation $C(D, \bar{D})$ between D_{ij} and $\bar{D}_{ij}$. More interestingly, one can compute the correlation $C(\bar{D}, D^g)$ between the reconstructed distances $\bar{D}_{ij}$ and the real geographical distances $D^g_{ij} = [(x_i^g - x_j^g)^2 + (y_i^g - y_j^g)^2]^{\frac{1}{2}}$. This gives a measure of the quality of geographical reconstruction obtained using only genetic data.

For Malagasy dialects we had as input lexical distances, D, with a correlation $C(D, D_g) = 0.675$ with geographic distances, D_g, and we have as output the reconstructed optimal positions $\bar{\mathbf{x}}, \bar{\mathbf{y}}$, whose distance matrix, $\bar{D}$, has a correlation with the original lexical distances $C(D, \bar{D}) = 0.835$ which means that the $N(N-1)/2$ entries of the matrix of lexical distances is very well represented by the coordinates $\bar{\mathbf{x}}, \bar{\mathbf{y}}$. More importantly, real geographical distances and reconstructed distances have a quite large correlation $C(\bar{D}, D_g) = 0.690$ indicating that geography is better recovered from the reconstructed optimal configuration than by lexical data. Let us stress that in the transition from D_{ij} to $\bar{D}_{ij}$ no information about geography was lost, on the contrary there was an increase of correlation, although small, from 0.675 to 0.690. This implies that geographically close dialects deeply influence each other and this horizontal transfer is, at least, as important as phylogenetics.

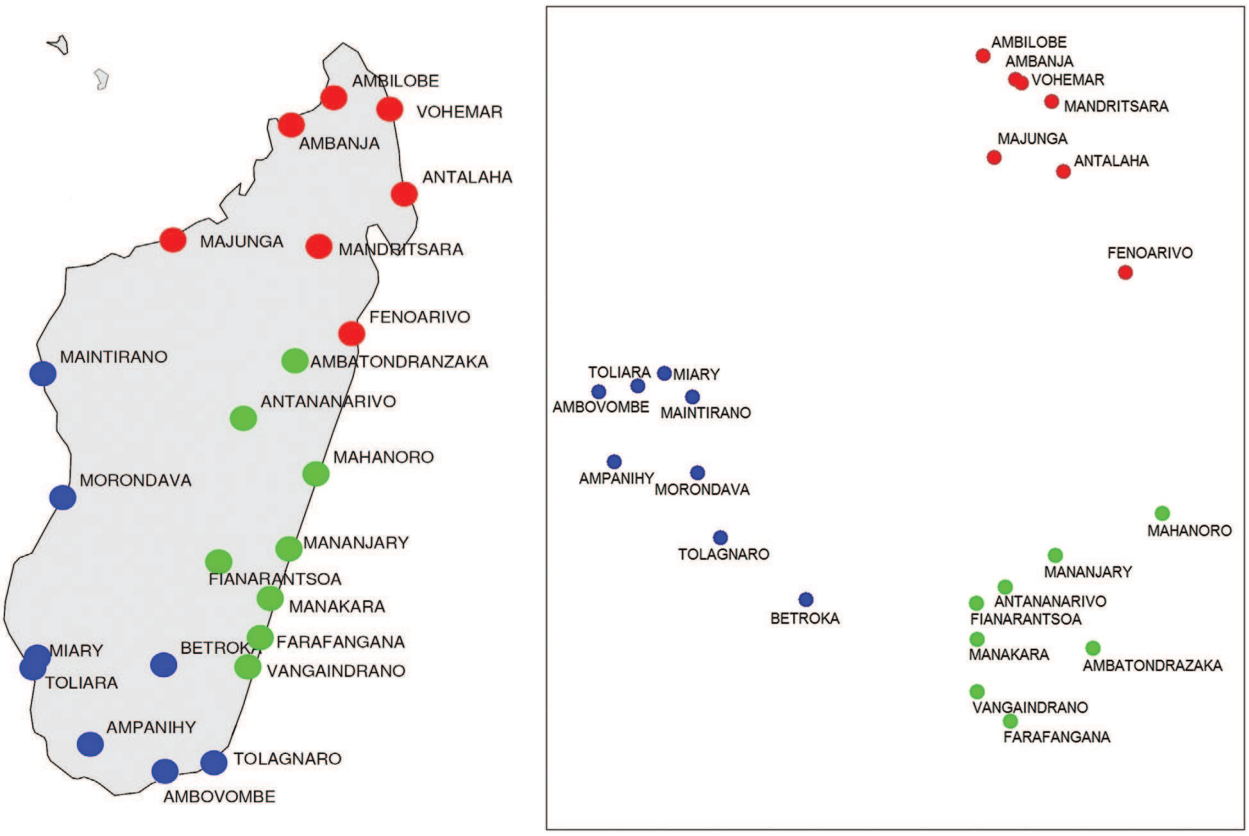

Fig. 1: (Color online) Geographical positions of the 23 towns in Madagascar (left) and optimal reconstruction of the positions from dialects (right).

Moreover, the comparison of the left side of fig. 1, where the towns are geographically located in $[x_i^g, y_i^g]$, and the right side, where the optimal positions $[\bar{x}_i, \bar{y}_i]$ are depicted, gives a qualitative perception of the accuracy goodness of the geographical reconstruction. Although the reconstruction is imprecise, there is a clear correspondence between the two pictures. Physical barriers (such as mountains and rivers) may partially explain the differences between geographical and reconstructed positions, most of the difference is due to phylogenesis which is the complementary phenomenon which explains genetic distances. Although the reconstruction is imprecise, it is remarkable that it has been obtained only from lexical data, totally neglecting geographical inputs. Using a colorful language, we could say that in case we ignored the geography of Madagascar we could have an idea of it simply collecting lists of words of various dialects.

It must be noted that the physical dimension of the reconstructed geography has to be the same as the physical dimension of the "true" geography (dimension two for towns or faunal areas, dimension three for stars,...). In this work we privileged dimension two since in most cases one has to handle individuals situated on a surface which is approximately plane (as in the case of Malagasy towns with corresponding dialects), but everything can be easily translated to different physical dimensions.

Finally, we would like to stress that the purpose of this geographical interpretation of genetic/lexical distances is different from other approaches, such as the Principal Component Analysis (PCA), the improved versions of PCA [29,30], and multidimensional scaling (MDS) also known as the Principal Coordinates Analysis ($PCoA$) [31]. For PCA the focus is in embedding data which are in a multidimensional space (matrix) in a lower-dimensional one which maintains most of the information contained in the matrix. $PCoA$ (or MDS) refers to an ordination technique aiming to place each object in N-dimensional space such that the between-object distances are preserved as well as possible by minimizing a stress function resembling that of (1). Limiting the dimension to two, applying $PCoA/MDS$ to the lexical distance matrix D_{ij}, one could obtain results similar to our results, but in our work we have the specific purpose of reconstructing the geographical locations of individuals from the available genetic and lexical differences and to study under which conditions it is possible to obtain good geographical information about distances on a surface from the minimum-distortion embedding of complex genetic and lexical relations into a physical landscape.

A simple model. – We consider here a simple, but not trivial, model which allows to precisely test when distances

are better represented by a geographical approach and when they are better represented by a phylogenetic tree.

Let us assume a population of N individuals with no differences in fitness and whose size N remains the same at all times. An individual is typically the population of a village/town (linguistics) or an animal/plant population in a given faunal area (biology).

Any generation is replaced by a new one at any time step and we assume that the time t is an integer which numbers the generations. The genetic distances between pairs of individuals i an j are the $N(N-1)/2$ entries of a symmetrical matrix ($D_{i,j}(t) = D_{j,i}(t)$) with vanishing diagonal elements ($D_{i,i}(t) = 0$). Moreover, any individual is identified by its fixed position on a unitary circumference so that i indicates the individual whose geographical position is $[x_i^g, y_i^g] = [\cos(2\pi i/N),\ \sin(2\pi i/N)]$, with $i = 1, \ldots, N$.

In place of simulating the evolution of the genetic (or linguistic) makeup of any individual [22,32–34], we equivalently chose to simulate directly the evolution of distances [35–39].

The initial state can be chosen assuming that all individuals are identical ($D_{i,j}(0) = 0$). The evolution of this matrix consists, at any generation step, of three steps: mutation, death/reproduction and gene-flow.

The first step concerns mutation and distances increase. This can happen at different and eventually random rates, but, for the sake of simplicity, we assume a constant rate:

$$D'_{i,j}(t) = D_{i,j}(t) + \gamma\big[1 - D_{i,j}(t)\big] \tag{3}$$

for any pair with $i \neq j$, while for diagonal elements $D'_{i,i}(t) = 0$. The parameter $0 \leq \gamma \leq 1$ is proportional to the mutation rate, while $[1 - D_{i,j}(t)]$ is the fraction of genome the two individual have still in common.

The second step (death/reproduction) implies that some of the individuals have no offspring and some other have more than one. We simply assume that at any time t each individual i has a single parent $\alpha(i,t)$, where $\alpha(i,t)$ are independent random variables for different individuals i and for different times t. With probability $1-p$ one has that $\alpha(i,t)$ equals i (parent is at the same location) and with probability p one has that $\alpha(i,t)$ takes at random one of the $N-1$ values $k \neq i$ (parent is in another location, meaning extinction of a local population and doubling of another one).

Thus for any pair with $i \neq j$ one has the following stochastic equation:

$$D''_{i,j}(t) = D'_{\alpha(i,t),\alpha(j,t)}(t), \tag{4}$$

while for diagonal elements $D''_{i,i}(t) = 0$. Notice that this passage sets some distances to zero since $\alpha(i,t)$ and $\alpha(j,t)$ can be equal even if i and j are not. Also notice that the average number of populations which extinguish in a time step is pN.

The third step (gene-flow) allows for some genetic flow between two nearest individuals. Therefore, for any couple of individuals with $i \neq j$,

$$D_{i,j}(t+1) = \sum_{i',j'} \epsilon(i,i')\epsilon(j,j')D''_{i',j'}(t), \tag{5}$$

where i' can be either i or a first neighbor of i and the same for j'. The coefficient $\epsilon(i,i')$ equals $1-q$ if i' coincides with i and it equals $q/2$ if i' is one a the two first neighbors of i. Also in this case the diagonal elements are zero, $D_{i,i}(t+1) = 0$. Notice that this passage tends to decrease those distances where i and j are first or second neighbors. In this case, in fact, i' may be equal to j' so that one (second neighbors) or two (first neighbors) elements in the sum vanish. This passage means that a fraction $q < 1$ of the genome of any individual is replaced by the genome of its two neighbors. In linguistics this horizontal transfer corresponds to lexical borrowings from geographically close languages or dialects.

It is important to note that the γ parameter must be very small in such a way the genetic distances increase almost continuously in time due to mutations. Such an assumption is quite common both in biology and linguistics and corresponds to the observed phenomenology. From a mathematical point of view γ has to be of the order of $1/N$ (or less) to ensure the proper infinite population size limit [40]. Moreover, at varying p and q the geographical reconstruction passes from being very good to being very poor, as discussed in the following. Finally, our choice for the value of N is arbitrary since it does not influence the geographic reconstruction.

After an initial transient T needed to reach a stationary state, any matrices $D_{i,j}(T+t)$ can be taken as a representative of $D_{i,j}$. An upper bound for T is $10 \cdot N/p + 2N^2/q$ according to the fact that time for coalescence (same ancestor for all individuals) is of the order of N/p (but it can be several time this value for some realizations) [35–37,39] while the diffusion time over the ring for the random walk underling the third passage is of the order of N^2/q.

Results. – Let us discuss the numerical results obtained from the model introduced in the previous section. As already mentioned, a prerequisite for geography reconstruction is a strong correlation between the genetic distances and the geographical distances, therefore we computed the correlation $C(D, D^g)$ between the $D_{i,j}(t)$ and the $D^g_{i,j}$ at different times t. Geographical distances are given by $D^g_{i,j} = 2\sin(\pi||i-j||/N)$ where $||i-j|| = \min(|i-j|, N-|i-j|)$.

In fig. 2 (left) we show $C(D(t), D^g)$ for $T \leq t \leq 2T$ for the stochastic evolution of the distance $D(t)$ according to eqs. (3), (4) and (5). The population is composed of $N = 25$ individuals and the model parameters are $q = 0.2$, $\gamma = 0.001$ and three different values of p. It can be seen that in all the considered cases a time $T = 10000$ is largely sufficient for reaching a stationary state, *i.e.*, in the range $T \leq t \leq 2T$ only fluctuations around a typical value appear and no trend is detectable. On the right side of fig. 2 we have plotted the value of the averaged

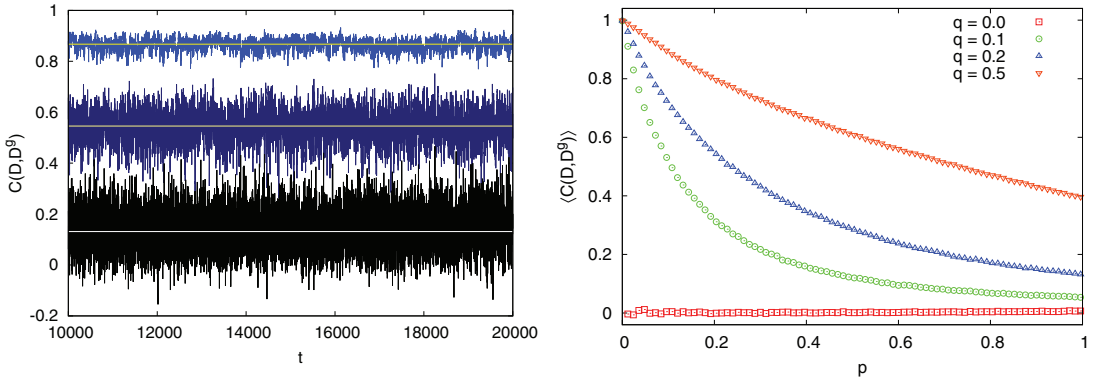

Fig. 2: (Color online) (Left) Correlation between genetic distances $D_{i,j}(t)$ and geographical ones $D^g_{i,j}$ as a function of $T \leq t \leq 2T$ for a population of $N = 25$ individuals with $q = 0.2, \gamma = 0.001$ and $T = 10000$. The values of p are: $p = 0.004$ (top), $p = 0.2$ (middle), $p = 1.0$ (bottom). The average values of the correlation over the window are 0.867 for $p = 0.004$, 0.546 for $p = 0.02$ and 0.134 for $p = 1$. (Right) Averaged correlation (again $N = 25, \gamma = 0.001$) as a function of p for different values of q.

correlation $\langle C(D(t), D^g)\rangle_t$ (again for the population $N = 25$, $\gamma = 0.001$) as a function of p for four different values of q. The average is made over the same time window of the left side of fig. 2. As expected the correlation decreases with p and increases with q. When $q = 0$ the process is purely phylogenetic and correlation between geographical and genetic distances is totally absent, nevertheless, when $q \neq 0$ even for $p = 1$ (all parents in a random location) some correlation survives.

From this preliminary investigation we can have an idea of the range of values of p and q which allow for good, or at least acceptable, reconstruction of geography. For example, for $q = 0.2$ only those value of p which are in the interval $[0, 0.2]$ should lead to a good geographical reconstruction since the correlation is sufficiently high.

In fig. 3, we show the reconstructed geographical positions $[\bar{x}_i, \bar{y}_i]$ and we remind that in the model there are $N = 25$ individuals whose "true" geographical positions are equally spaced on a unitary circumference. All the four reconstructed geographies are made choosing $\gamma = 0.001$ and $q = 0.2$. The figure contains four panels corresponding to $p = 1$, $p = 0.5$, $p = 0.2$ and $p = 0.04$. Notice that for $p = 1$ all individuals are replaced in a single generation but also in the case $p = 0.04$ the replacement rate is high since, on average, in a single generation one individual extinguishes and it is replaced. Colors of points are inserted in order to add a feeling of the goodness of the reconstruction, otherwise only the distance of points from circumference could be perceived.

The upper left panel in fig. 3 corresponds to $p = 1$. In this case, the input correlation $C(D, D^g) = 0.139$ is small and, therefore, reconstruction fails: $C(\bar{D}, D) = 0.342$ and $C(\bar{D}, D^g) = 0.025$. For the picture at the upper right panel of fig. 3, one has $p = 0.5$, for which $C(D, D^g) = 0.462$. The situation is similar since also in this case reconstruction fails: $C(\bar{D}, D) = 0.397$ and $C(\bar{D}, D^g) = 0.341$.

The scenario changes for the picture at the lower left side of fig. 3 with $p = 0.2$ for which $C(D, D^g) = 0.720$.

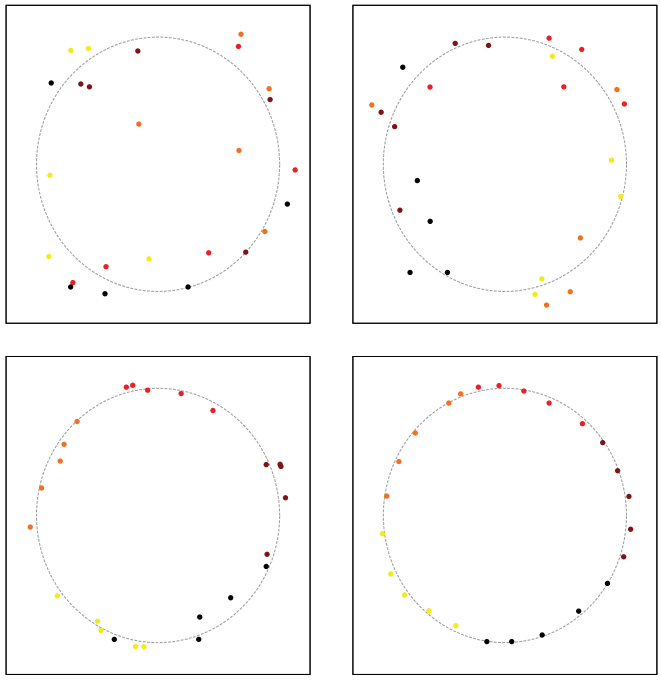

Fig. 3: (Color online) Reconstructed geographical positions $[\bar{x}_i, \bar{y}_i]$ for a a population of $N = 25$ individuals whose real geographical positions are equally spaced on a unitary circle. The four reconstructed geographies are related to a model with parameters $\gamma = 0.001, q = 0.2$ and $p = 1$ (upper-left), $p = 0.5$ (upper-right), $p = 0.2$ (lower-left), $p = 0.04$ (lower-right).

Our approach is able to identify the geometry since the output correlations are strong enough, $C(\bar{D}, D) = 0.753$ and $C(\bar{D}, D^g) = 0.920$. Finally, for the lower right side of fig. 3, with $p = 0.04$ and $C(D, D^g) = 0.916$, reconstruction is very accurate. In this last case, in fact, $C(\bar{D}, D) = 0.926$ and $C(\bar{D}, D^g) = 0.988$.

Notice how, for the last two cases, correlation $C(\bar{D}, D^g)$ between reconstructed distances and geographical distances is larger than correlation $C(D, D^g)$ between lexical

distances and geographical distances. For example, for $p = 0.2$ we have a correlation $C(D, D^g) = 0.720$ between the matrix entries and geographical distances while $C(\bar{D}, D^g) = 0.920$. Given that the $\bar{D}_{ij}$ are obtained only from the D_{ij} the result is somehow unexpected and it means that the method is able to extract the geometry which is hidden in lexical distances.

Conclusions. – We propose a feasible method for reconstructing the geographic positions of individuals only using genetic (or linguistic) distances available from genetic or linguistic data that are not *a priori* Euclidean. Our purpose is to extract the geographical information about distances on a surface by optimizing the embedding of complex genetic and lexical relations into a physical landscape.

We have strong evidence that recovering of the geographical positions from genetic or linguistic data is successful when horizontal transfer processes such as horizontal gene transfer between neighbors or vocabulary borrowings between different languages play a major role. When the process is purely phylogenetic (vertical), the correlation between geographical and genetic distances is absent and the reconstruction of geography from genetic or linguistic data fails.

We think that our method could be useful in linguistics as a complementary tool with respect to the phylogenetic approach. Representing the members of a linguistic family in terms of positions on a plane gives some information that is neglected by a tree representation and vice versa. We argue that the geographical approach could be especially useful when the languages of a family continuously modify one into the other as for example Romance languages where borders are artificial and mostly politically motivated.

* * *

We acknowledge Michele Pasquini for suggestions and advice concerning the numerical simulation and numerical optimization we have used in this research. We also thank Philippe Blanchard and Filippo Petroni for comments on the different technical and conceptual aspects of our work.

REFERENCES

[1] Cavalli-Sforza L. L. and Edwards A. W. F., *Evolution*, **21** (1967) 550.
[2] Nei M., Tajima F. and Tateno Y., *J. Mol. Evol.*, **19** (1983) 153.
[3] Hudson R. R., *Oxford Surv. Evol. Biol.*, **7** (1990) 1.
[4] Tagezaki N. and Nei M., *Genetics*, **144** (1996) 389.
[5] Excoffier L., Novembre J. and Schneider S., *J. Hered.*, **91** (2000) 506.
[6] Murphy W. J., Eizirik E., Johnson W. E., Zhang Y. P., Ryder O. A. and O'Brien S. J., *Nature*, **409** (2001) 614.
[7] Fernández M. H. and Vrba E. S., *Nature*, **80** (2007) 269.
[8] Prasad A. B. and Allard M. W., *Mol. Biol. Evol.*, **25** (2008) 1795.
[9] Swadesh M., *Proc. Am. Philos. Soc.*, **96** (1952) 452.
[10] Vérin P., Kottak C. P. and Gorlin P., *Ocean. Linguist.*, **8** (1969) 26.
[11] Gray R. D. and Jordan F. M., *Nature*, **405** (2000) 1052.
[12] Gray R. D. and Atkinson Q. D., *Nature*, **426** (2003) 435.
[13] Heggarty P., *Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data and to dating language?*, in *Phylogenetic Methods and the Prehistory of Languages*, edited by Forster P. and Renfrew C. (McDonald Institute for Archaeological Research, Cambridge) 2006, p. 183.
[14] Serva M. and Petroni F., *EPL*, **81** (2008) 68005.
[15] Petroni F. and Serva M., *J. Stat. Mech.* (2008) P08012.
[16] Starostin G., *J. Lang. Relationsh.*, **3** (2010) 79.
[17] Petroni F. and Serva M., *J. Stat. Mech.* (2010) P03015.
[18] Petroni F. and Serva M., *Physica A*, **389** (2010) 2280.
[19] Petroni F. and Serva M., *J. Quantit. Linguist.*, **18** (2011) 53.
[20] Sokal R. and Michener C., *Univ. Kansas Sci. Bull.*, **38** (1958) 1409.
[21] Saitou N. and Nei M., *Mol. Biol. Evol.*, **40** (1987) 406.
[22] Kalinowski S. T., *Heredity*, **102** (2009) 506.
[23] Prignano L. and Serva M., *Eur. Phys. J. B*, **69** (2009) 455.
[24] Cavalli-Sforza L. L., Menozzi P. and Piazza A., *The History and Geography of Human Genes* (Princeton University Press) 1994.
[25] Ramachandran S., Deshpande O., Roseman C. C., Rosenberg N. A., Feldman M. W. and Cavalli-Sforza L. L., *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005) 15942.
[26] Elhaik E., Tatarinova T., Chebotarev D., Piras I. S., Calò C. M., De Montis A., Atzori M., Marini M., Tofanelli S., Francalacci P., Pagani L., Tyler-Smith C., Xue Y., Cucca F., Schurr T. G., Gaieski J. B., Melendez C., Vilar M. G., Owings A. C., Gómez R., Fujita R., Santos F. R., Comas D., Balanovsky O., Balanovska E., Zalloua P., Soodyall H., Pitchappan R., Arunkumar G., Hammer M., Matisoo-Smith L., Spencer Wells R. and The Genographic Consortium, *Nat. Commun.*, **7** (2016) 13468.
[27] Serva M., Petroni F., Volchenkov D. and Wichmann S., *J. R. Soc. Interface*, **9** (2012) 54.
[28] Serva M., *PLoS ONE*, **7** (2012) e30666.
[29] Blanchard Ph. and Volchenkov D., *Mathematical analysis of urban spatial networks*, in *Springer Series: Understanding Complex Systems*, Vol. **XIV** (Springer) 2009.
[30] Blanchard Ph., Petroni F., Serva M. and Volchenkov D., *Comput. Speech Lang.*, **25** (2011) 679.
[31] Borg I. and Groenen P., *Modern Multidimensional Scaling: Theory and Applications*, 2nd edition (Springer-Verlag: New York) 2005, pp. 207–212, ISBN 0-387-94845-7.
[32] Derrida B. and Peliti L., *Bull. Math. Biol.*, **53** (1991) 355.

[33] Higgs P. G. and Derrida B., *J. Phys. A*, **34** (1991) L985.
[34] Derrida B. and Jung-Muller B., *J. Stat. Phys.*, **94** (1999) 277.
[35] Kingman J. F. C., *Stoch. Process. Appl.*, **13** (1982) 235.
[36] Serva M., *Phys. A*, **332** (2004) 387.
[37] Serva M., *J. Stat. Mech.* (2005) P07011.
[38] Serva M., *J. Stat. Mech.* (2006) P10013.
[39] Simon D. and Derrida B., *J. Stat. Mech.* (2006) P05002.
[40] Kingman J. F. C., *J. Appl. Probab.*, **19A** (1982) 27.