A stochastic local search approach to language tree reconstruction*

Francesca Tria, Emanuele Caglioti, Vittorio Loreto and Andrea Pagnani
Institute for Scientific Interchange (ISI) / Sapienza Università di Roma / Sapienza Università di Roma and Institute for Scientific Interchange (ISI) / Institute for Scientific Interchange (ISI)

In this paper we introduce a novel stochastic local search algorithm to reconstruct phylogenetic trees. We focus in particular on the reconstruction of language trees based on the comparison of the Swadesh lists of the recently compiled ASJP database. Starting from a generic tree configuration, our scheme stochastically explores the space of possible trees driven by the minimization of a pseudo-functional quantifying the violations of additivity of the distance matrix. As a consequence the resulting tree can be annotated with the values of the violations on each internal branch. The values of the deviations are strongly correlated with the stability of the internal edges; they are measured with a novel bootstrap procedure and displayed on the tree as an additional annotation. As a case study we considered the reconstruction of the Indo-European language tree. The results are quite encouraging, highlighting a potential new avenue to investigate the role of the deviations from additivity and check the reliability and consistency of the reconstructed trees.

Keywords: phylogeny, stochastic methods, noise and horizontal transfer, trees

* We are highly indebted to Søren Wichmann for providing us the data of the ASJP consortium, more recently made available as Wichmann et al. (2010), as well as for inspiring discussions. We also wish to thank Simone Pompei for many interesting discussions and suggestions. In addition, VL wishes to warmly thank Cinzia Mortarino and Sergei Starostin with whom he first got interested years ago in the problem of using Swadesh lists for phylogenesis. This research has been partly supported by the TAGora and ATACD projects funded by the European Commission under the contracts IST-34721 and 043415.

Diachronica 27:2 (2010), **341–358**. DOI 10.1075/dia.27.2.09tri
ISSN 0176–4225 / E-ISSN 1569–9714

1. Introduction

The reconstruction of phylogenetic trees belongs to a general class of inverse problems whose relevance is now well established in many different disciplines ranging from biology to linguistics and social sciences (Felsenstein 2004). In a generic inverse problem one is given a set of data and has to infer the most likely dynamical evolution processes that presumably produced the given data set. Historical linguistics (Renfrew et al. 2000) represents a clear example of an inverse problem. In this case the available data sets are list of homologous (lexical, phonological, syntactic) features or characters for many different languages: a parallel corpus whose compilation represents a paramount achievement in linguistics. In the 1950s Swadesh (1952, 1955) first proposed an approach to comparative linguistics that involves the quantitative comparison of lexical cognates, an approach named lexicostatistics. Since then many different approaches have been proposed to infer phylogenies (Felsenstein 2004) to overcome the difficulties related to this task. The main issue in inferring phylogenies concerns how to cope with the deviations of the underlying evolutionary process from a purely phylogenetic process, i.e., a process correctly represented by a tree. Mathematically one speaks of deviations from additivity of the phylogenetic tree. Additivity is a specific property of the distance matrix between taxa. Below we explain in detail the notion of additivity. Here recall that the sources of deviations from a purely phylogenetic process of the evolution of languages (i.e., the deviations from additivity of the associated distance matrix) are manifold: borrowings from other languages (a phenomenon dubbed horizontal transfer in evolutionary biology), inhomogeneous mutation rates of different characters, and the high probability, especially on extremely long phylogenies, that a given character may undergo multiple mutations. All these phenomena affect in different ways the natural evolution of languages concurring at a departure from the property of additivity.

Identifying the possible sources of non-additivity and their effects in a given data set is an open and challenging problem (Nakhleh et al. 2005). Additional difficulties are related to the accuracy of the data sets available as well as to the lack of suitable realistic benchmarks to be used to test the performances of the different algorithms. Also, when considering artificial benchmarks, the estimation of the performances can depend on the specific underlying artificial evolutionary scheme chosen.

In this paper we focus on distance-based methods whose basis is the computation of a suitable matrix of distances among all the taxa. This matrix is then typically analyzed by hierarchical clustering methods such as Neighbor-Joining (Saitou & Nei 1987), Fitch (Fitch & Margoliash 1967) or the more recent short-quartet method (Snir et al. 2008) and FastME (Desper & Gascuel 2002).

At the heart of our new algorithmic scheme are the notions of quartet and of quartet frustration. Both will be explained in detail later. Here it is enough to say that a quartet is a set of four taxa whose distances allow to define a mathematical relation

(i.e., the quartet frustration) that quantifies the deviations from a pure phylogenetic process. Quartets are thus a powerful mathematical tool to be exploited in novel algorithms able to cope with deviations from additivity. Following a concept already introduced in Snir et al. (2008), we propose to weight the different quartets according to their length in order to reduce the effect of those configurations that are more likely to have accumulated larger departures from additivity. It is important to stress here that we are not assuming any molecular clock hypothesis nor the independence of the rate of mutation of the different characters. We will show indeed that the performances of our algorithm on artificially generated dataset where neither molecular clock nor independent mutation rates are present, clearly outperform other state-of-the-art strategies.

In particulare we introduce a new algorithm belonging to the more general class of Stochastic Local Search (SLS) strategies (Hoos & Stützle 2005) that allow solving complex problems where the target is to find the best solution in a huge space of possible solutions to a given problem. Far from being random, the exploration of the solutions space is performed in a stochastic way driven by some insights on the structure of the space itself. We shall be more detailed in the following. Our new algorithm has been tested first on artificially generated phylogenies and finally on the reconstruction of the Indo-European language tree. In addition to the usual reconstruction we annotate the tree with two measures of violation of additivity and stability of the different tree partitions. The results are quite encouraging, highlighting a potential new avenue to investigate the role of the deviations from additivity and check the reliability and consistency of the reconstructed trees.

The outline of the paper is as follows. §2 discusses the violations from additivity and introduces the notion of four-points (or quartets) condition and the distance matrix based approach to calculate the tree length, which will be crucially used in our approach. §3 introduces our novel stochastic local search algorithm briefly reporting about its performances with respect to known state-of-the-art algorithms. §4 introduces the dataset used as well as the technique to compute the distance matrix. §5 reports our results for the reconstruction of our annotated language tree of Indo-European languages. Finally in §6 we discuss the relevance of our results and we draw some conclusions.

2. Violations of additivity

Additive trees can be characterized by a number of different, but equivalent, definitions that rely on the properties of the distances between taxa. Although different types of distance can be used in the context of languages, the following definitions do not depend on their specific choice. Let us consider a set of N taxa and the set of all their possible pairwise distances: these can be encoded in a $N \times N$ symmetric matrix D,

where the symmetry of the matrix reflects the symmetry of the distance between two taxa: the matrix element D(a,b) represents the distance between the taxa a and b and it is equal to the matrix element D(b,a) that represents the distance between b and a. The matrix D is said to be additive if any given pairwise distance D(a,b) can be recovered as the sum of the branch lengths in the path connecting the two taxa on the N-taxa tree. This definition is very intuitive, though it does not indicate any concrete strategy for deciding whether a given distance matrix D is additive or not: an interesting equivalent and more operative definition is the following.

2.1 Four-points condition

Let us consider a subset of four taxa, say a, b, c, d and let D be the corresponding $N \times N$ distance matrix. We can consider the following sums:

$$D_1 = D(a,b) + D(c,d),\ D_2 = D(a,c) + D(b,d),\ D_3 = D(a,d) + D(b,c). \quad (1)$$

The four taxa are said to satisfy the four-points condition if two of the above sums have the same and the greatest value. This is expressed in the following relations:

$$D_1 < D_2 = D_3 \text{ or } D_2 < D_1 = D_3 \text{ or } D_3 < D_1 = D_2. \quad (2)$$

A $N \times N$ distance matrix D is said to satisfy the four-points condition if the condition is satisfied by each possible group of four within N taxa. It is easy to realize that an additive distance (according to the definition in the previous subsection) satisfies the four-points condition. In order to show this statement, see Figure 1: here D_1 corresponds to the shortest of the three distances and it is indicated by the dashed blue path. The other two equal distances are those covering the path drown in red and are clearly larger than D_1, since they include the length of the link x.

Such a situation is general since it holds for any four taxa on a tree. The four-points condition provides a useful characterization of the additive property since, given a distance matrix, it is easy to scan all quartets (groups of four taxa) and to check whether the condition holds. When considering real data the additivity is typically always violated and so is the four-points condition. Therefore, in order to set up a robust method to reconstruct phylogenies based on the four-points condition, it is necessary to relax the requirements and to quantify violations from additivity in a suitable alternative way. This leads to the definition of the soft four-points condition.

2.2 Soft four-points condition

Let us again consider the example illustrated in Figure 1: focusing on the four taxa a, b, c, d, the branch marked with an x splits the tree in such a way that the taxa a, b are on one side while the taxa c and d are on the other. In the previous subsection we stated that, if D is additive, D(a, c) + D(b, d) = D(a, d) + D(b, c), and both D(a, c) + D(b, d) and

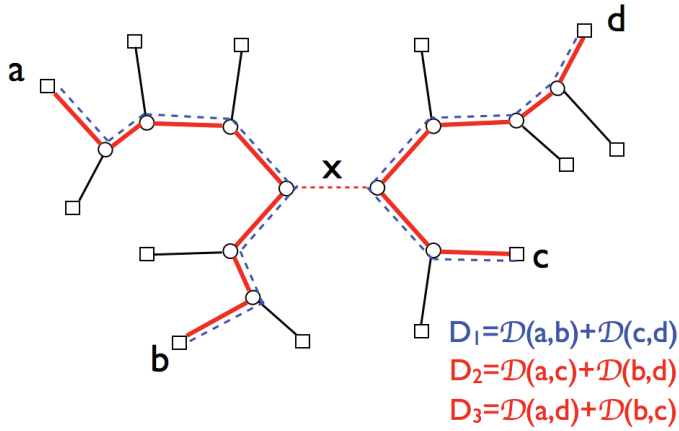

Figure 1. Definition of a generic quartet a,b,c,d. We refer to equation (1) and we consider an additive distance *D*. We represent directly the correct topology, i.e. a topology that does not violate the soft four-points condition (here also the four-points condition is not violated). We indicate with a blue dotted line the shorter distance D_1, and with a red line the two equal distances $D_2 = D_3$ (note that the two paths relative to D_2 and D_3 overlap completely). In this case, $D_2 = D_3 = D_1 + 2x$, where x is the length of the common edge transversed twice by both D_2 and D_3. If we considered a topology where the taxa b and c are swapped, the shorter distance D_1 is the sum of distances of couples of taxa, a and b, and c and d, where the elements of each couple do not lie in the same side of the tree; this configuration would thus violate the four-points condition and also the soft one.

D(a, d) + D(b, c) are greater than D(a, b) + D(c, d). In the soft four-points condition, we only require that the second relation holds. In practice we relax the requirement on the equality of two sums but we do require that the smallest sum correctly reflects the split of the tree: in our example we thus require that D(a, b) + D(c, d) is the smallest sum. We will say that a quartet is frustrated if it does not satisfy the soft four-points condition. Moreover, we can quantify the degree of dissatisfaction by introducing the quartet frustration. It is clear that, in order to check whether the soft four-points condition is satisfied, the distance matrix is not enough but one also needs a tree topology, which provides the split. Henceforth, we indicate with the shorthand notation (ab : cd) the situation in Figure 1, where a, b sit on one side of the tree and c, d on the other. Recalling the definitions of D_1, D_2, D_3 in Eq. 1, we thus define the quartet frustration, which quantifies the deviation from the situation in which D_1 is the lowest distance:

$$F_{(ab:cd)} = \max(0, D_1 - \min(D_2, D_3)) \tag{3}$$

The function max returns the maximum between 0 and the second argument, so that the frustration is always non-negative, and it equals zero when D_1 is the minimal distance, that is the soft four-points condition is satisfied. The function min returns the minimum of the two input arguments.

2.3 Calculation of the tree length from the distance matrix

We introduce a formula due to Pauplin (2000) whose computed value is equal to the tree length (defined as the sum of the lengths of all the tree branches) when the distance matrix considered is additive. In this formula the distances between any pair of taxa appear opportunely weighted. The rationale in using the distances between taxa to calculate the tree length is the following: when the distance matrix D is additive, the distance D(a,b) is the sum of the length of the branches of the tree in the path connecting a and b. Thus, if we consider all such paths, we cover the tree length more than once, and different branches are counted more than once. Pauplin's formula takes into account this effect by introducing weights in order to correctly estimate the tree length. The easiest way to introduce weights is the following: let us first define the topological distance τ(a,b) between any two taxa a and b as the number of nodes (or vertices) in the path connecting them. This number is equal to the number of branches in that path, minus one. Each distance D(a,b) is weighted by $2^{-\tau(a,b)}$.

A very simple example is described in Figure 2. In this case $D(a,b) = 4$, $D(a,c) = 3$, $D(b,c) = 3$, while the topological distance between every pair of taxa is 1, because one has to cross only one node to go from one taxa to any other.

Therefore Pauplin's distance is easily computed as:

$$L_P = D(a,b) \cdot 2^{-\tau(a,b)} + D(a,c) \cdot 2^{-\tau(a,c)} + D(b,c) \cdot 2^{-\tau(b,c)} = 4\tfrac{1}{2} + 3\tfrac{1}{2} + 3\tfrac{1}{2} = 5,$$

and we can immediately verify that it is equal to the length of the tree. In the general case, we can write:

$$L_P = \sum_{a \neq b} D(a,b) \cdot 2^{-\tau(a,b)} \tag{4}$$

where the sum runs over all the taxa. As stated above, Pauplin's formula equals the tree length only when the distance D is additive. The way of reweighting distances used in Pauplin's formula often makes it a particularly good approximation for tree length (Desper & Gascuel 2002) even for real data, where the condition of additivity is almost always violated. Furthermore, it is recognized that the minimization of Pauplin's

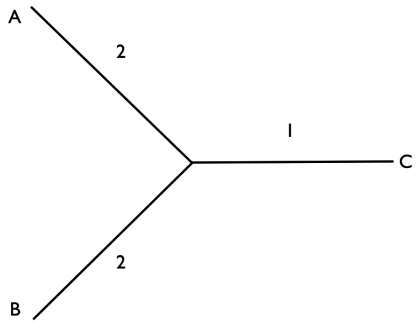

Figure 2. Definition of the Pauplin distance

formula, known as the Balanced Minimum Evolution (BME) criterion, is often a good criterion for reconstructing the correct phylogeny. This principle is used in an implicit way in the widely used Neighbor-Joining (Saitou & Nei 1987, Mihaescu et al. 2009) and more explicitly in a new generation distance-based algorithm, FastME (Desper & Gascuel 2002). When the departure from additivity is too strong, however, L_p no longer represents a good functional to be minimized in order to recover the correct tree. We introduce a novel distance-based algorithm for phylogeny reconstruction that crucially exploits the soft four-points condition as well as the notion of Pauplin's distance. We show that by combining these two ideas we end up with a method that is more robust against deviations from additivity and leads to a better inference of the correct phylogeny in a wide number of cases.

3. A stochastic local search strategy

3.1 Background

In this section we describe our algorithm. We use a Stochastic Local Search (SLS) strategy (Hoos & Stützle 2005) that allows us to find a tree topology that is particularly good with respect to our criterion (defined below). SLS algorithms have been widely used in solving complex combinatorial optimization problems such as Satisfiability, Colouring, MAX-SAT, and the Traveling Salesman Problem (Hoos & Stützle 2005), because of their ability to search in complex configurations spaces. The configuration space is typically parametrized in terms of a quantity called energy. In the context of our problem, energy is calculated as a function of the previously introduced quartet frustration measure. The Stochastic Local Search procedure works as follows: one starts from a randomly chosen initial configuration, that is, in our case, a random initial N-taxa tree topology. The initial configuration does not affect the performance of our algorithmic scheme. Starting form the initial configuration, one moves from the actual configuration to a neighboring one, where the notion of nearness has to be defined. Each move, i.e. each update of the tree topology, is determined by a decision based on local knowledge only and the decision is taken in a probabilistic way.

3.2 Updating the tree topology

Let us first define the notion of nearness, referring for simplicity to Figure 3. We focus on an internal edge (i.e., an edge not ending in a leaf) and refer to the four nodes neighboring the chosen edge as α, β, γ, δ. Let A, B, C, D be the four sub-trees rooted in α, β, γ, δ respectively. Our reference configuration is the one sketched in the left side of Figure 3, which we denote with the shorthand notation ((A,B),(C,D)), where the subtrees A and B lie on the left of the chosen edge and C and D on its right. This

configuration has two neighbors, represented in the right side of Figure 3: (i) ((A, D), (B, C)), where the subtrees A and D lie on one side of the chosen edge and B and C on the other, and (ii) ((A, C), (B, D)), where the subtrees A and C lie on the same side of the chosen edge and B and D on the other.

Updating the tree topology is the central step in the SLS algorithm we introduce. We can go from the present configuration to one of its neighbors according to the energy gain corresponding to this change. In particular, we first select randomly one of the two neighbors, then we calculate the energies of the present configuration and of the chosen neighbor. The energy is a function of the frustration of the quartets involved in the considered configuration and it will be defined in details below. Let us assume for a moment that we know these energies, which we call $E_{((A,B),(C,D))}$ and, say, $E_{((A,C),(B,D))}$. We change configuration with a probability proportional to the exponential of the energy difference: the probability of change is thus $e^{-\beta\Delta E}$, where $\Delta E = E_{((A,C),(B,D))} - E_{((A,B),(C,D))}$. The parameter β regulates how important the value of the energy difference is in the decision making process, and it is the equivalent to the inverse temperature parameter commonly used in statistical mechanics. For high values of β only neighboring configurations corresponding to lower energy with respect to the present configuration can be chosen. On the other hand for very low values of β one typically explores in a random way the configuration space. In our procedure, we start with a very low value of β (very high temperature) and progressively increase it until it reaches very high values (a procedure known as simulated annealing in statistical physics).

At each time step, one edge is considered and one decides whether the associated present configuration has to be changed or not. When no convenient moves are available the algorithm stops and returns the resulting tree topology.

3.3 Definition of energy

We now only need to define the energy of a configuration. Let us focus on the configuration ((A, B), (C, D)) again referring to Figure 3. We write $a \in A$ if a is a taxa of the subtree A and similarly for the taxa in the other subtrees. To obtain the energy of a configuration ((A,B),(C,D)) we sum up the frustration of all the quartets ((a,b),(c,d)), such that $a \in A$, $b \in B$, $c \in C$ and $d \in D$, each one with a weight borrowed from Pauplin's formula and with a suitable normalization factor:

$$E_{((A,B),(C,D))} = \sum_{((a,b),(c,d))} \frac{F_{((a,b),(c,d))}}{(D_1 + \min(D_2, D_3))^k} 2^{-[\tau(a,\alpha) + \tau(b,\beta) + \tau(c,\gamma) + \tau(d,\delta)]}, \quad (5)$$

where $\tau(a,\alpha)$ is the topological distance between the taxa a and the internal node α as defined in Eq. 4, and analogously for the other taxa. $F_{((a,b),(c,d))}$ is defined in Eq. 3 and the expression at the denominator, $(D_1 + \min(D_2, D_3))^k$, corrects the tendency for long-quartets (i.e., quartets corresponding to high values of D_1, D_2 or D_3) to violate

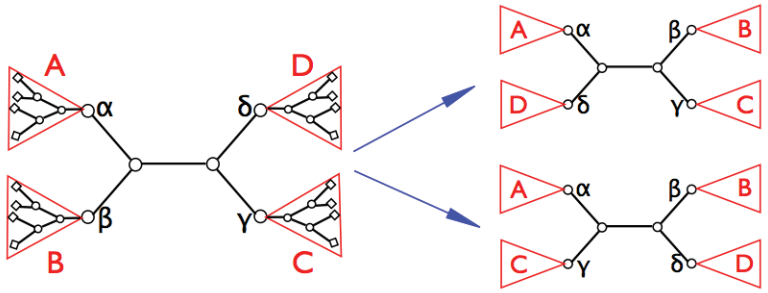

Figure 3. Illustration of the elementary move of the stochastic local search scheme.

the soft four-points condition by dampening their importance in the computation of the energy. The exponent k is a non-negative number, which regulates how strongly we dampen the effect of long quartets. In our results we use the value $k=5$, but we checked that the results are stable in a large range of values, going from $k=3$ to $k=10$. It is important to note that our algorithm does not automatically associate distances to the branches of the reconstructed topology. Once the topology is reconstructed, there exist different procedures to associate distances to the different branches. We adopted the least-square method implemented in the Fitch-Margoliash algorithm of the Phylip package (Fitch & Margoliash 1967). It is important to remark that the tree topology resulting from our algorithm is not independent of the distance matrix since our algorithmic procedure crucially depends on the evaluation of a energy-like functional computed in terms of the distances between taxa.

3.4 Results on artificially generated datasets

Note that our algorithmic scheme does not make any hypothesis about the evolutionary processes underlying the dataset used, e.g. an independent evolution of different characters. It is thus important to test its reliability and its performances in controlled situations. This is the aim of this section. We report the results concerning the performance of our algorithm as compared to other distance-based competing algorithms. To this end we use the following procedure: (i) We generate an artificial phylogeny by tuning the probability of back-mutation and horizontal transfer as a source of non-additivity, (ii) the correct-distance matrix calculated on the leaves of the generated tree is used as the input matrix given to the distance-based algorithms, and (iii) we compute the deviation of the inferred trees with respect to the true one using the standard Robinson Foulds measure (defined below). This is a crucial test since in this way we can compare the inferred phylogeny with the true one, which is in this case known. The evolutionary model we use to create the benchmark phylogenies is the simplest one taking into account both mutational and horizontal transfer events. We represent each taxon by a binary sequence of length L, but we check that more realistic representation,

i.e., q-state sequences, with $q > 2$, do not qualitatively change our results. A phylogeny is created as follows: we start with one sequence, for instance the sequence with all the bits equal to 0, from which N taxa are derived trough a branching process. At each step one of the leaves (nodes attached with only one branch to the tree) is selected. It then branches giving rise to two descendants. The iteration stops when an N-leaves tree is obtained. Superimposed on the branching processes are the mutational events. At each branching event, the two newborn descendants undergo two possible processes: (i) Mutation: with probability μ/L, and independently, each site of the sequence changes its value, where μ is the average number of mutations per sequence at each time step and (ii) horizontal transfer: with probability τ a part of the sequence of length $L/4$ is replaced with the corresponding part of another, randomly chosen, sequence of the tree (the value $L/4$ is arbitrary but again does not qualitatively affect the results).

We now introduce two matrices of distances. First we consider the Hamming distance between pairs of taxa, which is equal to the number of sites in which they differ. Under the assumption of independent and identically distributed (i.i.d) mutation probabilities on each site, in the limit of infinite length L of the sequence and without horizontal transfer, the average Hamming distance between two sequences after an evolution time t reads (Felsenstein 2004):

$$h = \frac{1}{2}\,(1 - e^{-2\mu t}). \tag{6}$$

Inverting this formula we can estimate the product μt, that defines the number of mutations that occurred between the two sequences. Based on this, we can define the corrected distance D_{corr} as:

$$D_{corr} = \frac{1}{2}\log(1 - 2h). \tag{7}$$

It is worth stressing that Eq. 6 expresses a relation between average values. The fluctuations around the mean value are much more important for the distance D_{corr} than for the Hamming distance h, so it is convenient to use the distance D_{corr} to infer phylogenies only when dealing with long enough sequences ($L > 1000$ in our evolutionary model for $\mu = O(1)$). When this condition is fulfilled, the use of this corrected distance D_{corr} turns out to give more accurate results. In order to assess the performances of the different algorithms to reconstruct the true phylogeny, we consider the standard Robinson-Foulds measure (RF) (Robinson & Foulds 1981) that counts the number of partitions (or splits) on which the inferred tree differs from the true one. Figure 4 illustrates the notion of the Robinson-Foulds measure. The two trees depicted allow for two partitions, represented by the transversal orange bars. One of the two partitions, e.g., the one separating the taxa (ed) from the taxa (abc) is present in both trees. On the other hand the partition on the left tree separating (bc) from (aed) is not present in the right tree. As a consequence the Robinson-Foulds measure of the two trees equals 1 in this case. In general one denotes as a positive partition a partition of the tree present in both the compared tree, while the opposite is true for a false positive partition. Note

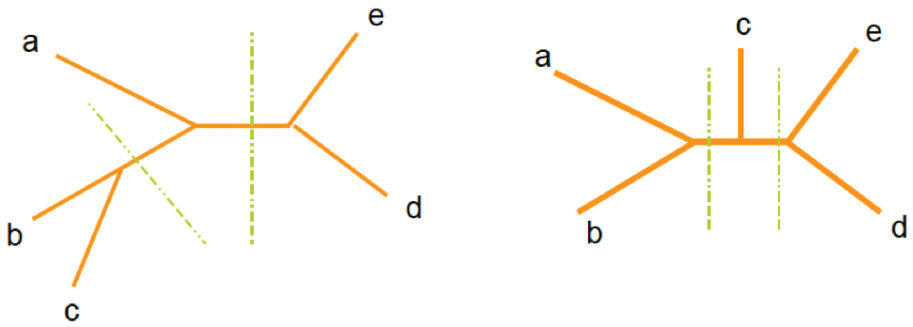

Figure 4. Illustration of the definition of the Robinson-Foulds measure. For the two trees considered RF = 1 since they differ of only one partition.

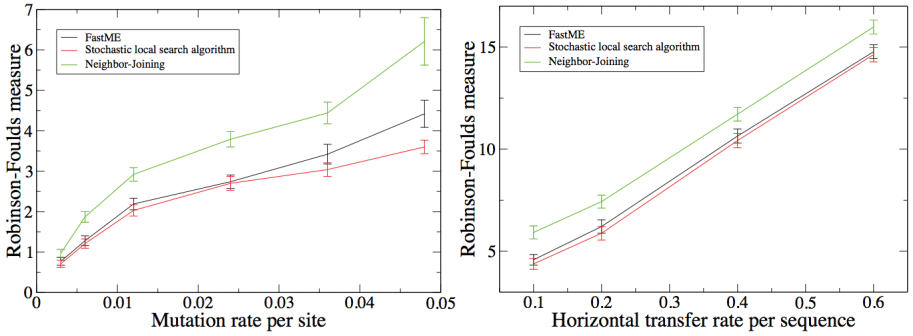

Figure 5. Comparison of the performances of our stochastic local search scheme with respect to the FastME and Neighbor-Joining algorithms. We show the Robinson-Foulds distance between the reconstructed and the true topology. Each point is an average over 100 samples from the generative model described in the text. Left: the horizontal transfer is absent ($\tau = 0$) and we study the performances of the three algorithms as a function of the mutation rate per site μ/L, where $L = 1000$ is kept fixed. Right: we fix $\mu = 30$ and $L = 10000$, and we study the performances of the three algorithms as a function of the rate τ of horizontal transfer. In both cases FastME outperforms Neighbor-Joining and our stochastic local search strategy outperforms FastME.

that for binary trees the number of the true positive partitions equals the number of the false positive partitions and both are equal to the Robinson-Foulds measure.

In Figure 5 we show the average RF distance of phylogenies reconstructed by NJ, FastME and our algorithm, where the average is performed over 100 samples generated as described above. The RF measure is shown as a function of μ/L when $\tau = 0$ and as a function of τ at fixed μ/L. We report here the results obtained by using the distance D_{corr} defined in Eq. 7. In both cases FastME outperforms Neighbor-Joining and our stochastic local search strategy outperforms FastME. We also checked that using the Hamming distance we still perform better than the other two algorithms but the performances are in this case collectively worse (data not shown).

4. The dataset used

The dataset we used for the present analysis is taken from the Automated Similarity Judgment Program (ASJP) (Brown et al. 2008, Holman et al. 2008, see: http://email.eva.mpg.de/wichmann/ASJPHomePage.htm). ASJP data orthography consists of 41 symbols, representing 7 vowels and 34 consonants, designed to cover all the commonly occurring sounds of the world's languages. The encoding is thus a phonetic one. We considered in particular the languages of the Indo-European group including 85 languages focusing on 40-items lists (see Table 1).

Table 1. List of the 40 meanings used.

I	Leaf	Knee	Star	You	Skin	Hand	Water	We	Blood
Breast	Stone	One	Bone	Liver	Fire	Two	Horn	Drink	Path
Person	Ear	See	Mountain	Fish	Eye	Hear	Night	Dog	Nose
Die	Full	Louse	Tooth	Come	New	Tree	Tongue	Sun	Name

Once we had fixed the 40-items lists for our set of languages, we defined the distance between two generic languages. For each list of meanings and for each pair of homologous words in the two lists, we computed the so-called edit distance (Levenshtein 1966) defined as follows. The edit distance between two strings, s1 and s2 is defined as the minimum number of point mutations required to change s1 into s2, where a point mutation is one of: (i) Change a letter, (ii) insert a letter, or (iii) delete a letter. In our case we normalized (Serva & Petroni 2008) the edit distance by the length of the longer of the two words compared. In this way the distance between two words is always in the range $[0:1]$. The total distance between two languages is thus computed as the average edit distance of all the homologous (and non empty) pairs.

5. Results

Figure 7 shows the annotated tree of Indo-European languages produced by our algorithm. Each internal branch has two numbers associated: the bootstrap measure and a measure of frustration. We give details of these measures in the following.

Bootstrap values An important question concerns the stability of the reconstructed tree with respect to small perturbations of the dataset. More exactly, we face the question whether minor changes in the word list produce a significant change in the reconstructed tree. To this end we introduce a bootstrap procedure by constructing 100 different lists of 35 meanings randomly selected out of the complete 40-items list. From each of these lists we construct a distance matrix in the way explained in §4 in such a way to obtain 100 different inferred trees of the Indo-European languages. The

trees are all inferred by making use of our stochastic local search algorithm. Now we can compare these trees with the one constructed using the complete dataset of 40 meanings. We count in particular in how many of the 100 reconstructed trees using 35 meanings each partition of our reference tree appears. This gives a bootstrap number associated to each partition of the reported tree.

A frustration measure A natural measure of frustration arising from our algorithm is the energy associated to each internal link in the local search strategy. This is in fact a measure of how much the soft four-points condition is violated due to the considered partition, each quartet being weighted by its Pauplin distance-like weight. The frustration measure is thus associated with the internal branches of the tree and it corresponds to E defined in §3.3. Figure 6 reports the values of the frustration measure vs. the bootstrap values. Interestingly, we note that the obtained bootstrap values are strongly correlated with the frustration values: branches which are stable with respect to the bootstrap measure show also little or no frustration. The Pearson coefficient (Wilcox 2005) between the two quantities turns out to be −0.786 (a value of 1 implying that all data points lie on a line for which one variable decreases as the other increases). This result implies that the p-value relative to the hypothesis that the two variables are uncorrelated is smaller than 10^{-16}. We show this correlation in Figure 6.

We finally note that the tree reported in Figure 7 features also non-binary divisions. This is due to the existence of branches of zero length in the reconstructed topology. In these cases we decided to show non-binary branching. These zero-length branches were also associated with very small bootstrap values as well as high violations of the four-points condition.

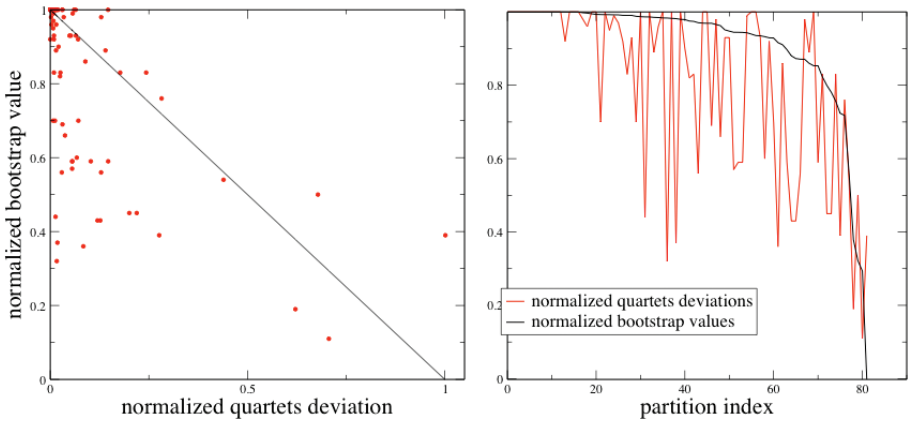

Figure 6. Correlation between the deviations from additivity and the bootstrap values. We normalized both the bootstrap values and the quartets deviations in order to reduce them on the same [0 : 1] interval. Left: normalized bootstrap values vs. normalized quartets deviations. Right: For each partition we plot both the normalized bootstrap value and the normalized quartets deviation. The overlap between the two curves is very good.

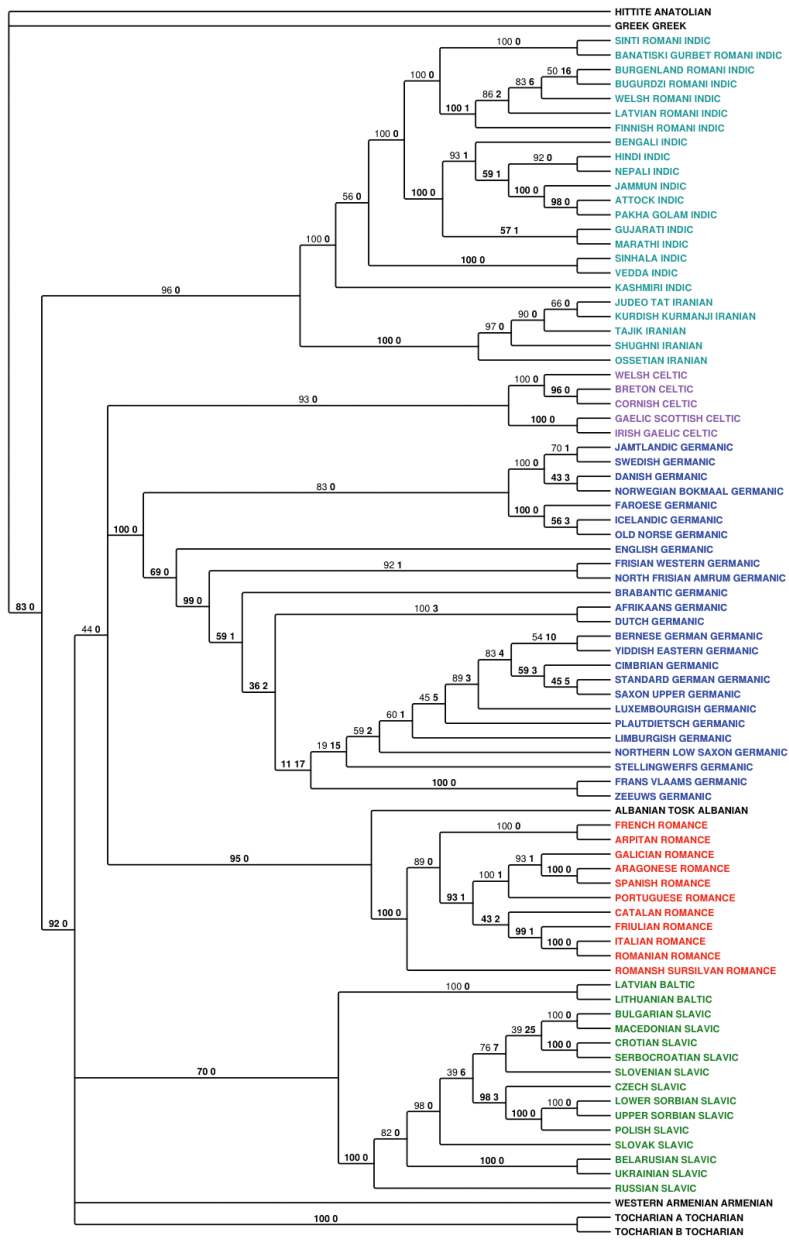

Figure 7. Annotated Indo-European language tree. Each internal branch has two numbers associated: the bootstrap measure (left) and a measure of frustration (right). The frustration measure shown on each branch is the E defined in §3.3 (multiplied by a factor of 10000 for graphical reasons). The higher the bootstrap measure and the lower the frustration measure the stronger the reliability of the specific internal branch. See the text for more comments.

6. Discussion and Conclusions

A few remarks about the tree reported in Figure 7 are in order. Our tree correctly reproduces the clustering of the major subgroups normally recognized: Romance, Germanic, Celtic, Balto-Slavic, Indo-Iranian. In order to make our tree more comparable with the Indo-European tree reported in Gray & Atkinson (2003) we also included extinct and old languages which lack modern descendants. Ancient languages with modern descendants should be correctly represented by internal nodes and not as leaves. An interesting area of research is represented by the reconstruction of trees when some additional information is available. Suppose for instance one knows that Latin is the extinct predecessor of all Romance languages. In the phylogenetic tree Latin should be more conveniently represented as an internal node instead of as a leaf. This is a general problem arising whenever the identity of some internal nodes (typically referred as Hypothetical Taxonomic Units) is known. Even more generally the information available could concern different aspects such as correlations among languages or the knowledge of the structure of specific subtrees. In all these cases it would be desirable to have suitable algorithms to fully exploit the information available. Again for the sake of comparison with the tree presented in Gray & Atkinson (2003) we outgrouped the tree on the Hittite language. The two trees are very similar although they exhibit some important differences, in particular the different localization of the Albanian Tosk on whose actual position there is not yet a shared consensus. The two trees differ also on the early divergence of Tocharian. We also obtained preliminary results for the estimate of the time divergence of the different branching but a more detailed account of these results will be presented in a forthcoming paper.

The results presented in this paper bring strong support to a new generation of algorithms (Tria et al. 2009) aiming at the reconstruction of language trees (and more generally of phylogenetic trees) while assessing their stability and reliability. In particular the measures of deviations from pure additivity are particularly promising since they allow for a quantification of the accuracy of the reconstruction possibly giving interesting hints about the underlying linguistic processes that have occurred. Obviously more systematic work is in order to assess the performances of our algorithm as well as the quality and the stability of the reconstructed trees on very different datasets. Crucial from this point of view is the nature and quality of the dataset used. For instance, it is still an open question how the words to be compared are best represented in terms of phonological transcription procedures. Also the choice of the list of meanings plays a very delicate role since the results could be strongly affected by the length of the list itself. A longer list presents obvious benefits from the statistical point of view. In addition, since the accuracy of correction to the Hamming distance (as defined in Eq. 7) turns out to increase for longer sequences, a longer list of meanings may in principle allows for a better estimate of the correction and a better reduction of the fluctuations associated with it. On the other hand a longer list increases the chances of inclusion

of borrowings, i.e., meanings with non-phylogenetic evolutionary histories, which can bias the reconstruction in an unpredictable way.

Another important research line we intend to pursue is that of looking for better functionals that could be sensitive to specific deviations from additivity, possibly disambiguating between horizontal transfer processes, heterogeneities of the evolutionary rates, back-mutation processes, etc. This research may lead to the development of new algorithms and strategies to correctly reconstruct and represent imperfect phylogenetic evolutionary processes.

References

Brown, Cecil H., Eric W. Holman, Søren Wichmann & Viveka Velupillai. 2008. "Automated classification of the world's languages: A description of the method and preliminary results". *STUF Language Typology and Universals* 61:4.285–308.

Desper, Richard & Olivier Gascuel. 2002. "Fast and accurate phylogeny reconstruction algorithms based on the Minimum-Evolution Principle". *Journal of Computational Biology* 9.687–705.

Felsenstein, Joseph. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates Inc.

Fitch, Walter M. & Emanuel Margoliash. 1967. "Construction of phylogenetic trees". *Science* 155:760.279–284.

Gray, Russell D. & Quentin D. Atkinson. 2003. "Language-tree divergence times support the Anatolian theory of Indo-European origin". *Nature* 426:6965.435–439.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller & Dik Bakker. 2008. "Explorations in automated language comparison". *Folia Linguistica* 42:2.331–354.

Hoos, Holger H. & Thomas Stützle. 2005. *Stochastic Local Search: Foundations and application*. San Francisco: Morgan Kaufmann.

Levenshtein, Vladimir I. 1966. "Binary codes capable of correcting deletions, insertions, and reversals". *Doklady Akademii Nauk SSSR* 163:4.845–848, 1965 (Russian). English translation in *Soviet Physics Doklady* 10:8.707–710, 1966.

Mihaescu, Radu, Dan Levy & Lior Pachter. 2009. "Why neighbor-joining works". *Algorithmica* 54:1.1–24.

Nakhleh, Luay, Don Ringe & Tandy Warnow. 2005. "Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages". *Journal of the Linguistic Society of America* 81:2.382–420.

Pauplin, Yves. 2000. "Direct calculation of a tree length using a distance matrix". *Journal of Molecular Evolution* 51:1.41–47.

Renfrew, Colin, April McMahon & R. Larry Trask. 2000. *Time Depth in Historical Linguistics*. Cambridge, UK: The McDonald Institute for Archeological Research.

Robinson, David R. & Leslie R. Foulds. 1981. "Comparison of phylogenetic trees". *Mathematical Biosciences* 53:1.131–147.

Saitou, Naruya & Masatoshi Nei. July 1987. "The neighbor-joining method: A new method for reconstructing phylogenetic trees". *Molecular Biology and Evolution* 4:4.406–425.

Serva, Maurizio & Filippo Petroni. 2008. "Indo-European languages tree by Levenshtein distance". *Europhysics Letters* 81.68005.

Snir, Sagi, Tandy Warnow & Satish Rao. 2008. "Short quartet puzzling: A new quartet-based phylogeny reconstruction algorithm". *Journal of Computational Biology* 15:1.91–103.
Swadesh, Morris. 1952. "Lexico-statistic dating of prehistoric ethnic contacts". *Proceedings of the American Philosophical Society* 96.452–463.
Swadesh, Morris. 1955. "Towards greater accuracy in lexicostatistic dating". *International Journal of American Linguistics* 21.121–137.
Tria, Francesca, Emanuele Caglioti, Vittorio Loreto & Andrea Pagnani. 2010. "A stochastic local search algorithm for distance-based phylogeny reconstruction". *Molecular Biology and Evolution*, doi:10.1093/molbev/msq154..
Wichmann, Søren, André Müller, Viveka Velupillai, Cecil H. Brown, Eric W. Holman, Pamela Brown, Matthias Urban, Sebastian Sauppe, Oleg Belyaev, Zarina Molochieva, Annkathrin Wett, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Robert Mailhammer & Helen Geyer. 2010. "The ASJP database (version 12)". http://email.eva.mpg.de/~wichmann/languages.htm.
Wilcox, Rand R., ed. 2005. *Introduction to Robust Estimation and Hypothesis Testing*. Oxford: Elsevier Academic Press.

Résumé

Dans cette contribution, nous introduisons un nouvel algorithme de recherche stochastique locale pour reconstruire des arbres phylogénétiques. Nous nous concentrons en particulier sur la reconstruction des arbres des langues basés sur la comparaison des listes de Swadesh de la base de données ASJP récemment compilée. En partant d'une configuration d'arbre générique notre algorithme explore stochastiquement l'espace des arbres possibles par la minimisation d'une pseudo-fonctionnelle qui quantifie les violations de l'additivité de la matrice des distances. Par conséquent l'arbre résultant peut être annoté avec les valeurs des violations sur chaque branche interne. Les valeurs des écarts sont fortement corrélées à la stabilité des branches internes, que l'on mesure avec une nouvelle procédure (*bootstrap*) et qui se voit affichée sur l'arbre comme une annotation supplémentaire. Comme exemple nous avons considéré la reconstruction de l'arbre des langues indo-européennes. Les résultats sont fort encourageants, mettant en lumière la possibilité d'une nouvelle voie pour examiner le rôle des déviations de l'additivité et vérifier la fiabilité et la cohérence des arbres reconstruits.

Zusammenfassung

In diesem Artikel wird ein neuartiger stochastischer lokaler Suchalgorithmus zur Rekonstruktion phylogenetischer Bäume vorgestellt. Wir konzentrieren uns dabei vor allem auf die Rekonstruktion von Sprachbäumen auf der Basis eines Vergleiches der Swadesh-Listen der kürzlich zusammengestellten ASJP-Datenbank. Ausgehend von einer generischen Baumkonfiguration untersucht unser Schema stochastisch den Raum möglicher Bäume, mit Orientierung auf die Minimierung einer Pseudo-Funktionalen, welche die Verletzungen der Additivität der Distanz-Matrix quantifiziert. Folglich kann der resultierende Baum mit den Werten der Verletzungen auf jedem internen Zweig annotiert werden. Die Werte der Deviationen korrelieren stark mit

der Stabilität der internen Kanten, sie werden mit einem neuartigen Boostrap-Verfahren gemessen und als zusätzliche Annotationen am Baum vermerkt. Als Fallstudie bieten wir die Rekonstruktion des indoeuropäischen Sprachbaumes. Die Resultate sind relativ ermutigend und weisen auf ein potentiell neues Verfahren zur Untersuchung der Rolle der Deviationen von Additivität und zur Überprüfung der Zuverlässigkeit der rekonstruierten Bäume.

Authors' addresses

Francesca Tria
Institute for Scientific Interchange (ISI)
Viale Settimio Severo 65, Villa Gualino
I-10133 Torino, Italy

tria@roma1.infn.it

Emanuele Caglioti
Dipartimento di Matematica
Sapienza Università di Roma
Piazzale Aldo Moro 5
00185 Roma, Italy

caglioti@mat.uniroma1.it

Vittorio Loreto
Dipartimento di Fisica
Sapienza Università di Roma
Piazzale Aldo Moro 5
00185 Roma, Italy
Institute for Scientific Interchange (ISI)
Viale Settimio Severo 65, Villa Gualino
I-10133 Torino, Italy

vittorio.loreto@roma1.infn.it

Andrea Pagnani
Institute for Scientific Interchange (ISI)
Viale Settimio Severo 65, Villa Gualino
I-10133 Torino, Italy

pagnani@roma1.infn.it