Evolutionary Bioinformatics

ORIGINAL RESEARCH

How Fitch-Margoliash Algorithm can Benefit from Multi Dimensional Scaling

Sylvain Lespinats[1], Delphine Grando[2], Eric Maréchal[2], Mohamed-Ali Hakimi[3], Olivier Tenaillon[1] and Olivier Bastien[2]

[1]UMR INSERM unité U722 and Université Denis Diderot—Paris 7, Faculté de médecine, site Xavier Bichat, 16 rue Henri Huchard, 75870 Paris cedex 18, France. [2]Laboratoire de Physiologie Cellulaire Végétale. UMR 5168 CNRS-CEA-INRA-Université Joseph Fourier, CEA Grenoble, 17 rue des Martyrs, 38054, Grenoble Cedex 09, France. [3]UMR5163, Laboratoire Adaptation et Pathogénie des Micro-organismes, Centre National de la Recherche Scientifique (CNRS), Université Joseph Fourier Grenoble 1, BP 170, 38042 Grenoble, Cedex 09, France.
Corresponding author email: olivier.bastien@cea.fr

Abstract: Whatever the phylogenetic method, genetic sequences are often described as strings of characters, thus molecular sequences can be viewed as elements of a multi-dimensional space. As a consequence, studying motion in this space (ie, the evolutionary process) must deal with the amazing features of high-dimensional spaces like concentration of measured phenomenon.
To study how these features might influence phylogeny reconstructions, we examined a particular popular method: the Fitch-Margoliash algorithm, which belongs to the Least Squares methods. We show that the Least Squares methods are closely related to Multi Dimensional Scaling. Indeed, criteria for Fitch-Margoliash and Sammon's mapping are somewhat similar. However, the prolific research in Multi Dimensional Scaling has definitely allowed outclassing Sammon's mapping.
Least Square methods for tree reconstruction can now take advantage of these improvements. However, "false neighborhood" and "tears" are the two main risks in dimensionality reduction field: "false neighborhood" corresponds to a widely separated data in the original space that are found close in representation space, and neighbor data that are displayed in remote positions constitute a "tear". To address this problem, we took advantage of the concepts of "continuity" and "trustworthiness" in the tree reconstruction field, which limit the risk of "false neighborhood" and "tears". We also point out the concentration of measured phenomenon as a source of error and introduce here new criteria to build phylogenies with improved preservation of distances and robustness.

The authors and the Evolutionary Bioinformatics Journal dedicate this article to the memory of Professor W.M. Fitch (1929–2011).

Keywords: Fitch-Margoliash, Multi Dimensional Scaling, Least Square methods, Sammon's mapping, molecular phylogeny

Evolutionary Bioinformatics 2011:7 61–85

doi: 10.4137/EBO.S7048

This article is available from http://www.la-press.com.

1. Introduction

Phylogenetic trees building methods

Since the early works of Hitchcock and Hitchcock (1840),[1] Darwin (1859, the only one illustration of the book)[2] and those of Ernst Haeckel (1866),[3] species are classified within trees. This idea has never been challenged and is supported by the evolutionary theory, which states that species genotypes/phenotypes drift from a common ancestor according to phylogenic proximity of the considered species. Trees are thereby convenient tools to express phylogenic relationships.

However, evolution has not been observed *per se*, and we must infer it from contemporary information (see Brocchieri[4] for a review on phylogenetic methods). Phylogenetic trees are often achieved from aligned sequences using character-based methods. Maximum parsimony methods minimize the number of changes.[5] Maximum likelihood methods are based on models of substitution process.[6]

Molecular character based methods can be launched from genomic or proteomic sequences but also from other molecular features like structural or thermodynamic features.[7,8] Conversely, distance based methods are often less powerful,[9] but they can be applied to any kind of matrices (of possibly large size) collecting distances between items (whatever the distances used—from aligned sequences to any measurable features). For example, UPGMA (Unweighted Pair Group Method with Arithmetic mean)[10] iteratively aggregates the most similar items. UPGMA has been said to suffer of long-branch attraction due to the (frequently false) hypothesis of the molecular clock (every sequence is supposed to evolve at the similar rate). In order to solve this problem, Neighbor Joining (NJ)[11] generates trees with the minimum cumulative branch lengths. We highly recommend the use of character based methods (such as maximum likelihood, minimum evolution, maximum parsimony, Bayesian approaches; they are more powerful[9] when the related hypotheses are valid). However, in some difficult cases, distance based methods are the only option[4,12] in large scale comparisons[13] and analyses on oligonucleotides frequencies (the so-called "genomic signatures"),[14,15] etc.

Note that every distance method assumes that a tree that preserves distances reflects the evolution between data. Fitch and Margoliash[16] followed this concept and described a method designed to generate a tree that preserves distances at best. The tree that minimizes squared difference between original and resulting distances is assumed to express the evolution of considered genes (or proteins, if it is the case). Here, we hold on this widely shared hypothesis.

Biological sequences and amino acids are clearly multi-dimensional objects.[17–20] It has been recently demonstrated that the high-dimensionality of biological sequences leads to emergent computational features like similarity measure and particular probability distribution of this similarity.[21–23] Distance methods are the first approach for phylogeny reconstructions and consist merely in considering the data as a point in an η-dimensional phase space (where η is in first approximation the length of the sequence). If one knows the relative position of sequences (distance matrices) and the velocity modulus corresponding to the sequence evolution, it should be possible to reconstruct the divergence history of the sequences (which corresponds to a trajectory in the η-dimensional space). However, starting from a point of this phase space, another point cannot be reached by a simple straight line because only a few likely paths between the two points in this phase space are allowed (Biological features of sequences must be preserved all along the path).[24] The topology of all *a priori* possible evolution process space has an E^{η} topology, where E is the space of the states of one residue position. Nevertheless, the topology of actual evolution process is clearly unknown, since the E^{η} subspace where the process occurs is also unknown. This is one of the reasons why maximum likelihood and parsimony methods are usually more accurate than distance methods. Indeed, ML and parsimony do take the intrinsic high-dimensional nature of the original data (ie, the sequences) into account, by modeling resp. optimizing character state changes. Conversely, distance-based methods such as Neighbor Joining or Fitch-Margoliash algorithm (which do not account for the data dimensionality) are commonly performed from genomic signature comparisons, very large data sets, such as DNA barcode libraries, or indeed large-scale genomic comparisons, and so on. However, these often are high dimensional data and we will show in the following why Fitch-Margoliash algorithm could be misled by features of high dimensional space, and how modifying it to avoid these traps.

Multi dimensional scaling

The general aim of the Multi Dimensional Scaling methods is the dimensionality reduction. The most used technique is by far the Principal Components Analysis (or PCA).[25,26] To obtain this representation, data are projected (by orthogonal projection) on a selected vector subspace according to a criterion equivalent to the maximization of distances preservation.[27] The "Multi Dimensional Scaling" (MDS) term was extracted from a method that allowed a mapping from a distance matrix[27] (this method is known thereafter as "Classical Multi Dimensional Scaling" or Classical MDS). When dealing with Euclidean distances, Classical MDS can be seen as a linear projection method. However, non-linear projection can be achieved as well. For that purpose, the preservation of small distances is favored through a weighting system. The Sammon's mapping[28] is one of the first non-linear MDS methods. Many MDS techniques have been developed since then, including Curvilinear Components Analysis (or CCA),[29] Kernel Principal Component Analysis (or KPCA),[30] Isomap,[31] Curvilinear Distances Analysis (CDA)[32] and Generative Topographic Mapping (GTM).[33]

We focus here on a new efficient method called Data-Driven High Dimensional Scaling (DD-HDS)[34] which specificities are i) the penalization of "false-neighborhoods" and "tears" and also ii) the consideration of the concentration of measured phenomenon. As many MDS methods (including Sammon's mapping, CCA and Local-MDS) DD-HDS is driven by a given criterion in order to iteratively converge toward a satisfying map (several examples of criteria are given in the beginning of section 3).

The present paper shows how close are Multi Dimensional Scaling and tree building methods fields and extracts from Multi Dimensional Scaling new ways to improve actual tree building methods from distance matrices. The article is organized as follows: Section 2 compares the Fitch-Margoliash algorithm with Multi Dimensional Scaling methods (with a special attention for Sammon's mapping)[28] Section 3 describes the possible contribution of the recently introduced Multi Dimensional Scaling methods to tree building methods through a new criterion. Section 4 sums up the resulting algorithm, used to generate results presented and analyzed in Section 5.

2. Existing links between Fitch-Margoliash and Sammon's Mapping criteria

Least Square (LS) methods are designed to generate a classification (through a tree) from distances between data. This distance matrix is noted d (d_{ij} is the element in the i^{est} row and the j^{est} column, which corresponds to the distance between item i and item j; d is then a $N \times N$ matrix where N is the number of items). In the field of molecular evolution, data are DNA (sometimes RNA) or protein sequences, represented as strings of characters, and distances are calculated between aligned sequences. Least Square methods purpose is to generate a tree that preserves distances between data "as much as possible" in the sense that it should minimize a selected criterion (eg, Eq. 1).

The general aim of Multi Dimensional Scaling is also to provide a configuration of points that preserves "as much as possible" distances between data (and a criterion is also selected to that purpose). Multi Dimensional Scaling can be then considered as a function that associates N items in a metric space to the original distance matrix. The output space is frequently a Euclidian two-dimensional space (which is a well-known metric space). However, any metric space could be obviously used here instead. Least Square methods (such as the Fitch-Margoliash method) can be then reformulated as Multi Dimensional Scaling where the distance in the output space is an additive distance (ie, a distance calculated on a tree).

In fact, striking parallel features can be drawn between Multi Dimensional Scaling and Least Square methods: inputs (distances matrices), purposes (preservation of distances "as much as possible") and outputs (configuration of items in an intuitive metric space) are similar. The only difference between Multi Dimensional Scaling methods (including Sammon's mapping) and tree building Least square methods (including the Fitch-Margoliash method) can be found in the "nature" of the output (and consequently, in the optimization methods). Moreover, we will show thereafter that Sammon and Fitch-margoliash's criteria are similar. Multi Dimensional Scaling methods might therefore provide a framework to better understanding criteria for tree building methods.

Note that we consider unrooted trees in the present paper. Moreover, branches with negative length, a common feature, are considered as

0-length branches,[35,36] distances in trees will have all the three mathematical properties of distances ($\forall$x,y,z, 1) d(x,y) = d(y,x); 2) d(x,y) = 0$\Leftrightarrow$x = y; 3) d(x,z)$\leq$d'x,y) + d(y,z)).[37] Note also that no hypothesis is made on the input distance; moreover all mathematical properties of distances are not required here (triangle inequality can possibly be transgressed). As a consequence, any dissimilarity could be considered (in tree reconstruction as well as in Multi Dimensional Scaling).[34]

Fitch-Margoliash criterion

Least Square methods are achieved in order to minimize difference between original distances and distances in the resulting tree. In particular, Fitch-Margoliash proposes a criterion (Eq. 1) to quantify the misfit between the original distance d and the distance matrix between items positioned in a given tree (we will designate this tree-distance matrix as ∂).

$$\varsigma = \sum_{i,j} w_{ij} \times \left(d_{ij} - \partial_{ij}\right)^2 \qquad (1)$$

where $w_{ij} = f(d_{ij})$. Several functions have been proposed for f (f is thereafter designed as the weighting function). Within the Least Square framework, w_{ij} is often considered to be proportional to the reverse of an estimation of the variance of d_{ij}, seen as an estimator of the "real" evolutionary distance between i and j.

Generally, f is chosen as $f_p(x) = 1/x^p$, where p is positive. For example, $w_{ij} = f_0(d_{ij}) = 1$ (*ie,* $P = 0$)[38] corresponds to the Ordinary Least Square methods (OLS). It assumes that every distance is measured with the same error rate. Conversely, by choosing a strictly positive value for p leads to Weighted Least Square methods (WLS), we can naturally consider that the larger the evolutionary distance, the poorer its estimation and then the weaker its weight. In that case, $w_{ij} = f_1(d_{ij}) = 1/d_{ij}$ could be considered appropriate (*ie,* $P = 1$). The most commonly used weight is $w_{ij} = f_2(d_{ij}) = 1/d_{ij}^2$ (*ie,* $P = 2$).[16] It is based on the observation that the variance of the error for the estimation of the evolutionary distance is often approximately proportional to d_{ij}^2 in simulated datasets (these datasets are aligned sequences following classical evolutionary models).[35] This estimation as been refined recently, to propose the following weight: $w_{ij} = f_{1.823}(d_{ij}) = 1/d_{ij}^{1.823}$ (ie, $P = 1.823$).[39]

Similarity between Fitch-Margoliash and Sammon's mapping criteria

The Sammon's mapping[28] has been designed to embed data in a low-dimensional space while preserving distances "as much as possible" with a special consideration for small distance. The method is known as one of the earliest non-linear projection proposed by John W. Sammon in 1969.

Striking parallel features can be drawn between Fitch-Margoliash and Sammon's mapping criteria (for $P = 1$ in Eq. 1 that corresponds to the case where $w_{ij} = 1/d_{ij}$). Indeed, Eq. 1 exactly expresses criterion that drives the Sammon's mapping if ∂ assigns distances in a Euclidean output space.

Multi Dimensional Scaling has been improved since it was first introduced (for an overview of the field, please report to Lee et al)[40] and many non-linear Multi Dimensional Scalings have outclassed Sammon's mapping. Progresses obtained in the field of non-linear mapping can be applied to Fitch-Margoliash method. In particular, we will focus on Data-Driven High Dimensional Scaling (DD-HDS),[34] an effective mapping method that follows the Sammon's mapping concept.

Finding the best configuration

While w_{ij} is chosen, (which lead to define the criterion ς, eq. 1), the best fitting tree has to be discovered. However, the number of possible tree topologies increases explosively with the number of items (number of possible topologies for N items = $(2N\text{-}5)!/[2 \times (N\text{-}2)!] = (2N\text{-}5) \times (2N\text{-}3) \times \ldots \times 5 \times 3$).[41] Of course, the exhaustive exploration is not an option, and finding the tree that best preserves distances is a NP-difficult problem.[42,43] Numerous algorithms have been proposed to approach this goal including genetic algorithm,[44–46] ant colony optimization[47] and TABU search.[48] In particular, the popular FITCH software from the PHYLIP package[49] iteratively introduces items (each possible place is tested for the new item, the best place from the criterion point of view is validated). After the introduction of each item, a "tree swapping" optimization process is run to reconsider the global organization of the tree: the Nearest Neighbor Interchanges (NNI).[50–52] Some alternative methods can be used in such framework rather than NNI. In particular, the "Tree Bisection-Reconnection" (TBR) and the "Subtree Pruning Regrafting" (SPR) can be used and

allow a wider tree reorganization.[51] However, such advantage is expensive in term of complexity: complexities are O(N) for NNI, O(N^2) for TBR and O(N^3) for SPR (N is the number of leaves).[53] For that reason, and to be mimetic with the original Fitch-Margoliash algorithm, we used NNI in the present work.

A corresponding problem occurred in Multi Dimensional Scaling. However, even when methods agree about the criterion that should be optimized, various ways can be followed to generate the best configuration of points in the output space. Among optimization methods used in this framework, we can cite Generalized Newton-Raphson algorithm,[28,54] TABU search,[55,56] genetic algorithms,[57,58] simulated annealing,[59,60] neural networks[29] and Force Directed Placement.[34,60–62]

3. Improvement of the Fitch-Margoliash Criterion

We have shown the proximity between Sammon's mapping and Fitch-Margoliash criteria. However, some drawbacks have been identified for the Sammon's mapping criterion. They obviously impact Fitch-Margoliash and have then to be avoided.

Penalization of "false neighborhoods" and "tears"

It has been found that Sammon's mapping poorly penalizes "false neighborhoods".[34,63–65] Indeed, let suppose that d_{ij} is high; $f(d_{ij})$ is small; $(d_{ij}-\partial_{ij})^2 \times f(d_{ij})$ does not contribute much to ς, even if ∂_{ij} is small. A case where d_{ij} is high and ∂_{ij} is small (which corresponds to a "false neighborhood") is then poorly penalized.

This situation can be avoided using the Curvilinear Components Analysis criterion[29] (Eq. 2) (the weight depends now on the distance in the representation space).

$$\varsigma = \sum_{i,j} (d_{ij} - \partial_{ij})^2 \times f(\partial_{ij}) \qquad (2)$$

Unfortunately, "tears" are now poorly penalized: if d_{ij} is small and ∂_{ij} is high (which corresponds to a "tear"), $(d_{ij}-\partial_{ij})^2 \times f(\partial_{ij})$ remains small and does not contribute much to ς. Note that, consideration of "tears" and "false neighborhoods" are sometimes addressed under concepts of "continuity" and "trustworthiness".[65,66]

However, "tears" and "false neighborhoods" can now be simultaneously penalized through a new criterion like the following:[34]

$$\varsigma = \sum_{i,j} (d_{ij} - \partial_{ij})^2 \times f(\min(d_{ij}, \partial_{ij})) \qquad (3)$$

$f_1(\min(d_{ij},\partial_{ij}))$ is high if and only if d_{ij} or ∂_{ij} is small: each small distance is considered as important (whatever if the distance is small in original or output space). This criterion proposed for Data-Driven High Dimensional Scaling has been found effective for avoidance of "tears" and "false neighborhoods" on several simulated and real datasets.

When considering trees, this analysis is still valid. In this context, two items that are connected whereas they are widely different is a "false neighborhood", when close items that are placed in different parts of the tree corresponds to a "tear". Naturally, within phylogenetic inference, both "false neighborhoods" and "tears" should be avoided. However, we demonstrated that Fitch-Margoliash (as well as Sammon's mapping) does not penalize efficiently "false neighborhoods". As a consequence, the consideration of criterion proposed in Eq. 3 could provide benefits in tree building like in Multi Dimensional Scaling.

Consideration on concentration of measure phenomenon (CMP)

The concentration of measure phenomenon

Dealing with high-dimensional data is complicated. Indeed, high-dimensional spaces have several surprising properties that discomfit human spirit (the famous "curse of dimensionality").[67,68] Among them, the concentration of measured phenomenon (illustrated in Fig. 1) is critical; it makes the relative difference between small and large distances tends to zero when the dimension increases.[68–71] Such phenomenon can be illustrated by the distribution of distances between data points randomly drawn in a hypercube according to the dimension of the space.[72]

The concentration of measured phenomenon has a main impact on Multi Dimensional Scaling: even if small distances are believed to be emphasized in the mapping (through the weighting function f_p defined in section 2.1), the weights for small and large distances could be very close.[15,34]

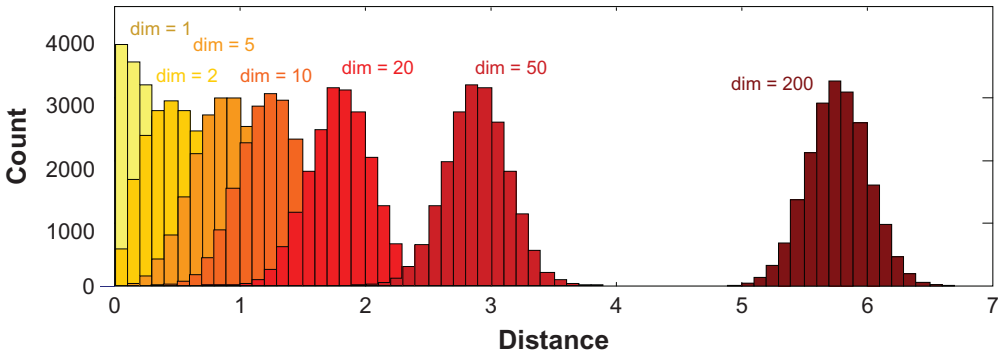

Figure 1. 200 random data are uniformly distributed in a unit cube (of a given dimensionality). Histograms of distances between every pairs of items (19 500 distances) are displayed according to space dimension. Distributions of distances for dimensions larger than 200 would have the same Gaussian-like shape, but their centers would be shifted to the right proportionally to the square root of the dimension.

As an illustration, let us considered two distances noted a and $a+\varepsilon$, where a is small and $a+\varepsilon$ is large. When the dimension of space increases, ε/a may tend to 0 ($\varepsilon/a \xrightarrow[n\to\infty]{} 0$). Then,

$$f_p(a+\varepsilon) = \left(\frac{1}{a+\varepsilon}\right)^p - \left(\frac{1}{a}\right)^p - f_p(a) = \left(\frac{\varepsilon}{a}\right)^p \times \sum_{i=1}^{p} \frac{\frac{p!}{i!\times(p-i)!}a^i\varepsilon^{p-i}}{(a+\varepsilon)^p} \xrightarrow[\varepsilon/a\to 0]{} 0 \quad (4)$$

Weights for a and $a+\varepsilon$ (noted $f_p(a)$ and $f_p(a+\varepsilon)$) are then very close for high-dimensional data. Here, p (the parameter of the weighting function f; in Fitch-Margoliash, p often equals 2) is supposed to be a strictly positive integer for the calculus simplicity (this could be generalized to $P \in [1,+\infty[$ using the gamma function).

Impact of CMP on phylogenic data

Proteins and DNA sequences are high dimensional objects (including when the only available material is the distance matrix between items).[17] The question of their dimensionality has been theoretically addressed through the question of intrinsic dimension (or degree of freedom) of data.[18,19] However, determining the dimensionality is not needed in the present framework. Still, the distance matrix may show properties close to ones of distances between high-dimensional data, ie, a low "relative contrast" (contrast can be defined as ([distance of the farthest to the origin of the vector space]—[distance of the nearest to the origin])/[distance of the nearest to the origin]) in the distance matrix.[69,73] In such case, the phenomenon described by equation 4 may occur: every distance has almost the same weight in contradiction with the spirit of Fitch-Margoliash method. In case of phylogenetic data, the origin of an eventual concentration of measured phenomenon could be discussed. However, such discussion can be avoided here, indeed we do not consider the dimensionality here, but the concentration of measure itself (whatever its origin).

The Multi Dimensional Scaling method DD-HDS[34] has treated this problem: a sigmoid function adapted on the distances distribution allows us to assign a high weight to smaller distances and a low weight to larger ones whatever the dimension. We propose to apply the same solution to Weighted Least Square methods: the weighting function f is modified according to equation 5.

$$f_{sigmo}(x,\lambda) = 1 - \int_{-\infty}^{x} g\left(u,\mu(\lambda),\sigma(\lambda)\right)du \quad (5)$$

where $g(u,\mu(\lambda),\sigma(\lambda))$ is the probability density function of a Gaussian variable with mean $\mu(\lambda)$ and standard deviation $\sigma(\lambda)$. $\mu(\lambda)$ and $\sigma(\lambda)$ are suggested to be:

$$\mu(\lambda) = \underset{i<j}{mean}(d_{ij}) - 2\times(1-\lambda)\times \underset{i<j}{std}(d_{ij}) \quad (6)$$

and

$$\sigma(\lambda) = 2\times\lambda\times \underset{i<j}{std}(d_{ij}) \quad (7)$$

where λ is a user-defined positive parameter, usually equal to 0.1, that reflects the trade-off between preservation of local versus global distances. λ can

be initially set at a high value (0.9 for example) and be decreased during the optimization, initially to take into account the global organization of data and progressively focus on local relationships.

Noticeably, the use of a Gaussian density function does not result from a hypothesis of Gaussian distribution of measured but only allows a continuous decrease of function f with high values for smaller distances and low values for larger distances. Equations 6 and 7 allow adjusting parameters μ and σ for equation 5; in case of an atypical distances distribution, the user can therefore adjust parameters.

Within the phylogenetic analysis framework, sequences are somehow related to each other and distances are usually expected to follow patterns. A similar argument applies to dimensionality reduction methods, regarding distances between data points (that are supposed to lie onto a low-dimensional manifold). However, we can observe that patterns are sometimes blotted out by the concentration of measure phenomenon when dealing with real data. Obviously, we cannot rely on this hypothesis in the context of tree reconstruction (of course, in the favorable case where patterns can be found, the goal will be reached more easily).

A new criterion

DD-HDS introduces a new criterion which is optimized $\varsigma = \sum_{i,j} \left| d_{ij} - \partial_{ij} \right| \times f_{sigmo}\left(\min\left(d_{ij}, \partial_{ij}\right), \lambda\right)$ (the absolute difference between distances allows to be consistent with an optimization process called Force Directed Placement that simulates a spring system, not used in the present tree building method). In order to develop a Weighted Least Square method (which is more common in tree building framework), we can use a corresponding criterion that rather considers square difference between distances (Eq. 8). For sure, close minimums should be found for these criteria. The optimized criterion is now:

$$\varsigma_{LS} = \sum_{i,j} \left(d_{ij} - \partial_{ij}\right)^2 \times \left(f_{sigmo}\left(\min\left(d_{ij}, \partial_{ij}\right), \lambda\right)\right)^2 \quad (8)$$

where λ is the user-defined parameter defined in the previous section. λ must be set between 0 (for maximally emphasizing of small distances) and 1 (where large distance still weight).

In Least Square methods, w_{ij} is often considered as the reverse of an estimation of the variance of d_{ij}, which is not the case for our criteria. Classically, w_{ij} is selected based on this perspective from observations on simulated data. These simulated data follow a definite form, since the alignment of artificial sequences is strictly based on assumption regarding the evolutionary model. When the user's data follow a similar model, such a criterion must be chosen. Conversely, when data do not follow the usual standard evolutionary models or when distances are not measured in the same way or when data are not aligned sequences etc, the classical Fitch-Margoliash criterion is not necessarily more suitable than any other one. In such cases, we gave important arguments to prefer using weights as presented in section 3, subsection "Consideration on concentration of measure phenomenon (CMP)". As a consequence, eq. 8 provides a model-free criterion (in the sense that it is based on no explicit model but it is designed to adapt to the distance distribution).

4. Algorithm

We exposed a new criterion and gave theoretical arguments to prefer its use instead of the classical Fitch-Margoliash's one. To test our assumptions, an algorithm using the new criterion is set up.

Optimization

Our algorithm mimics FITCH software[49] (except, naturally, for the criterion). Three randomly selected data are initially chosen as leaves of a 3-branch tree (one single possible topology). Remaining leaves are then randomly selected to be introduced one by one according to an iterative procedure: each position is tested for the new leaf; when the best position is chosen, the whole topology is challenged by testing branch permutations (the so called "tree swapping" procedure) according to the NNI (Nearest Neighbor Interchange) algorithm. When all leaves have been introduced, the procedure is stopped. The only difference between algorithms lies in the criterion (c.f. eq. 1 for example).

A modification is nevertheless necessary for the calculus of branch lengths. Originally, branch lengths depend on original distances and could then be analytically deduced from a given topology and chosen weighting simply using matrix algebra.[74] However,

weights can no more be fixed *a priori*, because they now also depend on distances in output tree. A slight adaptation using an incremental approach is then necessary. Firstly, weights are fixed as $w_{ij} = f_{sigmo}(d_{ij},\lambda)^2$ and branch lengths are calculated. We then get estimations for ∂_{ij}, which allows the modification of weights as $w_{ij} = f(\min(d_{ij},\partial_{ij}),\lambda)^2$ that generates new estimations for ∂_{ij} (each of these iterations is a linear operation). This process is iterated until stability of ∂_{ij} is reached. This modification does not strongly affect the calculus time. Indeed, few quasi-instantaneous steps (each one is a simple matrix algebra operation) are generally needed to converge.

The use of progressive reduction of λ (c.f. section 3.2), at worst multiplies the calculus time by the number of steps (here 2 steps). Overall these procedures increase the calculus time by less than a 2-fold factor (and, as a consequence, has no impact in term of complexity). It is worth to note that the overall complexity of the algorithm is $O(N^4)$, just as the classical Fitch-Margoliash (FITCH implementation).[75]

Expected benefits

If distances are additive there is a perfect solution to Least Square methods. Obviously, Fitch-Margoliash (just as most of classical methods like Neighbor Joining, etc) will solve this: no benefit can be expected in this ideal case. However, distances are often far from being additive, and it may be hard to reach the best tree in these cases.

Since additivity is a very restrictive property for distances, there are few chances to be in that case when dealing with real data. Fitch-Margoliash has shown efficiency in such a situation. However we highlight some drawback in this paper that may be a matter of concern. If, in addition, the distribution of distances is tight (which corresponds for example to distances in a high-dimensional space), we have shown that Fitch-Margoliash can be overtaken for theoretical reasons. As a consequence, better results can be expected in such case (and/or when the user's data do not follow classical evolution models, cf. section 3, subsection "A new criteria") because of a criterion that quantifies more appropriately the correspondence between the distances and the tree.

Moreover, during inference processes, the optimization to find the best tree is the major difficulty. We believe that the penalization of both "false neighborhoods" and "tears" will allow us to avoid the exploration of many poor solutions that were lightly penalized by classical criteria. As a consequence, the risk of reaching a local minimum is reduced: our method may therefore drive us more closely to the optimal tree.

Eventually, we propose the choice of a parameter λ that balances the matter of small versus large distances in order to allow a control by user.

5. Results

Methods and criteria

We evaluated trees within six criteria to test the benefits obtained by each one of our modifications (Fig. 2). The benefit supplied by every proposed modification can thus be sized up.

These criteria allow us to build trees. T_i corresponds to trees generated by minimizing ζ_i. T_{FM} corresponds to the tree based on the original Fitch and Margoliash's criterion ζ_{FM} and T_{Sa} corresponds to the tree based on the Sanjuán and Wróbel[39] criterion ζ_{Sa}. Several variations around the criterion that have been previously discussed are tested (criteria ζ_1, ζ_2, ζ_3, ζ_4, ζ_5 and ζ_6, described in Table 1). Note that T_2 (respectively T_5) is calculated from T_1 (respectively T_4) thanks to NNI transformations, within the process of progressive reduction of λ. T_{NJ} corresponds to trees generated by Neighbor Joining,[11] which is an agglomerative method for tree building from the distance matrix. Neighbor Joining (often noted NJ) is designed to generate the tree topology that gives the least total branch length. It is known to be a very fast method for phylogeny, which often provides a good preservation of distances even in maximum likehood methods, which are slower methods, were shown to be much more accurate than NJ. As a consequence, Neighbor Joining will be also compared to our method. We will then test here eight criteria and nine ways to create trees.

Evaluation plots: "continuity" and "trustworthiness"

Within Multi Dimensional Scaling framework, the distance preservation is often analyzed thanks to the Shepard's dy-dx diagram.[29,76] However, in tree-building framework, an equivalent diagram is not as easily interpretable (data not shown). To explore the

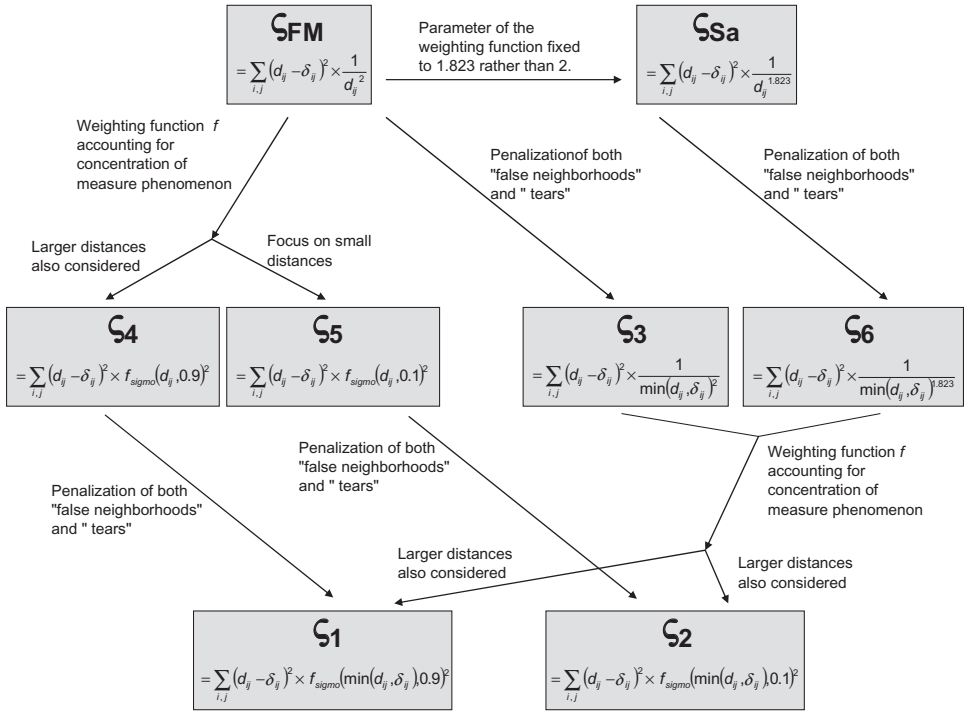

Figure 2. Summary of the eight tested criteria and links between them. Every combination of criterion components is tested to evaluate each improvement. Components allows: penalizing either "tears" (as in original criteria) or "tears" and "false neighborhoods" together (as it is suggested here); emphasizing small distances either thanks to the traditional reverse function or while considering the concentration of measure phenomenon with more (cases ζ_2 and ζ_5) or less (cases ζ_1 and ζ_4) focusing on smaller distances. ζ_{FM} corresponds to the original Fitch-Margoliash's criterion and ζ_{Sa} to the Sanjuán and Wróbel's one. ζ_1 and ζ_2 avoid both risks of "false neighborhoods" and risks related to the concentration of measure phenomenon. ζ_3 and ζ_6 avoid risks related to the concentration of measure phenomenon and risks of "tears" (but do not considered the risk of "false neighborhoods"). ζ_4 and ζ_5 avoid risks related to the concentration of measure phenomenon (while highly focusing on smaller distances in case of ζ_5) but lightly penalize "false neighborhoods".

distance preservation in a more appropriate way, we split the distance preservation through two curves that display "continuity" and "trustworthiness" (concepts introduced by Venna and Kaski).[65]

On the one hand, "continuity" quantifies the preservation of small original distances. As a consequence, a "tear" (close data in original space that are displayed widely separated) is a violation of the continuity. On the other hand, "trustworthiness" corresponds to the ability of the tree to express small distances. As a consequence, a "false neighborhood" (far distanced data in original space that are displayed as neighbors) is a violation of the trustworthiness.

Table 1. Kappa coefficients (c.f. section 5.3) for each method (in column) and for various *k* (in row). Note that the whole number of distances between 15 species is 105.

	T_1	T_2	T_3	T_4	T_5	T_{FM}	T_6	T_{Sa}	T_{NJ}
1	0.85	0.98	0.91	0.85	0.98	0.9	0.88	0.89	0.86
2	0.83	0.96	0.9	0.84	0.95	0.89	0.90	0.89	0.87
3	0.8	0.96	0.87	0.81	0.95	0.87	0.86	0.86	0.88
4	0.81	0.96	0.88	0.82	0.94	0.88	0.87	0.87	0.85
5	0.82	0.94	0.87	0.83	0.92	0.87	0.86	0.86	0.86
10	0.85	0.88	0.87	0.85	0.87	0.86	0.86	0.86	0.87
15	0.85	0.85	0.86	0.85	0.85	0.86	0.86	0.86	0.87

Trustworthiness and continuity are displayed here as two curves: 1) the curve that draws the square differences between original and output distances as a function of the original distance allows evaluating the continuity, and 2) the curve that draws the square differences between original and output distances as a function of the output distance allows evaluating the trustworthiness. Of course, these curves are often not slick; according to the user's purpose, these curves can be smoothened or not (a smooth curves allows the visualization of the general behaviors while an unsmooth curve reveals default). In the present paper, most figures show smoothened curves of "continuity" and "trustworthiness".

Some data randomly embedded (uniform distribution) in a two-dimensional Euclidian space were analyzed here (Fig. 3, upper insert). Trees are first generated from the matrix distance according to the Fitch-Margoliash method and then by the new

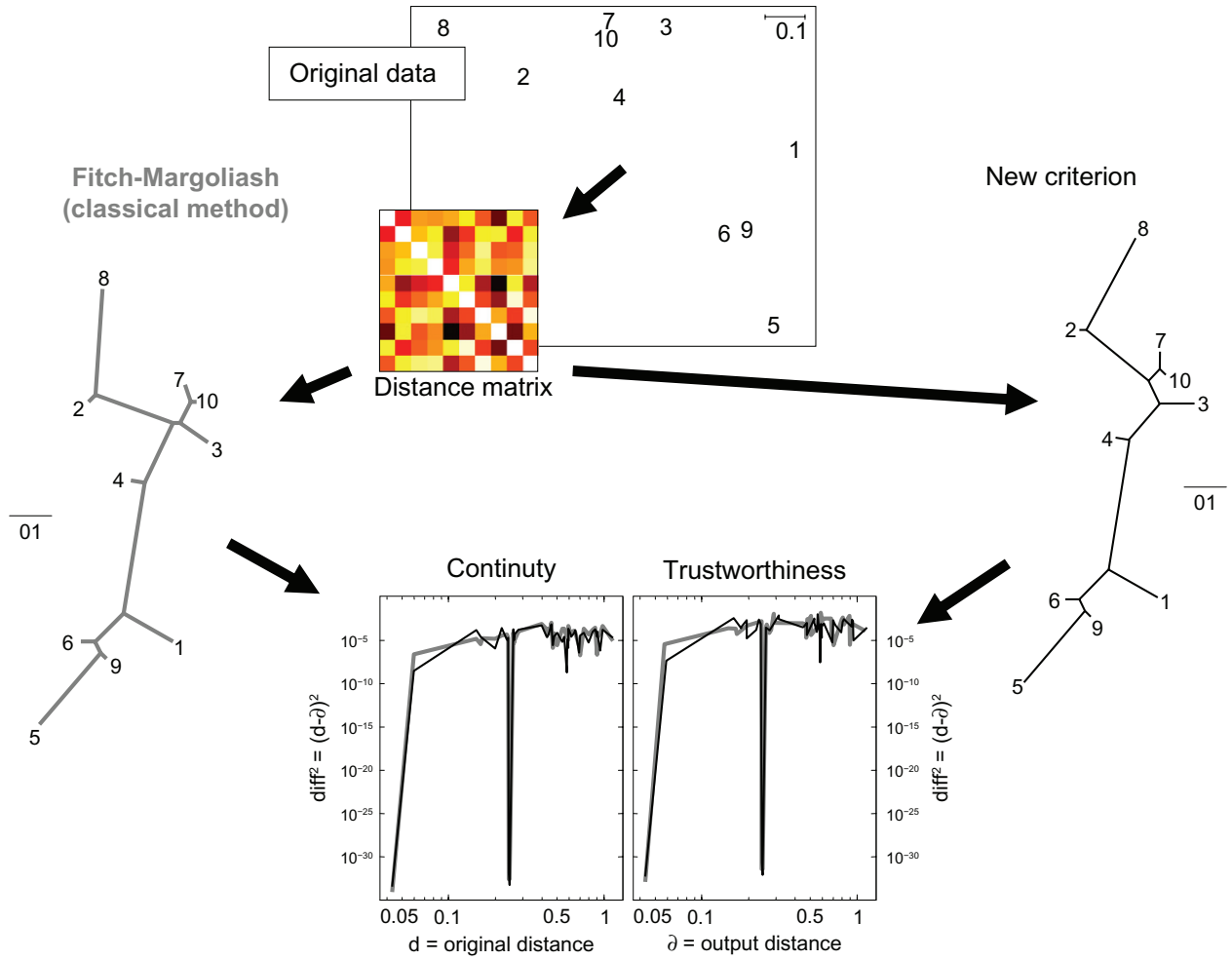

Figure 3. Example of comparison between two tree-building methods: Tree built according to the classical Fitch-Margoliash method versus the tree built thanks to the new criterion. Original data (and the associate distance matrix) are displayed in the upper insert. Left and right inserts express the two trees. "Continuity" and "trustworthiness" can then be compared on the lower insert (the grey curve corresponds to the Fitch-Margoliash method and the black curve is related to the new method).

method. Some differences can be noticed between the topology of the two resulting trees (lateral inserts). The comparison between original and resulting tree distances are presented in the lower insert in terms of "continuity" and "trustworthiness" (lower insert). Figure 3 illustrates the algorithm progress but does not allow conclusion about the characteristics of each criterion.

Kappa coefficients: preservation of k smallest distances

In order to compare the preservation of short distances in each method, the concordance between the k smallest original distances and the k smallest distances in trees is analyzed by mean of Kappa coefficients.[77] Kappa measures the degree of concordance on a scale from minus infinity to 1. A Kappa of one indicates full concordance, a Kappa of zero indicates that there is no more concordance than expected by chance and negative values are observed if concordance is weaker than expected by chance (a very rare situation). The measure could be seen as another mean to quantify the quality of resulting trees.

Kappa values are generated for each method and for various size of neighborhood (ie, various k).

Simulated data

Randomly positioned data

To test the capability of each criterion to generate fine trees from matrix of no additive distances, we created distance matrices from data randomly embedded in a Euclidean space. We tested sets of 15 data (ie, species) in two-dimensional and

100-dimensional spaces. Using 15 data points only could seem low, but it is clearly large enough to compare algorithms. Indeed, there is 7.9×10^{14} possible unrooted trees (which provide many chances for differences between results). Moreover, using 15 data points fits with the order of magnitude of real phylogenic problems. Lastly, 15-leaves trees are computable in a reasonable time; a large number of datasets can be simulated, which ensure to limit biases due to atypical situations.

Simulated sequences

Biological sequence evolution was simulated using a branching process. So, they are always related to a tree topology. For each time unit, all sequences present in the process can both be duplicated (with a fixed probability of duplication) and be mutated. The process starting from i) a sequence a0 ii) a probability distribution with the a0 amino acids position as support for this distribution reflecting the profile from which is derived the a0 protein family and iii) a duplication rate which can be allowed to vary with time or not. This program named EvolSeq was written in language C and can be requested to the authors.

In the present paper we mainly test the tree building methods with randomly positioned data; but several results reached with simulated sequences are also shown as supplementary material. Why was randomly positioned data preferred? Because randomly positioned data can maintain a perfect control on the topology and the dimensionality of the dataset. Conversely, as far as we know, whereas degrees of freedom are often controlled in case of simulated sequences, the sequences can however lie onto a subspace that dimensionality and topology are unknown. Indeed, let us consider biological sequences under evolution process as data that follow paths in a sequence space (the so called Configuration Space of Homologous Proteins, CSHP, after Bastien et al).[18] The dimensionality of such a space is the number of varying sites, but sequences more often live in a subspace with a much lower dimension (due to the co-varying sites and so on). In such a case, the possible configuration of sequences allowed by the process (which corresponds to the topology of the subspace) is unknown. However, highlighting the critical impact of the subspace dimensionality is one of the main claims of the present paper.

Two-dimensional space

A distance matrix has been created from 15 points embedded in a two-dimensional Euclidean space (random points, uniform distribution). Nine trees have been produced according to the nine tested criteria (with the same order for the insertion of data). We then observe various topologies according to the considered criterion. The average of Robinson and Foulds distances (Robinson and Foulds)[50] allows observing the proximity of tree topology obtained by each method (up and left insert in Figure 4 display these distances on a two-dimensional map thanks to the Data-Driven High Dimensional Scaling algorithm). For each method, "continuity" and "trustworthiness" are displayed as curves. To avoid considering the peculiarities of a given dataset, 200 sets of 15 data have been successively generated. For each dataset, trees have been calculated according to the various tested methods. This procedure led to 200 results that are merged in order to compare methods: the presented Robinson and Foulds distances corresponds to the average of the 200 Robinson and Foulds distances and trustworthiness and continuity curves are generated from every resulting couple of distances (that is $200 \times 15 \times 14/2 = 21000$ distances).

Methods lead to different results, even from the topology point of view. Criteria that use the weighting function with high consideration of small distance (criteria ζ_2 and ζ_5) provide the best "continuity". Note that they both lead to close results (c.f. Robinson and Foulds distance). However, ζ_5 does not provide a high "trustworthiness": many small distances in the tree correspond to large original distances which relies on "false neighborhoods" predicted by theoretical arguments (indeed ζ_5 has been said to few penalize "false neighborhoods"). Robinson and Foulds distance also reveals that other criteria (ζ_1 ζ_3 ζ_4 ζ_6 ζ_{Sa} and ζ_{FM}) lead to close results except for the Neighbor Joining (NJ). Such a simple dataset highlights the reality of the problem mentioned in section 3, and shows that some improvement can be done to circumvent these problems.

Kappa values are globally close to 1, which indicates that the distances sorting is generally well preserved (Table 1). Here again, criteria that focus on small distances (ζ_2 and ζ_5) clearly stand out. However, ζ_2 (the proposed criterion) seems to lightly outclass ζ_5. These results are clearly statistically

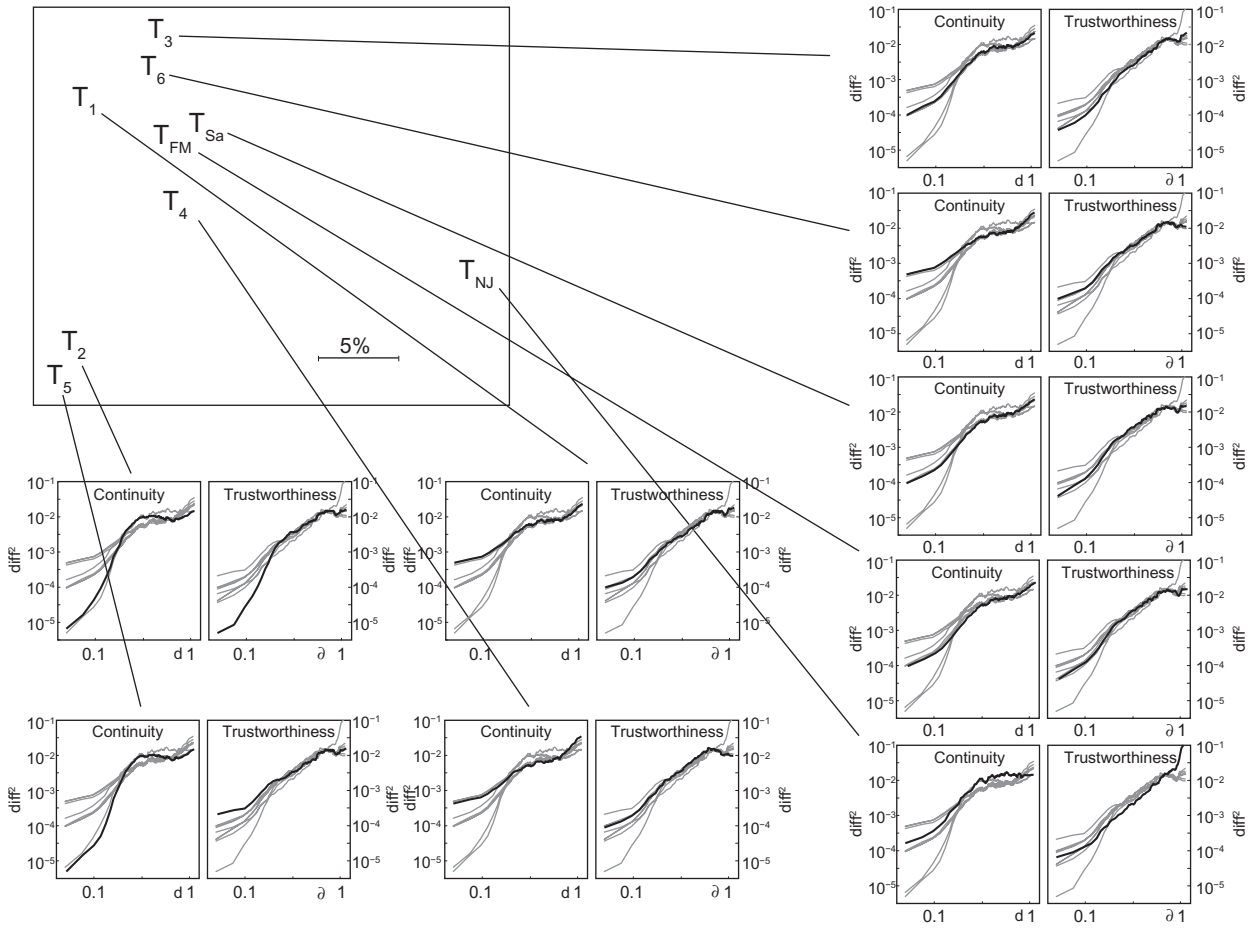

Figure 4. Evaluation of distance preservation by nine tree building methods (analysis on 200 sets of 15 random data in a two-dimensional space). The up and left insert shows the Robinson and Foulds distance between trees generated by the various methods: the distance that separates two methods in the graph accounts for the average Robinson and Foulds distance between methods' trees (see Hillis et al)[78] for a somewhat similar display of Robinson and Foulds distances). Other inserts express "continuity" (lefts inserts) and "trustworthiness" (right inserts). Every curve is reported on each graph to allow an easy comparison. On each insert, the related method corresponds to the black curve.

significant. Indeed, after reproducing the estimation of Kappa values procedure 30 times, we observed a standard deviation around the values equal to 0.01 (a difference higher than 0.02–2 × σ–can be considered as significant).

High-dimensional space

The same procedure has been used on datasets randomly generated in a 100-dimensional space (the quantity of data and the number of iterations remain stable) and the result is illustrated in Figure 5.

As previously, ζ_2 and ζ_5 show the best preservation of short original distance ("continuity"). ζ_5 provides a poor "trustworthiness" due to the presence of "false neighborhoods" predicted by theoretical arguments (same in the previous test). However, in case of high-dimensional data, results from ζ_2 and ζ_5 are somewhat different (high Robinson and Foulds distance). Moreover, results obtained with ζ_1 and ζ_4 can be discriminated from those obtained with ζ_3,ζ_{FM}, ζ_6,ζ_{Sa} and NJ. This difference is expected because ζ_1 and ζ_4 use the sigmoid weighting function: in case of high-dimensional data, only the sigmoid differentiates large and small distances (c.f. section 3).

Obviously, preserving the distances sorting is much harder for high dimensional data (values in Table 2 are weaker that the ones in Table 1). This situation can surely be related with the concentration of measured phenomenon that ensures that all distances are more or less similar (c.f. section 3).

Contrast between Kappa values is much weaker here, which indicates that most methods are more or less equivalent from the distance sorting point of view. However, methods using ζ_2 and ζ_5 are clearly ahead of the curve in quality. The standard deviation of the Kappa values estimation equals to 0.015: a

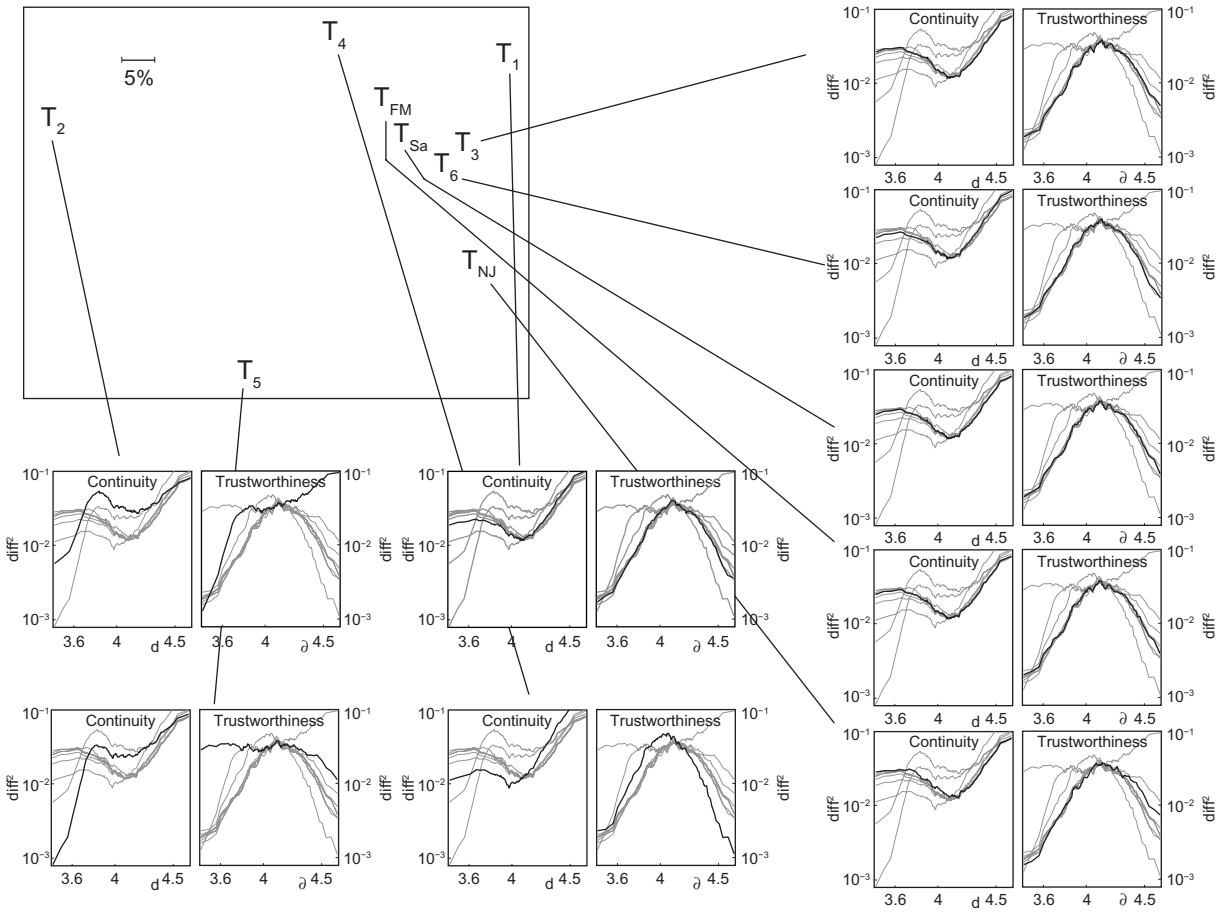

Figure 5. Evaluation of distance preservation (data in a 100-dimensional space). The up and left insert shows the Robinson and Foulds distance between trees generated by the various methods. Other inserts express "continuity" (lefts inserts) and "trustworthiness" (right inserts). Every curve is reported on each graph to allow an easy comparison. On each insert, the related method corresponds to the black curve.

difference higher than 0.03 can then be considered as statistically significant.

Test for the robustness of optimization processes

The NNI procedure can be seen as a gradient descent algorithm used to optimize the tree topology (monitored by various criteria). It is well known that a risk of gradient methods is to fall into local minimums. To quantify this risk according to each criterion, we used a randomly embedded dataset (8 data; two-dimensional at first, and then in 100-dimensional spaces). 20 random trees are then used as initializations of a NNI process (same data and same initializations are used for every criterion). The tree with the minimal value for the criterion is considered as the global minimum: the percentage of resulting trees corresponding to this solution quantifies then the robustness of optimization using a given criterion. To be free from the feature of a particular dataset, this test is iterated 1000 times, and results are averaged (Table 3).

Table 2. Kappa coefficients (c.f. section 5.3) for each method (in column) and for various *k* (in row). Note that the whole number of distances between 15 species is 105.

	T_1	T_2	T_3	T_4	T_5	T_{FM}	T_6	T_{Sa}	T_{NJ}
1	0.77	0.87	0.75	0.78	0.89	0.77	0.74	0.77	0.77
2	0.71	0.87	0.68	0.71	0.89	0.69	0.67	0.69	0.69
3	0.68	0.86	0.66	0.71	0.84	0.66	0.65	0.67	0.66
4	0.69	0.85	0.65	0.73	0.83	0.65	0.65	0.65	0.67
5	0.69	0.84	0.65	0.72	0.81	0.64	0.64	0.64	0.66
10	0.67	0.72	0.64	0.7	0.71	0.64	0.64	0.64	0.65
15	0.68	0.64	0.65	0.69	0.67	0.65	0.66	0.65	0.65

Because NNI is run from a tree containing all items, we were able to treat here an unfavorable case: indeed, the progressive introduction of items in the tree should often help to reach the optimal minima. This can explain the somewhat low robustness for

Table 3. Robustness of NNI according to the criterion that drives optimization. Each method is evaluated by the percentage of generated tree that corresponds to the optimal tree according to its criterion.

	ζ_1	ζ_2	ζ_3	ζ_4	ζ_5	ζ_{FM}	ζ_6	ζ_{Sa}
dim = 2	40.9%	61.7%	44.8%	30.9%	56%	33.4%	39.1%	33.9%
dim = 100	45.2%	48.8%	45%	41.3%	48.6%	42.1%	45.2%	42.2%

most methods (Table 3). However, obviously, an iterative introduction of items (as implemented by most methods) cannot guarantee that the optimal minima will be reached and the robustness is then an important property of tree building methods.

It can be seen that the use of the sigmoid weighting function with a low value for λ (that allows focusing on shortest distances) strongly increases the efficiency of the algorithm, as well as its robustness. Indeed, In Tables 1 and 2, we can observe the benefice in terms of efficiency though a gap between results reached from criterion 2 and the ones from criterion 3 as well as results reached from criterion 5 and the ones from the FM criterion (in both cases original Fitch-Margoliash weighting function is replaced by the sigmoid weighing function, please report to Fig. 2). Similarly in Table 3, the robustness of criterion 2 (respectively 5) is clearly higher than the robustness of criterion 3 (respectively, the FM criterion).

The penalization of both "false neighborhoods" and "tears" also clearly increases the robustness (by comparing criterion 3 to FM criterion or criterion 2 to criterion 5 in Table 2), but less strongly than the use of a sigmoid weighing function. However, it is not clear if it improves or damages the efficiency: trees from low-dimensional data seems to be very lightly improved (Table 1) but trees from high-dimensional data seems to be very lightly decreased (Table 2) when both defaults are similarly penalized. In order to evaluate the statistical significance of results presented in Table 3, we reproduced the estimation of robustness evaluation 30 times: the standard deviation around the values equal to 0.5% (a difference higher than 1% can thus be considered as significant).

6. Conclusion

Several properties derived from Multi Dimensional Scaling can be used within tree-building framework. In particular, we highlight the importance of taking "false neighborhoods" and "tears" into account while computing phylogeny. A new criterion is then proposed to overcome these risks and new evaluation processes using "continuity" and "trustworthiness" concepts allow us to check in it.

We also described the concentration of measured phenomenon and its possible influence when a tree is calculated according to classical methods. We then rectified the criterion in order to take account of the curse of dimensionality.

The analysis of simulated datasets highlighted that topology of trees may be different according to the chosen criterion. The new criterion outperforms other ones from distances preservation point of view. Moreover, a better robustness is observed.

As a consequence, the proposed criterion should be considered as a worthy alternative to the Fitch-Margoliash's one. Future works should i) include simulations conducted in a more realistic way, for example by adding noise to the evolution process, ii) be evaluated on several real datasets and iii) be extended to maximum-likehood approaches.

Acknowledgements

OB was supported by the Agence Nationale de la Recherche, as part of the PlasmoExplore project.

Disclosure

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Hitchcock E, Hitchcock CH. *Elementary Geology*. University of Michigan Library: Scholarly Publishing Office; 1840.
2. Darwin C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray; 1859.
3. Haeckel E. *Generelle Morphologie der Organismen*. Charleston, Carolina: Nabu Press; 1866.

4. Brocchieri L. Phylogenetic inferences from molecular sequences: review and critique. *Theor Popul Biol*. 2001;59:27–40.
5. Edwards AWF, Cavalli-Sforza LL. The reconstruction of evolution. *Ann Hum Genet*. 1963;27:105–6.
6. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 1981;17:368–76.
7. Sun FJ, Caetano-Anollés G. The origin and evolution of tRNA inferred from phylogenetic analysis of structure. *J Mol Evol*. 2008;66(1):21–35.
8. Wang M, Caetano-Anollés G. The evolutionary mechanics of domain organization in proteomes and the rise of moduarity in the protein world. *Structure*. 2009;17:66–78.
9. Steel M. A basic limitation on inferring phylogenies by pairwise sequence comparisons. *J Theor Biol*. 2009;256:467–72.
10. Sneath PHA, Sokal RR. *Numerical taxonomy. The Principles and Practice of Numerical Classification*. San Francisco: W.H. Freeman and company; 1973.
11. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Evol Biol*. 1987;4(4):406–25.
12. Salemi M, Vandamme AM. T*he phylogenetic handbook: A practical approach to DNA and protein phylogeny*. London: Cambridge University Press; 2003.
13. Decrypthon Project. 2005. Available at: http://www.decrypthon.fr/english/index.php. Accessed December 14, 2010.
14. Bastien O, Lespinats S, Roy S, Métayer K, Fertil B, Codani JJ, Maréchal E. Analysis of the compositional biases in *Plasmodium falciparum* genome and proteome using Arabidopsis thaliana as a reference. *Gene*. 2004; 336(2):163–73.
15. Lespinats S. *Style du génome exploré par analyse textuelle de l'ADN*. Paris VI University: PhD thesis; 2006.
16. Fitch WM, Margoliash E. Construction of phylogenetic trees. *Science*. 1967;155:279–84.
17. Agrafiotis DK. A new method for analyzing protein sequence relationship based on Sammon maps. *Protein Science*. 1997;6:287–93.
18. Bastien O, Ortet P, Roy S, Maréchal E. A configuration space of homologous proteins conserving mutual information and allowing a phylogeny inference based on pair-wise Z-score probabilities. *BMC Bioinformatics*. 2005;6:49.
19. Grishin VN, Grishin NV. Euclidean space and grouping of biological objects. *Bioinformatics*. 2002;18(11):1523–33.
20. Zimmermann K, Gibrat JF. Amino acid "little Big Bang": Representing amino acid substitution matrices as dot products of Euclidian vectors. *BMC Bioinformatics*. 2010;11:4.
21. Bastien O. A Simple Derivation of the Distribution of Pairwise Local Protein Sequence Alignment Scores. *Evolutionary Bioinformatics*. 2008;4: 41–5.
22. Bastien O, Maréchal E. Evolution of biological sequences implies an extreme value distribution of type I for both global and local pairwise alignment scores. *BMC Bioinformatics*. 2008;9:332.
23. Ortet P, Bastien O. Where Does the Alignment Score Distribution Shape Come From? *Evolutionary Bioinformatics*. 2010;6:159–87.
24. Aravind L, Mazumder R, Vasudevan S, Koonin EV. Trends in protein evolution inferred from sequence and structure analysis. *Curr Opin Struct Biol*. 2002;12(3):392–9.
25. Jolliffe I. *Principal Component Analysis, 2nd edition*. New York: Springer-Verlag; 2002.
26. Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*. 1901;2:559–72.
27. Torgerson WS. Multidimensional scaling: 1. Theory and method. *Psychometrika*. 1952;17:401–19.
28. Sammon JW. A nonlinear mapping for data structure analysis. *IEEE Trans Comput*. 1969;18(5):401–9.
29. Demartines P, Hérault J. Curvilinear component analysis: A self organizing neural network for nonlinear mapping of datasets. *IEEE Trans Neural Netw*. 1997;8(1):148–54.
30. Schölkopf B, Smola AJ, Müller KR. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput*. 1998;10:1299–319.
31. Tenenbaum JB, De Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. *Science*. 2000;290:2319–23.
32. Lee JA, Lendasse A, Verleysen M. Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis. *Neurocomputing*. 2004;57:49–76.
33. Bishop CM, Svensén M, Williams CKI. GTM: The generative topographic mapping. *Neural Comput*. 1998;10:215–34.
34. Lespinats S, Verleysen M, Giron, A, Fertil B. DD-HDS: a tool for visualization and exploration of highdimensional data. *IEEE trans. Neural Netw*. 2007;18(5):1265–79.
35. Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under inequal evolutionary rates. *Mol Biol Evol*. 1994;11: 459–68.
36. Susko E. Confidence regions and hypothesis testing for topologies using generalized least-squares. *Mol Biol Evol*. 2003;20:62–868.
37. Zaretsky K. Reconstruction of a tree from the distances between its pendant vertices. *Uspekhi Math Nauk (Russian Mathematical Survey)*. 1965; 20:90–2.
38. Cavalli-Sforza LL, Edwards AWF. Phylogenetic analysis: models and estimation procedures. *Am J Hum Genet*. 1967;19:223–57.
39. Sanjuán R, Wróbel B. Weighted least-squares likelihood ratio test for branch testing in phylogenies reconstructed from distance measures. *Syst Biol*. 2005;54(2):218–29.
40. Lee JA, Lendasse A, Verleysen M. *Nonlinear dimensionality reduction*. New York: Springer; 2007.
41. Felsenstein J. The number of evolutionary trees. *Syst Zool*. 1978;27:27–33.
42. Day WHE. Computational complexity of inferring phylogenies by dissimilarity matrices. *Bull Math Bio*. 1987;49:461–7.
43. Day WHE. Complexity theory: an introduction for practitionners of classification. In Arabie P, Hubert LJ, De Soete G, editors. *Clustering and Classification*. London: World Scientific. 1996:199–233.
44. Matsuda H. Construction of phylogenetic trees from amino acid sequences using a genetic algorithm. In: Hagiya M, Suyama A, Takagi T, Nakai K, Miyano S, Yokomori T, editors. *Proceedings of Genome Informatics Workshop 1995; Yokohama (Japan)*. Tokyo: Universal Academy Press. 1995:19–28.
45. Poladian L. A GA for maximum likelihood phylogenetic inference using neighbour-joining as a genotype to phenotype mapping. In: *Proceedings of the 2005 conference on Genetic and evolutionary computation (GECCO '05)*. New York: ACM Press. 2005:415–22.
46. Reijmers TH, Wehrens R, Daeyaert FD, Lewi PJ, Buydens LM. Using genetic algorithms for the construction of phylogenetic trees: application to G-protein coupled receptor sequences. *Biosystems*. 1999;49:31–43.
47. Qin L, Chen Y, Pan Y, Chen L, Guo J. A novel approach to phylogenetic tree construction using clustering and optimization strategies based on the ant colony algorithm. *BMC Bioinformatics*. 2006;7:S24.
48. Lin M, Fang SC, Thorne JL. A TABU search algorithm for maximum parsimony phylogeny inference. *Eur J Operational Research*. 2007;176: 1908–17.
49. Felsenstein J. PHYLIP: Phylogeny inference package, version 3.5c. *Department of Genetics, University of Washington, Seattle*, 1118. 1993. Available from: http://evolution.genetics.washington.edu/phylip.html. Accessed March 02, 2011.
50. Robinson DF, Foulds L. Comparison of weighted labeled trees. *Lectures Notes in Mathematics*. 1979;748:119–26.
51. Swofford DL, Olsen GJ. Phylogeny Reconstruction. In: Hillis DM, Moritz C, editors. *Molecular Systematics*. Sunderland, Massachussetts: Sinauer Associates. 1990:411–501.
52. Waterman MS, Smith TF. On the similarity of dendograms. *J Theor Biol*. 1978;73:789–800.
53. Li J, Guo M. A New Approach to Evolutionary Tree Reconstruction Combining Particle Swarm Optimization with p-ECR. *Int J Comput Intel Res*. 2008;4(2):187–95.
54. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge: Cambridge Univ. Press; 1992.
55. Glover F, Laguna M. Tabu Search. In: Reeves CR, editors. *Modern Heuristic Techniques for Combinatorial Problems*. London: McGraw-Hill. 1995: 70–150.

56. Takane Y, Young FW, de Leeuw J. Nonmetric individual differences multidimensional scaling: An alternating least-squares method with optimal scaling features. *Psychometrika*. 1977;42:7–67.
57. Goldberg DE. *Genetic algorithms in Search, Optimization, and Machine Learning*. Massachusetts: Addison-Wesley Publishing Co. 1989.
58. Reeves CR. Genetic Algorithms. In: Reeves CR, editors. *Modern Heuristic Techniques for Combinatorial Problems*. London: McGraw-Hill. 1995: 151–96.
59. Dowsland KA. Simulated Annealing. In: Reeves CR, editors. *Modern Heuristic Techniques for Combinatorial Problems*. London: McGraw-Hill. 1995:377–419.
60. Li JX. Visualization of high-dimensional data with relational perspective map. *Inf Visualization*. 2004;3:49–59.
61. Chalmers M. A linear iteration time layout algorithm for visualizing high-dimensional data. In: Yagel R, Nielson GM, eds. *Proceedings of the 7th conference on Visualization'96; SanFrancisco/LosAlamitos (California)*. 1996:127–32.
62. Morrison A, Ross G, Chalmers M. Fast Multidimensional Scaling through Sampling, Springs and Interpolation. *Inf Visualization*. 2003;2:68–77.
63. Aupetit M. Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing*. 2007;70(7–9):1304–30.
64. Lee JA, Verleysen M. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*. 2009;72:1431–43.
65. Venna J, Kaski S. Local multidimensional scaling. *Neural Netw*. 2006;19: 889–99.
66. Venna J, Kaski S. Neighborhood preservation in nonlinear projection methods: An experimental study. *Lecture Notes in Computer Science*. 2001; 2130:485–91.
67. Bellmann R. *Adaptive Control Processes: A Guided Tour*. New Jersey: Princeton University Press. 1961.
68. Donoho DL. High-dimensional data analysis: The curses and blessings of dimensionality. In: *Amer Math Soc Lecture: "Math challenges of the 21st century"; Los Angeles*. 2000. Available at: http://www-stat.stanford.edu/~donoho/. Accessed March 02, 2011.
69. Aggarwal CC, Hinneburg A, Keim DA. On the Surprising Behavior of Distance Metrics. *Lecture Notes in Computer Science*. 2001;1973:420–34.
70. Gromov M. Metric Structures for Riemannian and Non-Riemannian Spaces. *Progress in Mathematics*. 1999;152.
71. Milman VD. The heritage of P.Lévy in geometric functional analysis. *Astérisque*. 1988;157(158):273–301.
72. François D, Wertz V, Verleysen M. The Concentration of Fractional Distances. *IEEE T on Knowledge and data engineering*. 2007;19(7):873–86.
73. Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is Nearest Neighbors Meaningful? *Lecture Notes in Computer Science*. 1999;1540:217–35.
74. Bulmer M. Use of the method of generalized least-squares in reconstructing phylogenies from sequence data. *Mol Biol Evol*. 1991;8:868–83.
75. Felsenstein J. An alternating least-squares approach to inferring phylogenies. *Sys Biol*. 1997;46:101–11.
76. Demartines P. Mesures d'organisation du réseau de Kohonen. In: Cottrell M, editors. *Congrès Satellite du Congrès Européen de Mathématiques: Aspects Théoriques des Réseaux de Neurones*. Paris. 1992.
77. Fleiss JL. *Statistical methods for rates and proportions*. *2nd edition*. New York: John Wiley & Sons. 1981.
78. Hillis D, Heath T, St. John K. Analysis and Visualization of Tree Space. *Sys Biol*. 2005;54(3):471–82.

Supplementary Data

Example of a simulation with a subspace equal to the original sequence space and with the same topology

1. Common parameters for the simulation

Probability to have a mutation in a sequence per evolutionary time
#define MUTATION 0.05
Probability to have a duplication of a sequence per evolutionary time
#define DUPLICATION 0.0002
Number of step of the simulation
#define EVOLTIME 10000

2. Simulation results with two dimensions allow varying with time

a) Amino acids sequence a0

>Set12 domain, Toxoplasma Gondii
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAREREQRYNRSKVPMGSFMFYFKNG-SRMMAIDATDEKQDFGPARLINHSRRNPNMT PRAITLGDFNSEPRLIFVARRNIEKGEELLVDY

b) Distribution of probabilities on the amino acids sequence a0 positions

{0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.4, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.6, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0};

c) Results of the simulation process

The 15 sequences named A to O are the resulting sequences at the end of the evolution process.

1 (father: 0)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAREREQRYNRSKVPMGSFMFYFKNGSRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 1001
2 (father: 1)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAKEREQRYNRSKVPMGSFMFYYKNGSRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 8999 name A
3 (father: 1)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAREREQRYNRSKVPMGSFMFYFKNGSRMMAIDATDEKQDFGPARLINHSRRNPNMTP RAITLGDFNSEPRLIFVARRNIEKGEELLVDY 624
4 (father: 3)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYLKNGSRMMAIDATDEKQDFGPARLINHSRRNPNMTP RAITLGDFNSEPRLIFVARRNIEKGEELLVDY 1446
5 (father: 4)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYFKNGSRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 6929 name B

6 (father: 4)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYLKNG-
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 90
7 (father: 6)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYLKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 1167
8 (father: 7)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYL
KNGSRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
272
9 (father: 8)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYLKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 135
10 (father: 9)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYT-
KNGSRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
5265 name C
11 (father: 9)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYLKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
5265 name D
12 (father: 8)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYLKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 4164
13 (father: 12)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYLKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 182
14 (father: 13)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYLKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
1054 name E
15 (father: 13)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYLKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
1054 name F
16 (father: 12)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYL
KNGSRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
834
17 (father: 16)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYLKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
402 name G
18 (father: 16)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYLKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
402 name H

19 (father: 7)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYLKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 1496
20 (father: 19)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYQKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 2109
21 (father: 20)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYQKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
2067 name I
22 (father: 20)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYR
KNGSRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
1949
23 (father: 22)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYRKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
118 name J
24 (father: 22)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYRKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
118 name K
25 (father: 19)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYG-
KNGSRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
4176 name L
26 (father: 6)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAWEREQRYNRSKVPMGSFMFYLKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
6839 name M
27 (father: 3)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREARE REQRYNRSKVPMGSFMFYFKNGS
RMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 8014
28 (father: 27)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREARE REQRYNRSKVPMGSFMFYFKNG-
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
361 name N
29 (father: 27)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAKEREQRYNRSKVPMGSFMFYFKNG
SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
361 name O

d) Distance matrices corresponding to the process. Numbers are evolutionary time units.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
0	17998	17998	17998	17998	17998	17998	17998	17998	17998	17998	17998	17798	17998	17998
17998	0	13858	13858	13858	13858	13858	13858	13858	13858	13858	13858	13858	16750	16750

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
17998	13858	0	10530	10800	10800	10800	10800	11344	11344	11344	11344	13678	16750	16750
17998	13858	10530	0	10800	10800	10800	10800	11344	11344	11344	11344	13678	16750	16750
17998	13858	10800	10800	0	2108	2472	2472	11344	11344	11344	11344	13678	16750	16750
17998	13858	10800	10800	2108	0	2472	2472	11344	11344	11344	11344	13678	16750	16750
17998	13858	10800	10800	2472	2472	0	804	11344	11344	11344	11344	13678	16750	16750
17998	13858	10800	10800	2472	2472	804	0	11344	11344	11344	11344	13678	16750	16750
17998	13858	11344	11344	11344	11344	11344	11344	0	4134	4134	8352	13678	16750	16750
17998	13858	11344	11344	11344	11344	11344	11344	4134	236	0	8352	13678	16750	16750
17998	13858	11344	11344	11344	11344	11344	11344	8352	8352	8352	0	13678	16750	16750
17798	13858	13678	13678	13678	13678	13678	13678	13678	13678	13678	13678	0	16750	16750
17798	16750	16750	16750	16750	16750	16750	16750	16750	16750	16750	16750	16750	0	722
17798	16750	16750	16750	16750	16750	16750	16750	16750	16750	16750	16750	16750	722	0

e) Phylogenetic tree corresponding to the process

(A:8999,((B:6929,((((C:5265,D:5265):135,((E:1054,F:1054):182,(G:402,H:402):834):4164):272,((I:2067,(J:118,K:118):1949):2109,L:4176):1496):1167,M:6839):90):1446,(N:361,O:361):8014):624):1001

3. Simulation results with 100 dimensions allow varying with time

a) Amino acids sequence a0

>Set12 domain, Toxoplasma Gondii

CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAREREQRYNRSKVPMGSFMFYFKNG-SRMMAIDATDEKQDFGPARLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY

b) Distribution of probabilities on the amino acids sequence a0 positions

{0.005, 0.01, 0.02, 0.005, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0};

c) Results of the simulation process

1 (father: 0)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAREREQRYNRSKVPMGSFMFYFKNGSRMMAISATDEKQDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 797

2 (father: 1)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAREREQRYNRSKVPIGSFMFYFKNGSRMMAISATDEKQTFGPSRLINHSRRSPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 6555

3 (father: 2)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAREREQRYNRAKVPIGSFMFYFKNGSRMMAISATDEKQTFGPSRLINHSRRSPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 1148

4 (father: 3)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREAREREQRYNRAKVPIGSFMFYFKNGSRM-MAISLTDEKQTFGPSRLINHSRRSPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 1500 name A

5 (father: 3)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREARE REQRYNRAKVPIGSFMFYFKNGS
RMMAISATDEKQTFGPSRLINHSRRSPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 456
6 (father: 5)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREARE REQRYNRAKVPIGSFMFYFKNG-
SRMMAINATDEKQTFGPSRLINHSRRSPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
1044 name B
7 (father: 5)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREARE REQRYNRAKVPIGSFMFYFKNGS
RMMAISATDEKQTFGPSRLINHSRRSPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 822
8 (father: 7)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREARE REQRYNRAKVPIGSFMFYFKNGSRM-
MAISATDEKQTFGPSRLINHSRRSPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 222 name C
9 (father: 7)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREARE REQRYNRAKVPIGSFMFYFKNGS
RMMAISATDEKQTFGPSRLINHSRRSPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 222 name D
10 (father: 2)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREARE REQRYNRSKVPIGSFMFYFKNGS
RMMAISATDEKQTFGPSRLINHSRRSPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 141
11 (father: 10)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREARE REQRYNRSKVPIGSFMFYFKNG-
SRMMAISATDEKQTFGPSRLINHSRRSPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
2507 name E
12 (father: 10)
CIHLTKVPGKGRAVFAADTILKDDFVVEYKGELCSEREARE REQRYNRSKVPIGSFMFYFKNG-
SRMMAISATDEKQTFGPSRLINHSRRSPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
2507 name F
13 (father: 1)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSEREARE REQRYNRSKVPMGSFMFYFKNGS
RMMAISATDEKQDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 1079
14 (father: 13)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSEREARE REQRYNRSKVPMGSFMFYFKNGS
RMMAISATDDKQDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 252
15 (father: 14)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSEREARE REQRYNRSKVPMGSFRFYFKNGS
RMKAKSVTDDKQDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 6514
16 (father: 15)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSEREARE REQRYNRSKVPMGSLRFYFKNGS
RMKAKSVTDDKQDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
1358 name G
17 (father: 15)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSEREARE REQRYNRSKVPMGSFRFYFKNGS
RMKAKSVTDDKQDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
1358 name H
18 (father: 14)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSEREARE REQRYNRSKVPMGSFMFYFKNGS
RMMAISATDDKQDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY 4295

19 (father: 18)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSERESREREQRYNRSKVPMGSWMFYFKNG-
SRMLAISATDDKQDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
3577 name I
20 (father: 18)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSERESREREERYNRSKVPMGSFMFYFKNDS
RMMAISATDDKQDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
3577 name J
21 (father: 13)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSEREARE REQRYNRSKVPMGSFMFYF
KNGSRMMAISATDEKQDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
1500
22 (father: 21)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSEKEARE REQRYNRSKVPMGSFMFY
FKNGSRMMAISAIDEKQDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
4779
23 (father: 22)
CIHLTKVPGKGRAVFAPQTILKDDFVMEYKGELCSEKEARE REQRYNRSKVPMGSFMFYFKNG-
SRMMAIPAIDEKQDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
1845 name K
24 (father: 22)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSEKEARE REQRYNRSKVPMGSFMFYFKNG-
SRMMAISAIDEKQDFGPSRLINHSRRSPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
1845 name L
25 (father: 21)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSEREARE REQRYNRSKVPMGSFMFYFK
NGSRMMAISATDEKHDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
2042
26 (father: 25)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSEREARE REQRYNRSKVPMGSFMFYF
KNGSRMMAISATDEKHDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
2993
27 (father: 26)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSEREARE REQRYNRSKVPMSSFMFHFKN
GSRMMAISATDEKHDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
1589 name M
28 (father: 26)
CIHLTKVPGKGRAVFAPQTILKDDFVAEYKGELCSEREARE KEQRYNRSKVPMGSFMFYF
KNGSRMMAISATDEKHDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
1589 name N
29 (father: 25)
CIHLTKVPGKGRAVFAPQTILKDDFVVEYKGELCSEREARE REQRYNRSKVPMGSFMFYFKNG-
SRMMAISATDEKHDFGPSRLINHSRRNPNMTPRAITLGDFNSEPRLIFVARRNIEKGEELLVDY
4582 name O

d) Distance matrices corresponding to the process. Numbers are evolutionary time units.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
A	0	3000	3000	3000	5296	5296	18406	18406	18406	18406	18406	18406	18406	18406	18406
B	3000	0	2088	2088	5296	5296	18406	18406	18406	18406	18406	18406	18406	18406	18406
C	3000	2088	0	444	5296	5296	18406	18406	18406	18406	18406	18406	18406	18406	18406
D	3000	2088	444	0	5296	5296	18406	18406	18406	18406	18406	18406	18406	18406	18406
E	5296	5296	5296	5296	0	5014	18406	18406	18406	18406	18406	18406	18406	18406	18406
F	5296	5296	5296	5296	5014	0	18406	18406	18406	18406	18406	18406	18406	18406	18406
G	18406	18406	18406	18406	18406	18406	0	2716	15744	15744	15744	15744	15744	15744	15744
H	18406	18406	18406	18406	18406	18406	2716	0	15744	15744	16248	16248	16248	16248	16248
I	18406	18406	18406	18406	18406	18406	15744	15744	0	7154	16248	16248	16248	16248	16248
J	18406	18406	18406	18406	18406	18406	15744	15744	7154	0	16248	16248	16248	16248	16248
K	18406	18406	18406	18406	18406	18406	16248	16248	16248	16248	0	3690	13248	13248	13248
L	18406	18406	18406	18406	18406	18406	16248	16248	16248	16248	3690	0	13248	13248	13248
M	18406	18406	18406	18406	18406	18406	16248	16248	16248	16248	13248	13248	0	3178	9164
N	18406	18406	18406	18406	18406	18406	16248	16248	16248	16248	13248	13248	3178	0	9164
O	18406	18406	18406	18406	18406	18406	16248	16248	16248	16248	13248	13248	9164	9164	0

e) Phylogenetic tree corresponding to the process

(((A:1500,(B:1044,(C:222,D:222):822):456):1148,(E:2507,F:2507):141):6555,(((G:1358,H:1358):6514,(I:3577,J:3577):4295):252,((K:1845,L:1845):4779,((M:1589,N:1589):2993,O:4582):2042):1500):1079):797

4. Comparisons Simulation Results