

Concerning the NJ algorithm and its unweighted version, UNJ

Olivier Gascuel

⁽¹⁾ GERAD, Ecole des HEC,
5255 Av. Decelles, Montréal (Québec) - H3T 1V6 - CANADA
olivierg@crt.umontreal.ca

⁽²⁾ Département d'Informatique Fondamentale,
LIRMM, 161 rue Ada, 34392 - Montpellier - FRANCE
gascuel@lirmm.fr

⁽¹⁾ Current address, until July 31, 1997.

⁽²⁾ From August 1997.

Abstract

In this paper we will present UNJ, an unweighted version of the NJ algorithm (Saitou and Nei 1987; Studier and Keppler 1988). We will demonstrate that UNJ is well suited when the data are of the $(\delta_{ij}) = (d_{ij} + \varepsilon_{ij})$ type, where (d_{ij}) is a tree distance, and when the ε_{ij} are independent and identically distributed noise variables. Simulations confirm this theory. On a more general level, we will study the three main components of the agglomerative approach, applied to the reconstruction of tree distances.

(i) We will demonstrate that the selection criterion for the pair to be agglomerated, used by NJ and UNJ, retains its meaning whatever the variances and covariances of the δ_{ij} estimates. We will also provide a new proof of the correction of this criterion, based on an interpretation in acentrality terms proposed by Mirkin (1996).

(ii) Using the results of Vach (1989), of which we will provide a simple new demonstration, we propose an analytical formula which enables the correct least-squares estimation of edge lengths in $O(n^2)$ time, where n is the number of objects.

(iii) We will provide a class of admissible reduction formulae which guarantee the finding of the true tree with additive data. We propose to choose, among these formulae, the minimum variance reduction, so that at each step we use estimates which are as reliable as possible in choosing the pair to be agglomerated. We will present the general solution, and apply it to the particular data model retained here.

Résumé

Nous décrivons ici UNJ qui est une version non-pondérée de l'algorithme NJ (Saitou et Nei 1987; Studier et Keppeler 1988). Nous montrons que UNJ est bien adapté lorsque les données sont du type $(\delta_{ij}) = (d_{ij} + \varepsilon_{ij})$, où (d_{ij}) est une distance d'arbre, et où les ε_{ij} sont des variables de bruit indépendantes et identiquement distribuées. Des simulations confirment cette analyse théorique. Plus généralement, nous revenons sur les trois principales composantes de l'approche agglomérative, appliquée à la reconstruction des distances d'arbre.

- Nous montrons que le critère de sélection de la paire à agglomérer, utilisé par NJ et par UNJ, conserve son sens quel que soit les (co)variances des estimateurs δ_{ij} . Nous apportons également une nouvelle preuve de la correction de ce critère, en nous basant sur son interprétation en terme d'acentralité proposée par Mirkin (1996).

- En nous appuyant sur les résultats de Vach (1989), dont nous apportons une démonstration nouvelle, nous proposons une formule analytique permettant d'estimer efficacement les longueurs de branche d'un arbre de structure fixée, de manière optimale au sens des moindres carrés.

- Nous donnons une classe de formules de réduction admissibles, au sens où elles garantissent de retrouver l'arbre vrai avec des données additives, et nous proposons de choisir au sein de celles-ci la réduction de variance minimum, de manière à disposer à chaque étape d'estimateurs aussi fiables que possible pour choisir la paire à agglomérer. La solution générale est donnée, puis appliquée au modèle de données retenu ici.

1. Introduction

Let $\mathbf{D} = (d_{ij})$ be a tree distance over n objects ; \mathbf{T} the unique valued tree allowing the representation of \mathbf{D} , also referred to as the true tree ; $\mathbf{\Delta} = (\delta_{ij})$ a dissimilarity matrix of which each element δ_{ij} is an estimate or a measure, generally imperfect, of the distance d_{ij} . In this paper, we propose to find \mathbf{T} from the observation $\mathbf{\Delta}$. In other words, we try to construct a valued tree $\hat{\mathbf{T}}$, associated with the tree distance $\hat{\mathbf{D}} = (\hat{d}_{ij})$, which should be as close as possible to \mathbf{T} . Of course, there are several ways of defining the proximity between trees. In this paper, we will focus on the structure of the trees \mathbf{T} and $\hat{\mathbf{T}}$, rather than on the length of their edges. This problem is encountered in domains where one tries to construct an inheritance phenomenon as, for example, the history of manuscripts in Archaeology (Buneman 1971), or the evolution of the species in Biology (Swofford *et al.* 1996). The tree \mathbf{T} represents the history, the distances (d_{ij}) represent the divergence times between these objects or these species, and the dissimilarities (δ_{ij}) are estimates of these divergence times.

A classical approach consists in following the least-squares criterion in constructing the positively-valued tree which best represents $\mathbf{\Delta}$ according to this criterion (Cunningham 1978; De Soete 1983; Roux 1988; Gascuel and Levy 1996). Although simulation results are good (Kuhner and Felsenstein 1994; Gascuel and Levy 1996; Kumar 1996), this approach is not entirely satisfactory, since, to the best of our knowledge, very few results establish a link between the structure of the tree $\hat{\mathbf{T}}$ thus inferred and the structure of the true tree \mathbf{T} . This lacuna is partially filled by another approach, which is widely used in the domain of Evolution (Kidd and Sgaramella-Zonta 1971; Saitou and Nei 1987), and which is called the minimum evolution principle (ME). This principle consists in seeking among all the possible tree structures, that which leads to the "shortest" valued tree. The length of a valued tree is the sum of the lengths (valuations) of its edges, and within the ME principle, these are estimated using the least-squares criterion, without the positivity constraint. The structure of $\hat{\mathbf{T}}$ being fixed, the length of the edges are thus obtained by minimizing the Euclidean distance between $\hat{\mathbf{D}}$ and $\mathbf{\Delta}$. This minimization problem has a unique solution (described below, Section 2), and the length associated with a tree structure is thus well defined. Rzhetsky and Nei (1993) provide a thorough justification of the ME principle. They demonstrate that if the estimates (δ_{ij}) are unbiased, *i.e.*, $E(\delta_{ij}) = (d_{ij})$, then the structure of the true tree \mathbf{T} has among all the possible tree structures, the shortest expected

length. This property justifies the ME principle which, in order to find the true tree, seeks the shortest tree. Thus, the probability of finding a tree of the same structure as the true tree is maximized.

Saitou and Nei (1987) proposed an agglomerative algorithm, called NJ (Neighbor-Joining) which is based on the ME principle. A second version of this algorithm was proposed by Studier and Keppler (1988). It is equivalent to the original version (Gascuel 1994), but simpler. NJ may be described as follows. At each step, one disposes of r nodes, each representing an object or a group of objects agglomerated during the previous steps, on which the tree $\hat{\mathbf{T}}$ remains to be constructed (Figure 1a). A criterion, denoted S , enables one to choose, among the nodes, the pair to be agglomerated. This criterion is, in a way, equal to the least-squares length estimate of the tree represented in Figure 1b. Once the pair, $\{x, y\}$ say, is chosen, we create the node u (Figure 1b) and we determine the length of the edges (x, u) and (y, u) . Then we replace the nodes x and y by the node u in the dissimilarity matrix, by setting $\delta_{ui} = (\delta_{xi} + \delta_{yi})/2$ for each i different from x and y . The process is iterated until r is equal to 3. Finally, we join the 3 remaining objects to a central node and we compute the length of the last 3 edges. This algorithm thus follows the classical scheme of bottom-up hierarchical clustering, already used for tree distances by Sattath and Tversky (1977). Inside this scheme, NJ is defined by the three basic components: (i) the selection criterion for the pair to be agglomerated; (ii) the calculation method for the edge lengths; (iii) the formula enabling the reduction of the dissimilarity matrix.

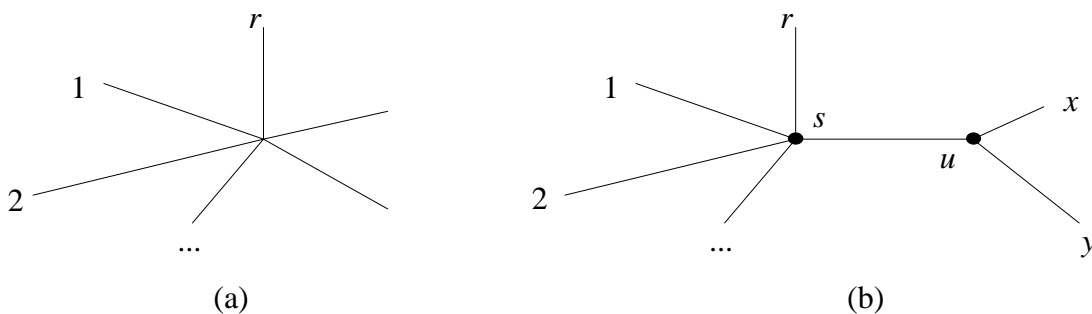


Figure 1: (a) star tree representing absence of structure ; (b) tree in which the pair $\{x, y\}$ has been agglomerated.

Although the NJ algorithm is widely used and has yielded satisfactory simulation results (Nei 1991; Gascuel and Levy 1996), certain questions remain concerning each of its components.

(i) When applied to a tree distance \mathbf{D} , the selection criterion is expected to systematically designate a true pair of neighbors of the corresponding tree. Saitou and Nei's (1987) proof concerning the correctness of the criterion S has been contested by Studier and Keppler (1988) who provide a new proof, which in turn has been contested by Mirkin (1996). Charleston *et al.* (1993) present an interesting study concerning this criterion, but they do not provide a complete proof of its correctness either.

(ii) The manner in which the edge lengths are estimated by NJ is inexact, in terms of the least-squares, when the agglomerated nodes represent not individual objects but rather groups of objects. This deficiency has inspired several authors (Sattath and Tversky 1977; Brölsch 1983; Vach 1989; Vach and Degens 1991; Rzhetsky and Nei 1993) to propose other formulae which are exact, but which are difficult to integrate into the agglomerative procedure.

(iii) Finally, one may question the true meaning of the NJ reduction formula which, systematically, gives identical importance to nodes x and y , even when one corresponds to a group comprising several objects and the other corresponds to a single object. In terms of hierarchical classification, such a reduction is said to be "weighted", meaning that the objects will not have the same influence depending on whether they belong to a large group or are isolated.

This paper will provide answers to each of these questions. In order to make the context precise, we place a model on the data making the hypothesis that the estimates (δ_{ij}) are unbiased, mutually independent and have the same variance. A slightly more specific version of this hypothesis consists in setting $\delta_{ij} = d_{ij} + \varepsilon_{ij}$ for every i and j , where the noise variables ε_{ij} are i.i.d. (independent and identically distributed) and of null expectation. This hypothesis is commonly applied when the estimates are a result of real observations, read measurement errors or imprecise equipment being solely responsible for the noise which hampers the data. Within the scope of this model, we will demonstrate that it is coherent to use an unweighted approach which allocates the same level of importance to each of the initial objects. Furthermore, within this model it is justified to use the "ordinary" least-squares criterion (to estimate the length of the edges) as opposed to the "generalized" least-squares criterion which takes into account the variances and

covariances of the estimates (Searl 1971; Bulmer 1991). Taking into consideration all of the above factors leads to the unweighted version of NJ, which will be called UNJ in what follows.

This paper is organized as follows. Firstly, we will provide a certain number of definitions, notations, and previous results (Section 2). Then, we will describe the UNJ algorithm and its main properties (Section 3). This algorithm follows the same agglomerative scheme as NJ and is defined by the three components described above which will be studied sequentially. Section 4 demonstrates that the selection criterion used by UNJ (identical to that of NJ) retains its meaning whatever the variances and covariances of the δ_{ij} estimates, and that it is correct, in that it always selects a true pair of neighbors when the data are additive. In Section 5, we demonstrate that the formula used by UNJ is correct in the least-squares sense. In order to establish this property, we use a fundamental yet relatively unknown result of Vach (1989), for which we provide a new, very simple proof. In Section 6, we show that the unweighted reduction used by UNJ is coherent with our data model. Section 7 compares performances of NJ and UNJ on simulated data, while Section 8 is devoted to discussion.

2. Preamble

The $\Delta = (\delta_{ij})$ dissimilarity is over the set E of n objects and we denote $E = \{1, 2, \dots, n\}$. The trees considered here have n leaves which are labeled with each of the objects of E . Throughout this paper, a distinction will be made between a valued tree and its structure. Any valued tree will be denoted \mathbf{S} , while its structure will be denoted $\dot{\mathbf{S}}$. The structure of the \mathbf{S} tree is defined by the set of bipartitions of E corresponding to each of its edges (by removing an edge from \mathbf{S} , we cut E into two disjoint subsets). Each of these bipartitions constitutes a pair $\{X, \bar{X}\}$ where X may be viewed in two different ways: as a subset of E (thus we have $\bar{X} = E - X$); or as a rooted subtree of \mathbf{S} whose root is situated at the extremity of the edge in question (\bar{X} is then the subtree situated on "the other side extremity" of the edge). Depending on the context, we will use one or other of these notions. We will discriminate between trivial bipartitions which separate a unique object E from all the remaining objects and which are always contained in the trees studied herein, and non-trivial bipartitions which separate subsets containing at least two objects. The cardinality of X (or equivalently the number of leaves of the tree X) will be denoted n_X . The root of a tree will be

indicated by the corresponding lower case letter, *i.e.*, x will be the root of X , and n_x will sometimes be denoted n_x depending on the context.

The trees inferred by the agglomerative methods such as NJ are binary, *i.e.*, all their internal nodes are of degree 3. Every tree distance may be represented by a tree of this type, provided the zero-valued edges are accepted, so that this characteristic is not restrictive. Nonetheless, such a representation is not always unique, since a 4-degree node may, for example, be represented in three different ways by two 3-degree nodes separated by a zero-valued edge. In order to avoid this type of difficulty, which slows down the demonstrations, we suppose in what follows that the true tree \mathbf{T} is itself a binary tree and contains only edges which are strictly positive. This is restrictive compared with the general case, however it has little practical effect, given that every tree distance may be approached as closely as desired by a tree respecting this condition.

One of our objectives here is to demonstrate a simple expression of the least-squares estimation of the edge lengths of a tree whose structure is fixed. We will refer only to estimation without the positivity constraint, which is justified in the case of the minimum evolution principle. A tree which has negative valuations does not define a distance, but an "unsigned tree dissimilarity" (Bandelt and Steel 1995) in which the dissimilarity between two nodes is simply the sum of the valuations of the path linking these nodes. A Δ matrix being given, the least-squares estimation associates with the tree structure $\dot{\mathbf{S}}$, a valued tree which we will refer to as the "adjusted" tree of $\dot{\mathbf{S}}$, and which we will denote as \mathbf{S} for simplification purposes, Δ being implicit. This adjusted tree is itself associated with an unsigned tree dissimilarity, also simply denoted $\mathbf{S} = (s_{ij})$. The matrix expression of \mathbf{S} as a function of $\dot{\mathbf{S}}$ and of Δ is well known (Sattath and Tversky 1977; Barthélemy and Guénoche 1991). Let us call q the number of edges of $\dot{\mathbf{S}}$, and let us suppose that an order has been chosen for the edges, which need not be stated explicitly for our purpose here. Thus we may represent all the edge lengths by a vector $\mathbf{B} = (b_1, b_2, \dots, b_q)$. \mathbf{S} may also be represented by a vector, and we have

$$\mathbf{S}' = \mathbf{A}\mathbf{B}' , \tag{1}$$

where \mathbf{A} is a 0-1 matrix $(n(n-1)/2) \times q$ which represents $\dot{\mathbf{S}}$.

This matrix is defined in the following manner:

$$\mathbf{A}_{(ij)k} = 1 \text{ if the } k\text{th edge (or bipartition) of } \dot{\mathbf{S}} \text{ separates } i \text{ and } j,$$

$$\mathbf{A}_{(ij)k} = 0 \text{ otherwise,}$$

where (ij) represents the rank of s_{ij} in the vector representation of \mathbf{S} . In this framework, \mathbf{S} is the projection of Δ on the sub-space generated by the bipartitions of $\dot{\mathbf{S}}$, and we get

$$\mathbf{S}^t = \mathbf{A}(\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t \Delta^t. \quad (2)$$

The drawback in this expression is that it requires the computation of the matrix product $\mathbf{A}^t \mathbf{A}$, whose computational complexity in time is $O(n^2 q^2)$. Since the trees under consideration are binary, we get $q = 2n - 3$, and the computation is thus in $O(n^4)$. However, as we will see below (Section 5), this expression (2) is fundamental to analytical formulae which enable a more rapid computation of edge lengths, in $O(n^2)$.

Before concluding this section, we introduce some notation. Let \mathbf{S} be an adjusted tree, and $\{X, \bar{X}\}$ and $\{Y, \bar{Y}\}$ two bipartitions of $\dot{\mathbf{S}}$. When $X \cap Y = \emptyset$, we set by extension

$$\delta_{XY} = \sum_{i \in X, j \in Y} \delta_{ij} \quad \text{and} \quad \overline{\delta_{XY}} = \frac{1}{n_X n_Y} \sum_{i \in X, j \in Y} \delta_{ij},$$

as well as

$$s_{XY} = \sum_{i \in X, j \in Y} s_{ij} \quad \text{and} \quad \overline{s_{XY}} = \frac{1}{n_X n_Y} \sum_{i \in X, j \in Y} s_{ij}.$$

We will also refer to the "flow" in \mathbf{S} of a rooted subtree X . This quantity, denoted f_X , is the sum of the lengths of the paths between each leaf of X and the root of X . Let x be this root, we get

$$f_X = \sum_{i \in X} s_{ix} \quad \text{and} \quad \overline{f_X} = \frac{1}{n_X} f_X.$$

3. The UNJ algorithm

The UNJ algorithm is summarized in Figure 2. At each step, it uses and then reduces a running matrix of size $r \times r$ that we have denoted $\Lambda = (\lambda_{ij})$, in order to avoid confusion with the initial data matrix. This matrix is of course initialized with value Δ . UNJ is defined by the following three formulae:

selection criterion

$$Q_{xy} = R_x + R_y - (r-2)\lambda_{xy}, \text{ with } R_z = \sum_{i=1}^r \lambda_{zi}, \quad (3)$$

estimation formula

$$\hat{d}_{xu} = \frac{1}{2} \lambda_{xy} + \frac{1}{2(n-n_u)} \sum_{\substack{i=1 \\ \neq x,y}}^r n_i (\lambda_{xi} - \lambda_{yi}), \hat{d}_{yu} \text{ obtained by symmetry, and} \quad (4)$$

reduction formula

$$\lambda_{ui} = w_x \lambda_{xi} + w_y \lambda_{yi} - w_x \hat{d}_{xu} - w_y \hat{d}_{yu}, \text{ where } w_x = \frac{n_x}{n_u} \text{ and } w_y = \frac{n_y}{n_u}. \quad (5)$$

Initialize the running matrix: $\Lambda = (\lambda_{ij}) \leftarrow (\delta_{ij})$;

Initialize the number of remaining nodes: $r \leftarrow n$;

Initialize the numbers of objects per node: $n_i \leftarrow 1, i \in \{1, \dots, n\}$;

While the number of nodes r is greater than 3 :

{ Compute the sums $R_i, i \in \{1, \dots, r\}$; (a)

Find the pair $\{x, y\}$ to be agglomerated by maximizing Q_{xy} (3) ; (b)

Create the node u , and set: $n_u \leftarrow n_x + n_y$;

Estimate the lengths of edges (x, u) and (y, u) using (4) ; (c)

Reduce the running matrix Λ using (5) ; (d)

Decrease the number of nodes: $r \leftarrow r-1$ } ;

Create a central node, and compute the last three edge-lengths using (4) ;

Output the tree found.

Figure 2: The UNJ algorithm.

These three formulae may be easily compared with those proposed by Studier and Keppler (1988) in their simplified version of NJ. The selection criterion is the same. The estimation formula of UNJ is an unweighted version of that of NJ. Indeed, the latter may be expressed as

$$\hat{d}_{xu} = \frac{1}{2} \lambda_{xy} + \frac{1}{2(r-2)} \sum_{\substack{i=1 \\ \neq x,y}}^r (\lambda_{xi} - \lambda_{yi}),$$

and is obtained from (4) by setting $n_i = 1, \forall i \neq x, y$, and by replacing $(n - n_u)$ by $\sum_{i \neq x,y} n_i$ which is then equal to $(r - 2)$. Similarly, the NJ reduction formula is obtained from (5) by setting $n_x = n_y = 1$. UNJ, as described here, thus appears as the unweighted version of the NJ algorithm proposed by Studier and Keppler (1988). Its properties are as follows.

Property 3.1: The complexity in time of UNJ is $O(n^3)$.

Property 3.2: The estimation formula (4), combined with the reduction (5), is optimal, in that whatever the structure of the inferred tree, the edge lengths obtained are identical to those obtained with the matrix solution (2). Furthermore, this property may be exploited within an algorithm in $O(n^2)$, allowing the least-squares estimation of edge lengths of any fixed structure binary tree.

Property 3.3: Given our hypothesis on the δ_{ij} estimates, the reduction (5) is optimal in that it minimizes the part of the variance of the running matrix (λ_{ij}) which influences the choice of the structure of $\hat{\mathbf{T}}$. In other words, at each step, it yields estimators which are as reliable as possible in choosing the pair to be agglomerated.

Property 3.4: When data are additive, *i.e.*, $\Delta = \mathbf{D}$, UNJ systematically finds the true tree \mathbf{T} .

Property 3.1 is immediate. In fact, the algorithm carries out $n - 3$ steps, and during each step the most costly operations correspond to lines (a) and (b) which are both in $O(r^2)$. The complexity of UNJ is thus the same as that of NJ.

The first part of property 3.2. will be shown in Section 5, while the second part is simple. Indeed, the UNJ algorithm may be transformed into an estimation algorithm of edge lengths of a binary tree with a fixed structure, whose complexity is in $O(n^2)$. To do this, one simply replaces lines (a) and (b) by a tree traversing algorithm which finds in $O(r)$ a pair of neighbors of the tree in

question; the most costly remaining lines are now (c) and (d), both being in $O(r)$. This $O(n^2)$ complexity is optimal, since it is linear in the size of the data. An analogous result may be obtained using one (5.3) of the formulae proposed by Vach and Degens (1991). It would appear however that this type of result is relatively unknown, and that users generally opt for the matrix solution, in $O(n^4)$, or for the algorithm of Rzhetsky and Nei (1993) which is in $O(n^3)$. We would like to point out however that formula (4) and this $O(n^2)$ algorithm only apply to binary trees. For non-binary trees, we have to use a more general formula (+) proposed by Vach (1989), and the corresponding algorithm remains to be studied.

Property 3.3 seems natural since the notions of unweighted mean, of i.i.d. variables and of minimum variance estimator are fundamentally linked. However, this type of argument based on a data model is rarely reviewed in the literature. For example, it does not figure in the comparison between UPGMA and WPGMA proposed by Sneath and Sokal (1973). This property (3.3) will be made precise and proved in Section 6.

Property 3.4 is special in that it requires three sub-properties in order to be demonstrated, relating to formulae (3), (4) and (5). Under the hypothesis that the data are additive, *i.e.* $\Delta = \mathbf{D}$, and that the tree \mathbf{T} representing \mathbf{D} is a binary tree comprising only strictly positive valuations, we will show that:

Property 3.5: Criterion (3) always selects a true pair of neighbors in the tree \mathbf{T} , *i.e.*, a pair $\{x, y\}$ of leaves linked by a path containing only one interior node, denoted u .

Property 3.6: The lengths of the edges (x, u) and (y, u) resulting from (4) are identical to those of the corresponding edges in \mathbf{T} .

Property 3.7: Reduction (5) applied to a true neighbor pair $\{x, y\}$ transforms \mathbf{D} into an additive matrix \mathbf{D}' which is represented by the subtree of \mathbf{T} obtained by deleting the nodes x and y and the corresponding edges.

The combination of these three properties enables us to obtain the result (3.4) by induction, each step of the algorithm consisting in reconstructing correctly (from the point of view of structure and length) two neighboring edges of \mathbf{T} . At the end, only three nodes remain, and hence only one possible structure, and the computation of the lengths is correct due to property 3.6.

Property 3.5 is shown in Section 4, which fills the lacuna indicated by Mirkin (1996) as outlined in our Introduction. Property 3.6 is an immediate consequence of property 3.2, given that the matrix solution (2) is obviously exact in the case of additive data. Finally, we will show, in Section 6, a generalization of the property 3.7 which enables the correction of an entire class of variants of NJ, including among others, NJ itself, UNJ and BIONJ (Gascuel 1996).

4. Concerning the Q selection criterion

UNJ retains the same criterion as NJ to select the node pair to be agglomerated. This criterion is open to several interpretations and several expressions which we will recall briefly (§4.1). Next (§4.2), we will present, in greater detail, the interpretation proposed by Mirkin (1996) and we will show how this interpretation has the advantage of retaining its full meaning when we drop the hypothesis that the δ_{ij} estimates are mutually independent and that they have the same variance. Finally (§4.3), we show that this interpretation leads to a simple proof of the correction of the Q criterion (Property 3.5).

4.1 The different expressions and interpretations of the Q criterion

At each step, we consider the dissimilarity matrix (λ_{ij}) where i and j represent objects or object clusters already agglomerated during the previous steps. Saitou and Nei (1987), at each step, assimilate these indices with unique objects, and they choose the pair $\{x, y\}$ which minimizes the least-squares length estimate of the tree represented in Figure 1b. The criterion thus defined is denoted S_{xy} , and is expressed

$$S_{xy} = \frac{1}{2} \lambda_{xy} + \frac{1}{2(r-2)} \sum_{\substack{i=1 \\ \neq x,y}}^r (\lambda_{1i} + \lambda_{2i}) + \frac{1}{r-2} \sum_{\substack{i=1 \\ \neq x,y}}^r \sum_{\substack{j=i \\ \neq x,y}}^r \lambda_{ij}. \quad (6)$$

Studier and Keppler (1988) propose replacing the S criterion by the Q criterion defined above (3), without attaching any specific interpretation to the latter. This new expression has the advantage of leading to a complexity in $O(n^3)$, as shown above (Property 3.1). Furthermore, it is equivalent to the original expression (6), criteria S and Q being linked by a negative slope linear expression (Gascuel 1994). Vach and Degens (1991) indicate that minimizing S (or maximizing Q) is equivalent to maximizing the least-squares length estimate of the edge linking u , the root of the

cluster in formation, and s , the center of the star (Figure 1b). We have also shown (Gascuel 1994) that S could be interpreted as a continuous version of the neighborliness criterion proposed by Sattath and Tversky (1977).

The tree 1b does not represent the structure of the true tree. By minimizing at each step its estimated length, one tends to find a short tree, which justifies the use of the S criterion within a greedy algorithm which follows the ME principle. However, the real reason why we systematically obtain the true tree (*i.e.*, the shortest tree) when the data are additive is, to a great extent, inexplicable. The same applies to the other interpretations proposed, which thus seem unsatisfactory. Moreover, the interpretations based on the length of the tree (Saitou and Nei 1987) or of the internal edge (Vach and Degens 1991) lead to a form of contradiction. Indeed, formula (6), used in estimating the length of tree 1b, is only correct at the first step, when each index designates a unique object. During the following steps, some indices represent object clusters, and this formula becomes approximate. The same applies to the length of the edge (u, s) , and Vach and Degens (1991) use an exact formula which is no longer linked, in a linear manner, to S . An optimal formula also exists which gives the least-squares length estimate of tree 1b at each step of the algorithm (available on demand). These optimal formulae are a priori more satisfying than the previous ones, if we accept the proposed interpretations. However, we observe with examples, that neither one nor the other guarantees the finding of the true tree when the data are additive, and hence the announced contradiction.

4.2 Interpreting Q in acentrality terms

Mirkin (1996) proposes another interpretation of Q , in acentrality terms. Let us consider for the moment that Q is applied to the tree distance \mathbf{D} . Now $Q'_{xy} = Q_{xy}/2$ may be expressed as

$$Q'_{xy} = d_{xy} + \sum_{\substack{i=1 \\ \neq x,y}}^r \frac{1}{2} (d_{xi} + d_{yi} - d_{xy}). \quad (7)$$

It is easy to see that the expression inside the sum, *i.e.*, $(d_{xi} + d_{yi} - d_{xy})/2$, is equal to the length of the path (i, u) , where u (Figure 3a) is the intersection of the paths (x, y) , (x, i) and (y, i) . In other words, this expression measures the acentrality of the path (x, y) for the node i , and Q' is equal to the sum of all these measures, to which d_{xy} is added, which expresses the acentrality of nodes x and y themselves. Q'_{xy} is thus an acentrality measurement of the pair $\{x, y\}$.

Let us consider the examples shown in Figures 3b and 3c. In the first case, we examine the pair of neighbors $\{i, j\}$, while in the second case, we examine the pair $\{i, k\}$. In the case of $\{i, j\}$, Q' is equal to the sum $(d_{ij} + d_{ik} + d_{il} + d_{im})$ in which each external edge is counted once, while the edge lengths d_{tu} and d_{uv} are counted, respectively, three times and twice. In the case of $\{i, k\}$, Q' is equal to the sum $(d_{ik} + d_{ij} + d_{ul} + d_{um})$ in which the external edges are always counted once, but the edge lengths d_{tu} and d_{uv} are now counted, respectively, once and twice. It is evident from this example that when applied to a pair of neighbors, the criterion counts the internal edges several times, whereas when applied to remote nodes in the tree, the criterion tends to count certain edges less often. By applying the criterion to the pair $\{i, l\}$, or to any pair situated at either extremity of the tree, each edge is only counted once.

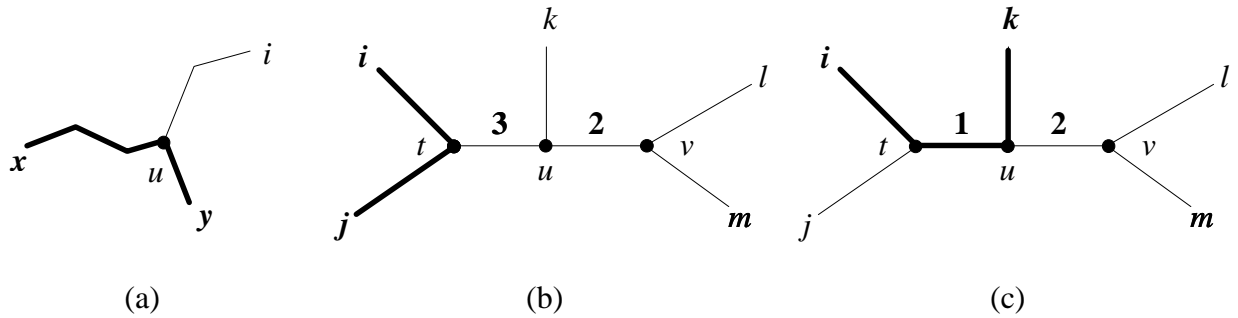


Figure 3: (a) the interior expression in (7) is equal to the length of the path (i, u) ; (b) when the pair $\{i, j\}$ is examined, d_{tu} is counted 3 times and d_{uv} twice; (c) when the pair $\{i, k\}$ is examined, d_{tu} is counted once and d_{uv} twice.

Observations made in this example are demonstrated below (§4.3) in the general framework. Q is therefore a numerical criterion which, when applied to any tree distance \mathbf{D} , designates a pair of neighbors of the tree \mathbf{T} which represent this distance. In practice, we do not dispose of the tree distance \mathbf{D} , but of its estimation $\mathbf{\Delta}$, and the formula (3) is applied to the estimates δ_{ij} (or λ_{ij}) rather than to the distances d_{ij} . In reality, we thus use an estimator \hat{Q} of the true value of the Q criterion. This estimator \hat{Q} has an obvious quality which has not been highlighted by Mirkin (1996) yet appears to be very important: it is obtained without making any hypotheses on the δ_{ij} estimates. In fact, it is easy to see (Bulmer 1991) that $(\delta_{i_1} + \delta_{i_2} - \delta_{i_1 i_2})$ is the generalized least-squares estimator of the sum $(d_{i_1} + d_{i_2} - d_{i_1 i_2})$. In other words, \hat{Q} is the sum of generalized least-squares estimators which each measure the acentrality of the pair to be agglomerated. This does not imply that the estimator \hat{Q} thus obtained is equal, for any given tree structure, to the

generalized least-squares estimator of Q . Indeed, for a given tree structure and a given variance-covariance matrix, it is usually possible to find an estimator of Q whose variance is smaller than the variance of \hat{Q} . However, we can prove that it is not possible to have an unbiased estimator other than \hat{Q} , without already knowing the structure of the tree. Consequently in our context, \hat{Q} seems to be the only possible estimator of Q , and whatever the nature of the variances and covariances of the δ_{ij} estimates, its use appears well founded. We therefore reach a different interpretation from that of Saitou and Nei (1987) and Vach and Degens (1991) who rely on ordinary least-squares, thus losing part of the meaning when the hypothesis of variance identity and covariance nullity is dropped. Thus the use of Q is justified in approaches which drop this hypothesis (Gascuel 1996). The use of Q is also justified within the scope of the hypothesis made in this paper, because after a certain number of steps the estimators λ_{ij} no longer have the same variance due to the mean process (5). Moreover, we can expect a certain robustness of the method, at least concerning its capacity to find the structure of the true tree.

4.3 Proving the correctness of the Q criterion (Property 3.5)

Let \mathbf{D} be a tree distance represented by the tree \mathbf{T} whose valuations are strictly positive, and whose internal nodes are at least of degree 3. Let us consider the Q' criterion ($= Q/2$) defined above (7), allowing the interpretation proposed in terms of path length of \mathbf{T} . Let us suppose that Q' designates the pair $\{1,2\}$ and that this does not consist of true neighbors in \mathbf{T} . The path (1,2) thus comprises at least two nodes different from nodes 1 and 2. Among these, let us consider nodes u and v , one edge distant from 1 and from 2 respectively (Figure 4). Each of these nodes is the root of the subtree denoted \mathbf{T}_u , respectively \mathbf{T}_v , and two cases can occur: either \mathbf{T}_u or \mathbf{T}_v contains only one leaf (Figure 4a); or these two subtrees each comprise several leaves (Figure 4b). We will demonstrate that in both cases a pair of neighbors exists, whose value for the criterion Q' is strictly greater than the value of the pair $\{1,2\}$.

Case 3a: Let us suppose that \mathbf{T}_u contains only one leaf (the argument is symmetrical for \mathbf{T}_v), and let us consider the notations introduced in Figure 4a. We will demonstrate that the pair of neighbors $\{1,3\}$ has a better score than $\{1,2\}$. We have

$$Q'(\{1,2\}) = d_{12} + d_{3u} + \sum_{i=4}^r d_{ii}.$$

Likewise, we also have

$$\begin{aligned}
Q'(\{1,3\}) &= d_{13} + d_{2u} + \sum_{i=4}^r (d_{ui'} + d_{ii'}) \\
&= d_{12} + d_{3u} + \sum_{i=4}^r (d_{ui'} + d_{ii'}) \\
&= Q'(\{1,2\}) + \sum_{i=4}^r d_{ui'}.
\end{aligned}$$

The \mathbf{T}_v tree comprises at least one leaf, therefore the last sum is greater than or equal to d_{uv} which, by hypothesis, is strictly positive, and the result is demonstrated.

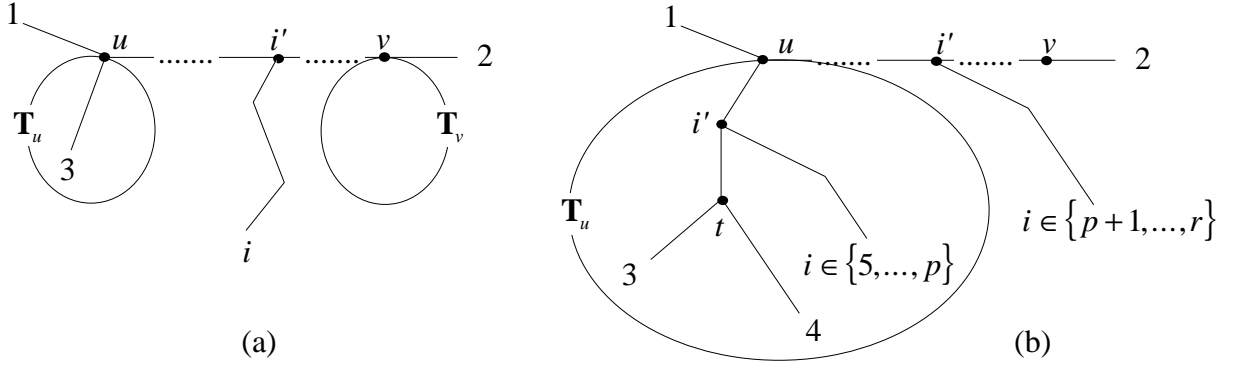


Figure 4: (a) \mathbf{T}_u comprises only the leaf denoted 3; $i \in \{4, \dots, r\}$ is thus any leaf and i' designates the intersection of the paths $(1,2)$, $(1,i)$ and $(2,i)$. (b) \mathbf{T}_u comprises a pair of neighbors, denoted $\{3,4\}$, as well as the leaves of index i varying from 5 to p ; i' is thus the intersection of paths $(1,3)$, $(1,i)$ and $(3,i)$; the remaining leaf indices vary from $p+1$ to r ; i' thus designates the path intersection $(1,2)$, $(1,i)$ and $(2,i)$.

Case 3b: If \mathbf{T}_u and \mathbf{T}_v each comprise at least two leaves, at least one of these subtrees contains at the most $(r-2)/2$ leaves. Let us suppose that it is \mathbf{T}_u (the argument is symmetrical for \mathbf{T}_v). \mathbf{T}_u necessarily contains a true pair of neighbors, denoted $\{3,4\}$. We will demonstrate that this pair is better than $\{1,2\}$. Let us consider the notations introduced in Figure 4b. We have

$$Q'(\{1,2\}) = d_{12} + d_{3u} + d_{4u} + \sum_{i=5}^p (d_{ui'} + d_{ii'}) + \sum_{i=p+1}^r d_{ii'}.$$

Likewise, we also have

$$\begin{aligned}
\mathcal{Q}'(\{3,4\}) &= d_{34} + d_{1t} + d_{2t} + \sum_{i=5}^p (d_{tu} - d_{ui'} + d_{ii'}) + \sum_{i=p+1}^r (d_{tu} + d_{ui'} + d_{ii'}) \\
&= d_{12} + d_{3u} + d_{4u} + \sum_{i=5}^r d_{ii'} + \sum_{i=p+1}^r d_{ui'} + (r-4)d_{tu} - \sum_{i=5}^p d_{ui'} \\
&= \mathcal{Q}'(\{1,2\}) + \sum_{i=p+1}^r d_{ui'} + A, \quad \text{where } A = (r-4)d_{tu} - 2\sum_{i=5}^p d_{ui'}.
\end{aligned}$$

As before, we have $\sum_{i=p+1}^r d_{ui'} \geq d_{uv} > 0$.

Therefore, all that needs to be shown is: $A \geq 0$. By hypothesis, we know that \mathbf{T}_u contains at most $(r-2)/2$ leaves, including leaves 3 and 4. Moreover, we know that for every $i \in \{5, \dots, p\}$ we have $d_{ui'} \leq d_{tu}$. Consequently

$$A \geq (r-4)d_{tu} - 2\left(\frac{(r-2)}{2} - 2\right)d_{tu} = 2d_{tu} \geq 0,$$

and the result is demonstrated. \square

We would like to point out that the case $A = 0$ may occur when the node u is of a degree greater than 3, and when the node t is equal to node u . However, this does not invalidate the proof, which holds when the degree of nodes internal to \mathbf{T} is at least 3. This latter hypothesis is used, effectively, since we assume that the trees \mathbf{T}_u and \mathbf{T}_v each comprise at least one leaf. The case where the internal nodes may be of degree 2 requires special treatment (as well as the redefinition of the notion of neighbor).

5. Least-squares estimation of edge lengths

It would seem that Sattath and Tversky (1977) were first to find an analytical expression for the least-squares estimation of edge lengths of a tree with fixed structure; however, this is not explicitly described in their article. Brölsch (1983), and subsequently Vach (1989) and Vach and Degens (1991) have demonstrated a certain number of general properties of this estimation, and

they explicitly propose several exact formulae. One of the formulae (5.3) provided by Vach and Degens (1991) is equivalent to formula (4), but more complex due to the use of a reduction formula which differs from (5). It is given below (Equation 10), as well as the proof of its equivalence with formula (4). Likewise, Rzhetsky and Nei (1993) independently found an exact formula which differs from formula (4), yet is identical with another (4.6) among the formulae of Vach and Degens (1991). This latter formula cannot be integrated into the agglomerative procedure, and it is used in a specific algorithm, in $O(n^3)$, to estimate the edge lengths once the structure has been fully determined.

As has been mentioned above, these studies seem to be relatively unknown and infrequently used. For this reason, we have tried to render this section sufficiently explicit and autonomous. First, in (§5.1), we will describe the fundamental result of Vach (1989) for which we provide a new (to the best of our knowledge) and simple proof. Then in (§5.2), using this result, we will demonstrate (Property 3.2) the correctness of formula (4).

5.1 Conservation properties

Let $\Delta = (\delta_{ij})$ be a dissimilarity, $\dot{\mathbf{S}}$ a tree structure, and \mathbf{S} the corresponding adjusted tree (or unsigned tree dissimilarity). Using the notation defined above (Section 2), we have (Vach 1989)

Property 5.1: For every bipartition $\{X, \bar{X}\}$ of $\dot{\mathbf{S}}$, we have $s_{X\bar{X}} = \delta_{X\bar{X}}$ (and $\overline{s_{X\bar{X}}} = \overline{\delta_{X\bar{X}}}$).

In other words, the mean dissimilarity between the elements of X and those of \bar{X} is identical in Δ and in the adjusted tree \mathbf{S} , this being valid for every bipartition of $\dot{\mathbf{S}}$. This property is established very simply from the matrix solution (2). Combining this with equality (1), we obtain directly

$$\mathbf{A}^t \mathbf{S}^t = \mathbf{A}^t \Delta^t. \quad (8)$$

The coefficients of \mathbf{A}^t are expressed as $\mathbf{M}_{k(ij)}$ and have value 1, if and only if the k th edge of $\dot{\mathbf{S}}$ separates i and j . The k th line of the equation (8) thus establishes equality between the sum of the dissimilarities s_{ij} on the one hand, and δ_{ij} on the other hand, provided that i and j are separated by the k th edge. In other words, Property 5.1 is established for the bipartition associated with the k th edge; and since each bipartition (or edge) of $\dot{\mathbf{S}}$ is represented by a line of \mathbf{A}^t , Property 5.1 is established. \square

There is another conservation property associated with degree 3 (or ternary) nodes, which has not been referred to by Vach (1989), but which is useful to derive edge length estimates. A ternary node u is the extremity of three edges (u, x) , (u, y) and (u, z) , and it is associated with the three rooted subtrees X , Y and Z of which x , y and z are the respective roots. We may thus express the following property relative to these subtrees :

Property 5.2: For all ternary nodes of $\dot{\mathbf{S}}$ and for every pair X, Y of subtrees associated with this node, we have $s_{XY} = \delta_{XY}$ (and $\overline{s_{XY}} = \overline{\delta_{XY}}$).

In other words, the mean dissimilarity between the subtrees associated with a ternary node is also preserved. This property is established simply from Property 5.1. Let X, Y and Z be the three subtrees associated with u . According to 5.1 we have

$$s_{X\bar{X}} = \delta_{X\bar{X}}, \quad s_{Y\bar{Y}} = \delta_{Y\bar{Y}} \quad \text{and} \quad s_{Z\bar{Z}} = \delta_{Z\bar{Z}}.$$

This result and the definition of these quantities enable us to write the three equations

$$\begin{aligned} s_{XY} + s_{XZ} &= \delta_{XY} + \delta_{XZ}, \\ s_{XY} + s_{YZ} &= \delta_{XY} + \delta_{YZ}, \\ s_{XZ} + s_{YZ} &= \delta_{XZ} + \delta_{YZ}, \end{aligned}$$

whose unique solution corresponds to Property 5.2. □

We would like to point out that it is essential to Property 5.2 that the node be ternary. When u is of degree $g (>3)$, a system of g equations is obtained (one per edge) with $g(g-1)/2$ "unknowns" (one per pair of subtrees), and this system may be satisfied without the mean dissimilarity between subtrees being conserved. It is also worth noting that this result is a generalization of a well known result in the case of ultrametric distances obtained by least-squares adjustment: the mean distance between two neighboring clusters is identical in the observed dissimilarity and in the ultrametric obtained, where it is represented by the level of formation. Finally, let us note that, as with ultrametrics, the dissimilarity between two neighbor leaves x and y linked by a ternary node remains unchanged, *i.e.*, $s_{xy} = \delta_{xy}$.

5.2 Formula (4) is correct (Property 3.2)

The results above directly yield correct least-squares formulae. Let u be a ternary node associated with the three subtrees X, Y and Z whose roots are x, y and z respectively. Using Property 5.2 and

the definitions given above (Section 2), we can write the three equations

$$\begin{aligned}\delta_{XY} &= s_{XY} = n_Y f_X + n_X n_Y (s_{xu} + s_{yu}) + n_X f_Y, \\ \delta_{XZ} &= s_{XZ} = n_Z f_X + n_X n_Z (s_{xu} + s_{zu}) + n_X f_Z, \\ \delta_{YZ} &= s_{YZ} = n_Z f_Y + n_Y n_Z (s_{yu} + s_{zu}) + n_Y f_Z.\end{aligned}$$

And solving these equations, we find

$$s_{xu} = \frac{1}{2} \overline{\delta_{XY}} + \frac{1}{2} \overline{\delta_{XZ}} - \frac{1}{2} \overline{\delta_{YZ}} - \overline{f_X}, \quad s_{yu} \quad \text{and} \quad s_{zu} \quad \text{obtained by symmetry.} \quad (9)$$

We will now consider an agglomerative procedure as described in Section 3, having as its objective the estimation of the edge lengths of $\dot{\mathbf{S}}$. At each step, the algorithm selects two neighboring edges of $\dot{\mathbf{S}}$, estimates the length of these edges using formula (4), then reduces the matrix using formula (5). At the p th step, it disposes of $r = n - p + 1$ subtrees situated at the "periphery" of $\dot{\mathbf{S}}$ whose edge lengths have already been computed, and all that remains to be estimated is the length of the edges situated at the "center" of $\dot{\mathbf{S}}$. Suppose that we are at step p , and that X and Y are the two trees to be agglomerated. Z represents the "rest" of the tree (whose structure is still unknown if considered from the point of view of UNJ). Taken as a set, Z regroups the objects of the $r - 2$ subtrees different from X and from Y and which have been already resolved, so that we have

$$Z = \bigcup_{I \neq X, Y} I,$$

where I is any one of the subtrees (subsets) already resolved. We may now rewrite the equation (9) as

$$s_{xu} = \frac{1}{2} \overline{\delta_{XY}} + \frac{1}{2(n - n_U)} \sum_{I \neq X, Y} n_I (\overline{\delta_{XI}} - \overline{\delta_{YI}}) - \overline{f_X}, \quad \text{where } n_U = n_X + n_Y = n - \sum_{I \neq X, Y} n_I. \quad (10)$$

This formula (10) is correct, and corresponds to formula (5.3) of Vach and Degens (1991). We now need only to show that, at each step of the algorithm, it coincides with formula (4), when the latter is combined with the reduction (5). At the first step, no agglomeration has been achieved, thus $\overline{f_I} = 0$ and $\lambda_{ij} = \overline{\delta_{IJ}} - \overline{f_I} - \overline{f_J}$ for every "subtree" I, J whose respective roots are i and j . Now it is easy to check that the two formulae coincide.

Let us consider the p th step, just before reduction (5), and suppose that

- (a) $\lambda_{ij} = \overline{\delta_{IJ}} - \overline{f_I} - \overline{f_J}$ for every resolved subtree I, J whose respective roots are i and j ;
- (b) formula (4) and equation (10) have coincided during the previous computations.

We will now demonstrate that applying the reduction (5) maintains (a), and that (a) being maintained, formulae (4) and (10) coincide during the next step. This means that the hypotheses (a) and (b) are maintained at step $p+1$ (just before the reduction). By induction, the desired result will follow for all the steps of the algorithm, including the last (the estimation of the last three edge lengths) which is really not different from the preceding steps. Let us now begin with the first point.

(a) is maintained: As above, the two agglomerated subtrees are denoted X and Y , and they form a new subtree denoted U . We must check that (a) is maintained for the new coefficients λ_{ui} which are obtained by application of reduction (5). We have

$$\begin{aligned}\lambda_{ui} &= w_x \lambda_{xi} + w_y \lambda_{yi} - w_x s_{xu} - w_y s_{yu} \\ &= w_x (\overline{\delta_{XI}} - \overline{f_X} - \overline{f_I}) + w_y (\overline{\delta_{YI}} - \overline{f_Y} - \overline{f_I}) - w_x s_{xu} - w_y s_{yu} \\ &= \overline{\delta_{UI}} - \overline{f_U} - \overline{f_I}.\end{aligned}$$

The transition from the first to the second line uses (a), whereas to establish the value of $\overline{f_U}$ we have used (b) which supposes, in particular, the correct computation of s_{xu} et s_{yu} . This completes the proof of the first point.

(b) is maintained: For the sake of simplicity, let us again denote as X, Y and U the two agglomerated subtrees and the tree thus formed, even though these are not the same trees as those mentioned above, since we are now at the subsequent step. Utilizing the fact that (a) is maintained, equation (10) can be rewritten in the form

$$\begin{aligned}s_{xu} &= \frac{1}{2} (\lambda_{xy} + \overline{f_X} + \overline{f_Y}) + \frac{1}{2(n-n_U)} \sum_{I \neq X, Y} n_I (\lambda_{xi} + \overline{f_X} + \overline{f_I} - \lambda_{yi} - \overline{f_Y} - \overline{f_I}) - \overline{f_X} \\ &= \frac{1}{2} \lambda_{xy} + \frac{1}{2(n-n_U)} \sum_{I \neq X, Y} n_I (\lambda_{xi} - \lambda_{yi}),\end{aligned}$$

which corresponds precisely to formula (4), and the proof is completed. \square

6. Reducing the dissimilarity matrix

In the previous section, we saw that the use of reduction (5) is fully justified from the point of view of the least-squares estimation of the edge lengths. We will now propose a second justification, based on the model placed on the data. First, in (§6.1), we show that reduction (5) belongs to a class of "admissible" reduction formulae, in the sense that they guarantee that the true tree to will be found with additive data, when combined with selection criterion (3) and with a correct estimation formula. Then, in (§6.2), we show that among these admissible formulae, reduction (5) corresponds, in some sense, to the minimum variance reduction.

6.1 An admissible reduction formula class

Let us consider Figure 5 : we dispose of a tree distance \mathbf{D} , represented by tree \mathbf{T} ; the selection criterion (3) systematically designates a pair of neighbors of this tree, denoted $\{x, y\}$, u being the internal node separating these two leaves ; then, the estimation formula consists in computing the lengths of the edges (x, u) and (y, u) . Let us assume that this formula is correct in the case of additive data, as is the case for formula (4) and for the formula used by NJ, and numerous other possible formulae. We will say that a reduction is *admissible* if it transforms distance \mathbf{D} into distance \mathbf{D}' which is represented by the subtree \mathbf{T}' in which u is now a leaf. Given the properties of the selection criterion and the estimation formula, it is clear that this prerequisite is sufficient to guarantee the finding of the true tree with additive data.

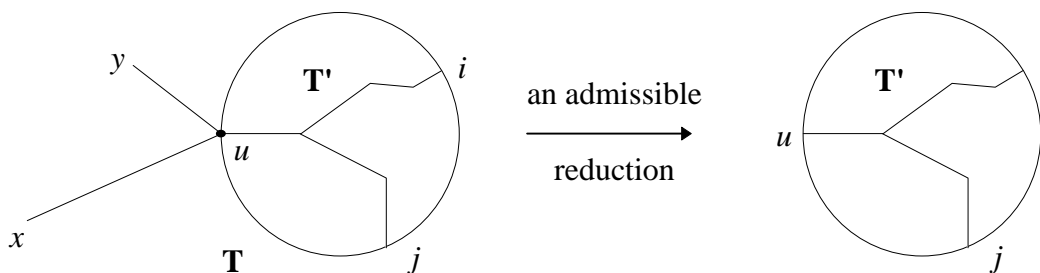


Figure 5: An admissible reduction transforms the tree distance represented by \mathbf{T} , into the tree distance represented by \mathbf{T}' .

An admissible reduction class is defined by the following generic formula

$$\lambda_{ui} = \mu\lambda_{xi} + (1-\mu)\lambda_{yi} - \mu\hat{d}_{xu} - (1-\mu)\hat{d}_{yu}, \quad (11)$$

where μ is any real number. Indeed, when this reduction is applied to the tree distance \mathbf{D} , distances d_{ij} ($i, j \neq x, y$) remain unchanged, while the dissimilarities λ_{ui} newly introduced satisfy

$$\begin{aligned} \lambda_{ui} &= \mu d_{xi} + (1-\mu)d_{yi} - \mu\hat{d}_{xu} - (1-\mu)\hat{d}_{yu} \\ &= \mu(d_{xi} - d_{xu}) + (1-\mu)(d_{yi} - d_{yu}) \\ &= d_{ui}, \end{aligned}$$

the transition from the first to the second line relying on the fact that the estimation formula is correct, whereas the transition from the second to the third is based on the fact that x and y are neighbors in \mathbf{T} .

The NJ reduction formula as formulated by Studier and Keppler (1988) is obtained from expression (11) with $\mu = 1/2$, and that of UNJ with $\mu = n_x / (n_x + n_y)$. An infinite number of other possible solutions exist. The result found here is analogous to that of Bandelt and Dress (1986) on the "convex" versions of the ADDTREE algorithm of Sattath and Tversky (1977). From our point of view, preference for a given NJ version over another version should be based on a data model. As we will see below, this enables us to choose the minimum variance reduction, in other words, the reduction which provides the more reliable estimates to select the pairs of taxa to be agglomerated during the next steps.

6.2 The minimum variance reduction

First of all, it is to be noted that in expression (11) there is a first part $(\mu\lambda_{xi} + (1-\mu)\lambda_{yi})$ which depends on i while the second part $(-\mu\hat{d}_{xu} - (1-\mu)\hat{d}_{yu})$ is identical for every i . Moreover, we can easily demonstrate that if we add a constant k to the reduction formula, then at the following step the selection criterion (3) is increased by $2k$ for every pair $\{x, y\}$, so that the addition of this constant does not affect the choice of the following agglomerations, and does not therefore influence the structure of the tree under construction. Consequently, only the variance of the first two terms has an influence on the structure of this tree.

This variance will be qualified as structural, and for every index i ($\neq x, y$) it is expressed

$$V(\lambda_{ui}) = \mu^2 V(\lambda_{xi}) + (1 - \mu)^2 V(\lambda_{yi}) + 2\mu(1 - \mu)\text{COV}(\lambda_{xi}, \lambda_{yi}). \quad (12)$$

When we try to minimize the sum of the structural variances induced by the reduction, we end up with a second degree polynomial in μ , whose minimum is achieved for

$$\mu^* = \frac{\sum_{\substack{i=1 \\ \neq x, y}}^r (V(\lambda_{yi}) - \text{COV}(\lambda_{xi}, \lambda_{yi}))}{\sum_{\substack{i=1 \\ \neq x, y}}^r (V(\lambda_{xi}) + V(\lambda_{yi}) - 2\text{COV}(\lambda_{xi}, \lambda_{yi}))}. \quad (13)$$

This is a very general formula. Let us now take into account the characteristics of the chosen data model. We know that the covariance terms are null, and that the variances of the initial dissimilarities δ_{ij} are equal. To simplify, let us assume they are equal to 1. At the first step, we thus have $\mu^* = 1/2$, which corresponds to reduction (5), and the structural variance of terms λ_{ui} , newly created, is $1/2$. We now consider step p , and we suppose that the minimum variance reduction (12) and reduction (5) have coincided up to now. Each index λ_{ij} has thus a structural variance equal to $1/n_i n_j$, with the result that the minimum variance reduction is obtained for

$$\mu^* = \frac{\sum_{\substack{i=1 \\ \neq x, y}}^r \left(\frac{1}{n_y n_i} \right)}{\sum_{\substack{i=1 \\ \neq x, y}}^r \left(\frac{1}{n_x n_i} + \frac{1}{n_y n_i} \right)} = \frac{n_x}{n_x + n_y},$$

which corresponds to reduction (5). By induction, we deduce that the minimum variance reduction and reduction (5) coincide throughout the algorithm. Note that the same result may be obtained if we try to minimize the maximum on i of structural variances induced by the reduction, and that this result may be extended to any moment function of the form $(\sum x_i^q)^{1/q}$.

Formulae (12) and (13) are very general. Within the scope of our model, they enable us to demonstrate the identity between the minimum variance reduction and reduction (5). However, the main interest of these formulae is elsewhere, this result being widely anticipated. What is

interesting is that these formulae lead naturally to a general version of NJ which is able to take into account any variance-covariance matrix of the δ_{ij} estimates. Indeed, formula (13) enables μ^* to be calculated, thus determining the minimum variance reduction. Formula (12), coupled with an analogous formula relating to covariances, enables the same variance-covariance matrix to be updated at each step, so that the process may be repeated iteratively throughout the agglomerative procedure. This is the method we adopted within the scope of a biological sequence data model, and, as expected, significant improvements were obtained concerning the capacity to find the true tree (Gascuel 1996).

7. Simulation results

In order to evaluate the gain obtained by UNJ, within the scope of our model, we conducted simulations based on a schema inspired by Pruzansky *et al.* (1982) and Vach and Degens (1991). These were conducted with the 6 tree structures shown in Figure 6. The first three comprise 12 leaves, while the other 3 comprise 24. We find two extremes in these structures: chains (Chain 12, 24) and perfectly balanced structures (Eql. 12, 24), as well as intermediary structures (Int. 12, 24). In the chains, each (correct) agglomeration consists in adding a unique object to a group which already may comprise several objects. In this case, UNJ and NJ should differ significantly, since the unweighted reduction (5) tends to diverge from the weighted reduction of NJ which is systematically based on $\mu = 1/2$. In the case of balanced structures, each (correct) agglomeration generally agglomerates two clusters comprising the same number of objects, and NJ and UNJ should be extremely close. In the case of intermediary structures, results are expected to be intermediary.

For each of these structures, we generated 500 valued trees by randomly drawing the length of the edges according to a uniform distribution on $[0,1]$. The tree distance corresponding to each of these valued trees was normalized in order to obtain unit variance. To this normalized tree distance was added a gaussian noise with a null expectation and a standard deviation σ , with $\sigma = 0.1, 0.3$ and 0.6 . Finally, in order to avoid excessively non-metric data, we added a constant to each dissimilarity, so that $\min(\delta_{ij}) = 0.5$.

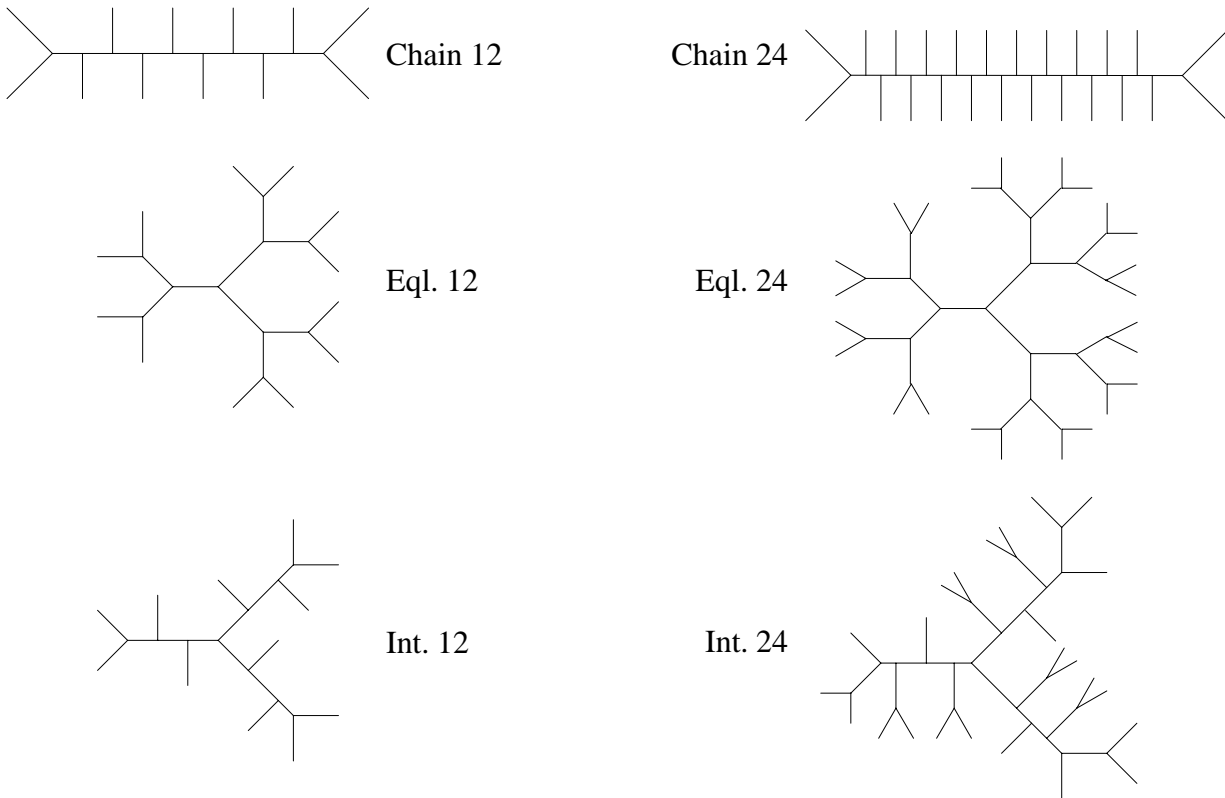


Figure 6: The six structures used in the simulations.

For each noisy tree distance obtained, we measured three criteria.

- The Robinson and Foulds (1981) distance between the inferred tree and the true tree. This distance corresponds to the number of bipartitions of the true tree, which are not found by the inferred tree, plus the number of bipartitions of the inferred tree which do not belong to the true tree. Since these quantities are always equal, we show in the table half of this distance, so that for n objects the criterion (RF) lies between 0 (when the trees are identical) and $n - 3$ (when none of the non-trivial bipartitions of the true tree have been found).

- The minimum evolution criterion (ME) which, as explained above, corresponds to the least-squares length estimate of the inferred tree. In the case of NJ (whose estimations are approximate) we used the algorithm described here (Equation 4, Section 3) in order to have the exact value of the criterion. So as to get an idea of the quality of the results obtained, we also applied this criterion to $\hat{\mathbf{T}}$, the structure of the true tree.

- The least-squares criterion (LS), in other words, the squared Euclidean distance between the data matrix Δ and the inferred tree matrix $\hat{\mathbf{D}}$. Due to the approximate formulae used by NJ, its performances are very poor insofar as this criterion is concerned. Therefore, in the table we show the results obtained by NJ, once the edge lengths have been correctly estimated by (4). Thus we obtain an idea of the quality of the tree structure inferred by NJ. As for the ME criterion, we also applied this criterion to $\hat{\mathbf{T}}$. Moreover, the least-squares criterion partially loses its meaning if we accept negative valuations. In order to relativise the results obtained following this criterion, we also measured, for each tree inferred and for $\hat{\mathbf{T}}$, the number of edges whose estimated length is negative. The greater their number, the worse the quality of the tree, even if with respect to least-squares, performance is good.

Results are given in Table 1.

- If we consider the capacity (RF) to recover the true tree structure, UNJ is systematically better than NJ. The error reduction is not very high: between about 10% for chains and a moderate noise ($\sigma = 0.1, 0.3$) and 1 to 2% for balanced trees, the reduction for intermediary trees being situated in between. The difference between the number of times where UNJ is better than NJ, and that where it is worse, is sometimes great, reaching 40% in the case of Chain 24 and $\sigma = 0.3$. In all cases, this difference is positive, which proves, if proof is needed, that UNJ should be given preference over NJ when assuming the data model chosen here. Generally, this difference is greater with 24 objects than with 12, which is easily explained because in this case reduction (5) has greater latitude in differing from $\mu = 1/2$, and approaching extreme values 0 and 1. In other words, with a large number of objects, UNJ differs more markedly from NJ, and takes advantage of the specificity of the model. As expected, the method performances decrease considerably when the noise σ increases. Consequently, the absolute gain obtained by UNJ tends to increase, whereas the relative gain tends to diminish.

- If we consider the minimum evolution criterion (ME), the tree inferred by NJ is generally better than the true structure $\hat{\mathbf{T}}$, whereas the tree inferred by UNJ is often better than that of NJ. Given that $\hat{\mathbf{T}}$ is itself likely to be close to the optimum, this proves that in terms of heuristic, UNJ and NJ are very efficient. On the other hand, this capacity to be "better" than $\hat{\mathbf{T}}$ is a handicap since, for example, if a heuristic was systematically better (in this sense) than $\hat{\mathbf{T}}$, it would never find

Tree	σ		RF	% RF<,>	% ME<,>	% LS<,>	#NEG	
Chain 12	0.1	NJ	0.51			36 6	26 15	0.3 0.4
		UNJ	0.47	8	8 5	15 1	13 2	0.2
	0.3	NJ	1.57			73 12	52 32	0.4 0.8
		UNJ	1.42	10	22 10	36 5	34 7	0.2
	0.6	NJ	3.44			95 4	72 28	0.6 1.6
		UNJ	3.34	3	23 16	47 9	45 10	0.4
Int. 12	0.1	NJ	0.33			24 4	20 7	0.5 0.6
		UNJ	0.31	4	4 3	8 1	8 0	0.4
	0.3	NJ	1.18			62 8	50 20	0.5 1.2
		UNJ	1.16	1	10 9	21 3	21 3	0.5
	0.6	NJ	2.62			91 6	73 24	0.7 1.5
		UNJ	2.55	2	17 11	34 6	32 8	0.6
Eq1. 12	0.1	NJ	0.20			14 3	14 3	0.5 0.6
		UNJ	0.20	1	0 0	1 0	1 0	0.5
	0.3	NJ	0.80			48 8	45 11	0.5 0.8
		UNJ	0.78	2	3 1	6 0	6 1	0.5
	0.6	NJ	2.00			74 9	71 12	0.7 1.5
		UNJ	1.97	2	9 5	21 3	19 4	0.6
Chain 24	0.1	NJ	2.18			81 11	40 51	0.5 0.8
		UNJ	1.92	12	35 16	53 12	51 12	0.0
	0.3	NJ	7.35			99 1	37 63	1.0 2.5
		UNJ	6.49	12	58 17	73 20	74 19	0.1
	0.6	NJ	13.3			100 0	47 23	1.2 4.7
		UNJ	12.7	5	50 17	73 23	74 21	0.3
Int. 24	0.1	NJ	1.00			50 14	44 17	0.2 0.6
		UNJ	0.92	7	13 6	20 4	18 5	0.1
	0.3	NJ	3.26			87 10	73 24	0.4 3.2
		UNJ	3.04	7	28 13	47 11	44 13	0.2
	0.6	NJ	7.73			99 1	75 25	0.8 3.1
		UNJ	7.44	4	34 20	69 16	61 19	0.4
Eq1. 24	0.1	NJ	0.67			38 9	37 8	0.2 0.6
		UNJ	0.66	1	2 1	4 1	3 1	0.2
	0.3	NJ	2.27			78 9	75 12	0.3 1.5
		UNJ	2.24	1	8 6	17 4	15 6	0.3
	0.6	NJ	5.26			93 5	87 12	0.6 2.6
		UNJ	5.23	1	15 13	39 11	36 12	0.4

Table 1: Results obtained for the six structures of Figure 6. RF is half of the Robinson and Foulds distance between the inferred tree and the true tree; the second item for UNJ indicates the percentage of error reduction obtained when comparing with NJ. %ME<,> refers to the minimum evolution criterion; in the case of NJ, the first item gives the percentage of times where the inferred tree seems better (according to ME) than the true structure \hat{T} , and the second item gives the percentage of times where it seems worse than \hat{T} ; in the case of UNJ the items have the same meaning, except that now we compare the inferred tree with that obtained by NJ. %LS adopts the same comparison scheme, but for the least-squares criterion. #NEG indicates the average number of negative edges; the second item for NJ corresponds to \hat{T} .

exactly $\hat{\mathbf{T}}$. We thus reach a well known problem in stochastic optimization. Given our results, the solution does not consist in improving the heuristic, but in refining the optimized criterion, so as to make a better selection among the trees close to $\hat{\mathbf{T}}$.

- Similar comments may be made about the least-squares criterion. We note, however, that the domination of NJ on $\hat{\mathbf{T}}$ is weaker than with the ME criterion; this is no doubt due to the fact that once adjusted, $\hat{\mathbf{T}}$ includes a rather large average number of negative edges. We have no explanation for this, and it merits attention in future developments. Otherwise, UNJ largely dominates NJ, while at the same time introducing less negative edges.

8. Discussion

We have presented a second version of NJ, which is unweighted and which we have called UNJ. We have shown that this version is coherent with a data model of the type $(\delta_{ij}) = (d_{ij} + \varepsilon_{ij})$, where (d_{ij}) is a tree distance, and where the ε_{ij} are independent and identically distributed noise variables. This new version derives from the original version of Saitou and Nei (1987) and Studier and Keppler (1988), and also from the results of Vach (1989) and Vach and Degens (1991) concerning the edge length estimation. The simulations show that an appreciable increase has been attained by using UNJ, when the data closely follow the chosen model.

It is not our intention to suggest a systematic preference for UNJ over NJ. Everything depends on the data. With biological sequence data (Swofford *et al.* 1996) we have noticed, for example, that NJ achieved better performances than UNJ, in terms of ability to recover the true tree structure. This is simply explained by the fact that these data are very far from the model chosen, particularly concerning the hypothesis of independence of the δ_{ij} estimates. However, other data exist which are closer to the model retained, for example, the ADN-ADN hybridization data, based on physical measures and in which the δ_{ij} estimates are basically independent (Felsenstein 1987). For these data and certainly for others, UNJ seems better adapted than NJ.

In fact, our intention is not to "defend" this new version of NJ, but instead to present a framework for the implementation of NJ versions, taking into account the specificity of the data. Within this framework, the model is expressed through the variance-covariance matrix of the δ_{ij} estimates.

As we have shown, selection criterion (3) retains its meaning whatever this matrix (§4.2), and this latter is used at each step to determine the minimum variance reduction (§5.2). By proceeding in this way throughout the algorithm, we get estimates which are as reliable as possible in choosing the pair to be agglomerated, and we increase the probability of finding the true tree. We have applied this schema here to a very simple classic model. Compared with NJ, the gains observed in the simulations may be qualified as modest, even though they are always positive (Section 7). However, we also applied this same framework to biological sequences (Gascuel 1996). As mentioned above, these data induce variances which vary considerably from one estimate to another, and also give important covariances. Compared with NJ, there is a considerable gain. For certain tree structures, we obtain up to a 50% error reduction, in terms of ability to recover the true tree structure. Generally speaking, we recommend that this approach be used and explored further, notably in the domain of edge length estimation, for which there is no general solution other than the matrix-based technique (Bulmer 1991) which is computationally very expensive.

Bibliography

- BANDEL, H.J., and A. DRESS. 1986. Reconstructing the shape of a tree from observed dissimilarity data. *Advances in Applied Mathematics*. **7**:309-343.
- BANDEL, H.J., and M. STEEL. 1995. Symmetric matrices representable by weighted trees over a cancellative abelian monoid. *SIAM J. on Discrete Math.* **8**:517-525.
- BARTHELEMY, J. P., and A. GUENOCHÉ. 1991. *Trees and proximity representations*. Wiley, Chichester.
- BRÖLSCH, J. 1983. Minimum-Quadrat-Schätzung von Evolutionsbäumen. In: I. Dahlberg, M. Schader (eds), *Automatisierung in der Klassifikation. Studien zur Klassifikation* **13**: 177-187.
- BULMER, M. 1991. Use of the Method of Generalized Least-squares in Reconstructing Phylogenies from Sequence Data. *Mol. Biol. Evol.* **8**:868-883.

- BUNEMAN, P. 1971. The recovery of trees from measures of dissimilarity. In: F.R. Hodson, D.G. Kendall and P. Tautu (eds), *Mathematics in Archeological and Historical Sciences*. Edinburgh University Press, Edinburgh, 387-395.
- CHARLESTON, M.A., M.D. HENDY and D. PENNY 1993. Neighbor-Joining uses the optimal weight for net divergence. *Mol. Phyl. Evol.* **2**:6-12.
- CUNNINGHAM, J.P. 1978. Free trees as representations of psychological distances. *Journal of Mathematical Psychology* **17**: 165-188.
- DE SOETE, G. 1983. A least-squares algorithm for fitting additive trees to proximity data. *Psychometrika* **48**: 621-626.
- FELSENSTEIN, J. 1987. Estimation of Hominoid Phylogeny from a DNA Hybridization Data Set. *J. Mol. Evol.* **26**:123-131.
- GASCUEL, O. 1994. A note on Sattath and Tversky's, Saïttou and Nei's and Studier and Keppeler's algorithms for inferring phylogenies from evolutionary distances. *Mol. Biol. Evol.* **11**:961-963.
- GASCUEL, O. and D. LEVY. 1996. A reduction algorithm for approximating a (non-metric) dissimilarity by a tree distance. *J. of Classification* **13**: 129-155.
- GASCUEL, O. 1996. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Cahier du GERAD G-96-59. To appear in *Mol. Biol. Evol.* (1997).
- KIDD, K.K. and L.A. SGARAMELLA-ZONTA. 1971. Phylogenetic analysis: concepts and methods. *Am. J. Human Genet.* **23**:235-252.
- KUHNER, M.K., and J. FELSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11**:459-468.
- KUMAR, S. 1996. A stepwise algorithm for Finding Minimum Evolution Trees. *Mol. Biol. Evol.* **13**:584-593.
- MIRKIN, B. 1996. *Mathematical Classification and Clustering*. Kluwer Academic Publishers, London.

- NEI, M. 1991. Relative efficiencies of different tree-making methods for molecular data. In: M. MIYAMOTO and J. L. CRACRAFT (eds) *Phylogenetic Analysis of DNA Sequences*. Oxford University press, Oxford, 90-128.
- PRUZANSKY, S., A. TVERSKY and J.D. CAROLL. 1982. Spatial versus tree representation of proximity data. *Psychometrika* **47**: 3-19.
- ROBINSON, D.F., and L.R. FOULDS. 1981. Comparison of Phylogenetic Trees. *Math. Biosci.* **53**:131-147.
- ROUX, M. 1988. Techniques of approximation for building two tree structures. In: C. Hayashi, E. Diday, M. Jambu, and N. Ohsumi (eds), *Recent Developments in Clustering and Data Analysis*. Academic Press, New York, 151-170.
- RZHETSKY, A., and M. NEI. 1993. Theoretical Foundation of the Minimum-Evolution Method of Phylogenetic Inference. *Mol. Biol. Evol.* **10**:1073-1095.
- SAITOU N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstruction of phylogenetic trees. *Mol. Biol. Evol.* **4**:406-425.
- SATTATH S., and A. TVERSKY. 1977. Additive similarity trees. *Psychometrika* **42**:319-345.
- SEARL, S.R. 1971. *Linear Models*. Wiley, New York..
- SNEATH, P.H.A., and R.R. SOKAL 1973. *Numerical Taxonomy*. Freeman, San Francisco.
- STUDIER, J.A., and KEPLER, K.J. 1988. A note on the neighbor-joining method of Saitou and Nei. *Mol. Biol. Evol.* **5**: 729-731.
- SWOFFORD, D.L., G.L. OLSEN, P.J. WADDELL, and D.M. HILLIS. 1996. Phylogenetic inference. In: D.M. Hillis, C. Moritz and B.K. Mable, eds, *Molecular Sytematics* (second edition), Chapitre 11. Sinauer, Sunderland (MA).
- VACH, W. 1989. Least-squares approximation of additive trees. In: O. Opitz (ed) *Conceptual and Numerical Analysis of Data*. Springer, Heidelberg, 230-238.
- VACH, W., and P.O. DEGENS. 1991. Least-squares Approximation of Additive Trees to Dissimilarities - Characterizations and Algorithms. *Computational Statistics Quarterly* **3**: 203-218.