The Anâtaxis phylogenetic method. I. Optimal trichotomies under fuzziness constraints : homoplasy and heterogeneity of evolutionary rate over phyletic lineages

Autor(en): **Bittar, Gabriel / Carter, Leigh**

Objekttyp: **Article**

Zeitschrift: **Archives des sciences et compte rendu des séances de la Société**

Band (Jahr): **50 (1997)**

Heft 2: **Archives des Sciences**

PDF erstellt am: **01.06.2024**

Persistenter Link: https://doi.org/10.5169/seals-740278

Archs Sci. Genève	Vol. 50	Fasc. 2	pp. 153-168	Septembre 1997

Communication présentée à la séance du 14 mai 1997

THE ANÂTAXIS PHYLOGENETIC METHOD. I. OPTIMAL TRICHOTOMIES UNDER FUZZINESS CONSTRAINTS, HOMOPLASY AND HETEROGENEITY OF EVOLUTIONARY RATE OVER PHYLETIC LINEAGES

BY

Gabriel BITTAR* & Leigh CARTER**

ABSTRACT

The Anâtaxis phylogenetic method. I. Optimal trichotomies under fuzziness constraints, homoplasy and heterogeneity of evolutionary rate over phyletic lineages.- A new phylogenetic method, named Anâtaxis, is proposed. It is a dissimilarity-matrix and outgroup-based, triadic trees-compatibility method that represents a new practical approach for phylogenetic inference. The first three steps of this method, as well as the fifth and last step, are presented here, the technically difficult fourth step being presented in part two. The first step in the method is the calculation of a dissimilarity matrix, which may include standard error and polymorphism-uncertainty parameters. The homologous sequences data must contain, as a starting point, an indisputable outgroup that is not too distant from the ingroup taxa which are to be analysed. The second step is the calculation of a dissimilarity matrix normalised in the sense that it is corrected for the heterogeneity of lineage-specific rates of evolution, on the basis of the information given by comparison of the ingroup taxa with the outgroup taxon(s). The third step consists in proposing for each ingroup triad a trichotomy that minimalises, in accordance with the outgroup-based information, *ad hoc* hypotheses of lineage-specific rate heterogeneity, and homoplasy. During this step, the fuzziness of the dissimilarity data due to its noisiness is taken into account in the process of determining the adequate trifurcating topologies. Once the dissimilarity matrix has been during the fourth step entirely and adequately analyzed for global tree reconstruction, the fifth and last step consists in inferring the best global tree, in different manners according to the way that noise in the data has been taken into account.

INTRODUCTION TO ANÂTAXIS

In the sixties, while the basic genetic code for proteins was being resolved, some researchers foresaw the enormous phylogenetic potential of analysing homologous molecular sequences (e.g. ZUCKERKANDL & PAULING, 1965; JUKES & CANTOR, 1969). Since then, much progress has been made in understanding the mechanisms of molecular

* Unité d'Investigation Clinique, Hôpitaux Universitaires de Genève, bâtiment La Seymaz, site psychiatrique de Bel-Air, CH - 1226 Thônex, Geneva, Switzerland

** Ionica plc, Cowley Rd, UK - CB4 4AS, Cambridge, England

Address for correspondence : Dr G. Bittar, Bioplan-Systeman Foundation, ch. du Pont-Noir 9A, CH-1226 Thônex, Geneva, Switzerland - fax: 41 22 349 2236; email: bittar@sc2a.unige.ch

evolution (e.g. BRITTEN, 1977; ROSE & DOOLITTLE, 1983; THOMAS & BECKENBACH, 1989; WOLFE *et al.*, 1992; BERNARDI *et al.*, 1993; OSAWA & JUKES, 1995). Moreover, the increasing availability of molecular sequences and of computer processing power has allowed researchers in phylogenetics to use larger and larger sets of molecular data. To help them in their task, more and more valid and efficient numerical procedures and computer tools for properly analysing the data obtained from molecular chains have been developed, and still are (e.g. FITCH & MARGOLIASH, 1967; HOLMQUIST *et al.*, 1972; HENDY & PENNY, 1982; SAITOU & NEI, 1987; DOOLITTLE, 1990; GOJOBORI *et al.*, 1990; SACCONE *et al.*, 1990; BLAISDELL, 1991; NEI, 1991; FELSENSTEIN, 1993; OLSEN *et al.*, 1994; HUELSENBECK & RANNALA, 1997). Indeed, the applications of molecular phylogenetics are vast and expanding (PALMER, 1992; CAVALLI-SFORZA, 1996).

The tools used in phylogenetics, which is a basal component of evolutionary science, are based on different methods, but they all present some problems, that we have briefly described in a preceding paper (BITTAR, 1996). Faced with these problems, we thought it would be useful to conceive of a new approach, and to implement it in a computer program to test it on different sets of real data. Numerical phenetic taxonomy methods, based on distances-matrix, being by far the quickest since they avoid direct numerical reference to the matrix of character states, we thought it was worth developing on them, but in a novel way such that they would no longer be affected by the systematic biases affecting them, and thus allowing the produced tree to be phylogenetically more accurate (BITTAR & CARTER, 1994).

We have accordingly developed, as an alternative to quick but often biased distances-matrix phenetical approaches, a new kind of trees-compatibility dissimilarity-matrix method, that groups taxa while taking into account both the possibility of homoplasy, on one hand, and the possibility of a wide spectrum of rates of evolution within the different branches, on the other hand (for example, see GOODMAN, 1981; GOJOBORI & YOKOYAMA, 1987; WOLFE *et al.*, 1987; CACCONE & POWELL, 1990; HONEYCUTT *et al.*, 1995; for a different opinion on the subject of evolutionary rates, see for example EASTEAL, 1990). In this way, while avoiding the pitfalls and biases of a phenetical distances-matrix approach, which are due to not accounting for homoplasy and the heterogeneity of evolutionary rates in the different lineages, our Anâtaxis program allows the user to perform a quick analysis of even a huge set of data by a novel, more phylogenetic, method of analysing the information inherent in a matrix of pairwise dissimilarities (it should be noted that the **Anâtaxis** trees indicate polychotomies when information is not judged sufficient for defining nodes, in order to avoid the unnecessary display of uncertain or weekly supported phyletic relationships).

Basically, we propose with Anâtaxis a new method integrating in a dissimilarity-matrix numerical approach the phylogenetic concept of outgroup, which, as we shall see, if rigorously applied, eliminates the biases associated with classical distances-matrix methods based on clustering. Giving a more phylogenetic quality to numerical phenetic techniques is a good approach, the validity of which has been exemplified 25 years ago

(FARRIS, 1972), but the importance of this hybrid approach has not been sufficiently recognized.

It must be emphasised that Anâtaxis is based on a direct analysis of the dissimilarities, which are *not* metric distances because of the possibility of *homoplasy*; quite evidently, neither are these dissimilarities additive distances, and Anâtaxis is not intended at optimalising any kind of scalar measure, such as minimalising the length of the global tree. Anâtaxis only tries to define a tree which is compatible with the dissimilarity matrix, and which minimalises *ad hoc* evolutionary hypotheses. Accordingly Anâtaxis is not *stricto sensu* a distance-matrix method (even though it may be considered as belonging to the very general category of phenetical methods, as opposed to characters analysis methods). And neither is Anâtaxis *stricto sensu* a quadruplet method, as will be made evident in this paper. In fact, from the point of view of a systematics of phylogenetic methods, Anâtaxis constitutes a new category by itself. Whatever, when faced with a substantial set of data, with program Anâtaxis it is no longer necessary to analyse the possible evolutionary story of each character, as cladistic parsimony methods do, hence avoiding the inevitable drastic slowdown and the peculiar pitfalls associated with these methods. Nor is it necessary to stand by in frustration waiting for a future with more rapid computers so as to be able to use the promising but very slow probabilistic methods, such as Maximum Likelihood Estimation methods.

The Anâtaxis computer program in which our method has been implemented, is the most general purpose program in the **Vivâras** phylogenetic package that we are developing (BITTAR, 1995); it can be applied to any kind of data obtained from evolving objects, with the sole conditions that one of these can be defined as outgroup to the others and that one can construct a symmetrical matrix of object-to-object dissimilarities. The Anâtaxis method is quick and efficient, and the validity and robustness of the trees produced by the program have already been demonstrated on different sets of data (NADOT *et al.*, 1995; SOUZA-CHIES *et al.*, 1996; PAWLOWSKI *et al.*, 1996; BITTAR *et al.*, 1996; VEUTHEY & BITTAR, in preparation).

Now let us see more precisely how the Anâtaxis method operates.

ANÂTAXIS FIRST AND LAST STEPS

The general procedure for using Anâtaxis is the following. The first step is conceptually simple, it consists in processing properly (!) aligned (e.g. with help from program Clustal W, v.1.7, THOMPSON *et al.*, 1994) homologous sequences, with Anâtaxis if the data is simple, otherwise with a more sophisticated package (such as **TakâmolE**, from the Vivâras package, under development), so as to produce

- either a single pairwise dissimilarity matrix Δ_{ij} (corrected for sequencing errors : $\hat{\Delta}_{ij}$) if the user does not want to integrate the notions of standard error and polymorphism/uncertainty (this implies using Anâtaxis in a way in which numerical clustering of the set of dissimilarity values plays an important role); i and j being the two terminal

taxa or OTUs (Operational Taxonomic Units) for which a homologous sequence of characters is being compared; the members of the main diagonal are all equal to zero ($\Delta_{ii} = 0 \quad \forall\ i$) and the matrix is symmetrical ($\Delta_{ij} = \Delta_{ji} \quad \forall\ i,j$);

- or, if one wants to integrate in the input data the notions of standard error (due to the paucity of dissimilar sites) and polymorphism-uncertainty (parameters υ_{ij}), which is advisable, to process the dissimilarity data in either of two different ways, according to the manner one wishes to tackle with Anâtaxis the noise intervals on the original $\hat{\Delta}_{ij}$.

In the first way, the input for Anâtaxis consists of three successive matrices, one containing the $\hat{\Delta}_{ij}(1-\upsilon_{ij})$ values, the second the $\hat{\Delta}_{ij}$ values, and the third the $\hat{\Delta}_{ij}(1+\upsilon_{ij})$ values (in the case where there is no missing data) : these are the three input matrices intended for the Anâtaxis "fuzziness" treatment done in the second step. Generally, and particularly if the unknown data is distributed rather evenly among the different sequences, in such a way that no missing-data zone can clearly be defined, and if the number of taxa is high, this is the method of choice in terms of both rapidity and reliability. After dealing with these three matrices, Anâtaxis outputs in the last and fifth step a single tree, in the usual nested-parenthesis symbolic notation form.

In the second way, the input for Anâtaxis consists of an appropriate number of dissimilarity matrices that have been randomly modified from the original matrix, each $\hat{\Delta}_{ij}$ varying within its proper $\hat{\Delta}_{ij}(1\pm\upsilon_{ij})$ noise limits (generally 30 to 50 randomising runs on the whole original matrix shall be appropriate) : in this case Anâtaxis outputs in the last step as many trees (in nested-parenthesis symbolic notation form) as there were input dissimilarity matrices, proceeding for each of these tree reconstructions in the same manner as for a single noise-free input matrix (accordingly, numerical clustering of the set of dissimilarity values also plays here an important role). All these Anâtaxis trees are then further subjected to a consensus rule : in practice, they can be input into the program Consense from the Phylip 3.57 package (FELSENSTEIN, 1993), which calculates, according to M_I majority rules, a consensus tree.

Finally, this consensus tree, or the sole tree that has been directly produced with Anâtaxis, can in an ultimate step be drawn with Phylip's programs Drawgram or Drawtree, or with MacClade 3.05 (MADDISON & MADDISON, 1992), or PAUP 3.1.1 (SWOFFORD, 1993), or M. Gouy's NJPlot, or any adequate other program.

ANÂTAXIS SECOND STEP

Out-group and normalisation

An important characteristic of Anâtaxis is that it produces, from the original dissimilarity matrix (all $\upsilon_{ij} = 0$) produced by TakâmolE (with or without taking into account the two upper and lower error-boundaries matrices), and, if desired, from the noise-derived semi-random matrices, a tree that is *rooted.*

The reason for this rooting is that, as a first and foremost condition, the data for Anâtaxis must contain at least one sequence of a taxon o that can be considered as an indisputable *outgroup* to all other taxa to be analysed, which together constitute the *in-group* I. The out-group is the basis for the definition of an OUT-IN vector $\hat{\Delta}_{oi}$, constituting e.g. the first line of the whole dissimilarity semi-matrix.

In the following example of a dissimilarity matrix, clade 5 constitutes the out-group o, taxa 1 to 4 constitute the in-group I.

sequence	1	2	3	4	
5	$\hat{\Delta}_{51}$	$\hat{\Delta}_{52}$	$\hat{\Delta}_{53}$	$\hat{\Delta}_{54}$	$\hat{\Delta}_{oi}$ OUT-IN vector
4	$\hat{\Delta}_{41}$	$\hat{\Delta}_{42}$	$\hat{\Delta}_{43}$		
3	$\hat{\Delta}_{31}$	$\hat{\Delta}_{32}$			$\hat{\Delta}_{ij}$ IN-IN sub-matrix
2	$\hat{\Delta}_{21}$				

Better, there can be many different out-groups, producing as many different dissimilarity matrices from which one produces an arithmetic (or algebraic) mean OUT-IN vector $\hat{\Delta}_{oi}$ (where the index $_i$ designates a member of the in-group I). Even better, each indisputable out-group can be a whole clade with known internal structure (again allowing the calculation of a weighted $\bar{\bar{\Delta}}_{oi}$ OUT-IN vector of dissimilarity); e.g., if the out-group o is constituted of three taxa, o_1, o_2 and o_3, phyletically forming a resolved trichotomy $(o_3, (o_2, o_1))$, we have $\bar{\bar{\Delta}}_{io} = (2\hat{\Delta}_{io3} + \hat{\Delta}_{io2} + \hat{\Delta}_{io1})/4$.

In the next step, all the $\hat{\Delta}_{oi}$ are made identical to a normalising value, so that the effect of the heterogeneous contribution to the IN-IN sub-matrix (constituted by the whole matrix minus the OUT-IN vector) of the unequal rates of evolution among the different lineages can, to a good approximation, be eliminated – this normalisation step is not mandatory, Anâtaxis can also work in a different, but much more complex, way (BITTAR & CARTER, 1994) –. Empirically, using the median of the $\hat{\Delta}_{oi}$ for normalising gives good results. Hence, all the difference values between this median and the Δ_{oi} are calculated :

$$diff_{oi} = med(\hat{\Delta}_{oi}) - \hat{\Delta}_{oi}$$

Then a new, normalised, IN-IN* matrix of dissimilarity is calculated, in which (i and j $\in$ I)

$$\hat{\Delta}_{ij}^{*} = \hat{\Delta}_{ij} + diff_{oi} + diff_{oj}.$$

This kind of normalisation procedure is derived from the proposition of KLOTZ *et al.* (1979), but here we use, on the basis of tests done on many kinds of data, a median

rather than a mean function. Moreover, we have empirically found that this procedure may be applied directly to primary dissimilarity values, rather than to secondary distance values. Though the importance of such a normalisation procedure has been recognised long time ago (FARRIS, 1972), it has rarely been applied in phenetical studies, despite its powerful usefulness in allowing the avoidance of evolutionary rate heterogeneity between lineages. Other methods for correcting for unequal rates of substitution in the absence of known root or basal outgroup have been proposed (LI, 1981), they basically correspond to mid-point rooting of which we have come to the conclusion that it is a dangerous method, often giving phyletically absurd results. Consequently we are now convinced that to do proper phylogenetic inference of a given set, it is necessary to have the data from an outgroup, described as such from another procedure than the phylogenetic one that is used. Otherwise, the phylogenetic reasoning is easily circular and the phyletic inference *en définitive* arbitrary.

This fundamental point being clarified, the importance of the first (initial) out-group cannot be over-emphasised. It must be phylogenetically a clear out-group, but it must not be too distant from the in-group we wish to analyse, otherwise all $\hat{\Delta}_{oi}$ would tend to be equal to the maximum mean dissimilarity value of diverging sequences, i.e. $(U_{0oi}-1)/U_{0oi}$, where U_{0oi} is the mean efficient size of the universe of possible states for the $2n$ characters composing diverging sequences i and o (e.g. $U_0 = 4$ for equiprobable bases and in the absence of gaps; n is the length of the homologous aligned sequences). If the out-group was phyletically that distant from the in-group, it would be no more useful than a randomly composed sequence o : it could help to artificially root the tree, but without the possibility of any normalisation process taking into account the heterogeneity of substitution rates among the different lineages (thus implicitly assuming that all lineages have evolved at the same rate).

The dissimilarity values as "fuzzy" objects

There is a standard error-uncertainty interval (due to paucity of dissimilar sites - γ_{ij} parameter - and to polymorphism - π_{ij} parameter -) for each $\hat{\Delta}_{ij}$ dissimilarity (corrected for sequencing errors), defined as

$$\hat{\Delta}_{ij} \in [\hat{\Delta}_{ij}(1-\upsilon_{ij}) \; ; \; \hat{\Delta}_{ij}(1+\upsilon_{ij})] ,$$

with $\upsilon_{ij} \approx \pi_{ij} + \gamma_{ij}$ if π_{ij} and γ_{ij} are small, and $\upsilon_{ij} \in [0 \; ; \; 1]$.

For any normalised dissimilarity, we have

$$\hat{\Delta}_{ij}^{*} \in [\hat{\Delta}_{ij}^{*}(1-\upsilon_{ij}) \; ; \; \hat{\Delta}_{ij}^{*}(1+\upsilon_{ij})].$$

If there are missing data, $\hat{D}_{ij}$ being the sum of zone-component dissimilarities known from a direct comparison of sequences i and j, d_{ij} being the sum of inferred zone-component dissimilarities ($\hat{\Delta}_{ij} = \hat{D}_{ij} + d_{ij}$),
and $\upsilon_{ij}^{\#} = \upsilon_{ij}.\hat{D}_{ij}/\hat{\Delta}_{ij} + 1.d_{ij}/\hat{\Delta}_{ij}$,

$$\hat{\Delta}_{ij} \in [\hat{\Delta}_{ij} (1 - \upsilon_{ij}^{\#}) ; \hat{\Delta}_{ij} (1 + \upsilon_{ij}^{\#})],$$

For any normalised dissimilarity, in presence of missing data, we thus have

$$\hat{\Delta}_{ij}^{*} \in [\hat{\Delta}_{ij}^{*} (1 - \upsilon_{ij}^{\#}) ; \hat{\Delta}_{ij}^{*} (1 + \upsilon_{ij}^{\#})].$$

In practice, three matrices are successively input in the Anâtaxis program if the "fuzziness" treatment is chosen :

the matrix containing the lower-boundary normalised dissimilarity values,
$[\hat{\Delta}_{ij}^{*} (1 - \upsilon_{ij}^{\#}]$;

the matrix containing the middle-point (original) normalised dissimilarity values,
$[\hat{\Delta}_{ij}^{*}]$;

the matrix containing the upper-boundary normalised dissimilarity values,
$[\hat{\Delta}_{ij}^{*} (1 + \upsilon_{ij}^{\#})]$.

Then the procedure simply consists in treating the $\hat{\Delta}_{ij}^{*}$ intervals as "fuzzy" numbers which are considered as approximately identical when they overlap.

A much better, really "fuzzy"-style approach, would be to allocate to each interval a probability function $f(\Delta_{ij}^{*})$, of which the integral between the lower-boundary and the upper-boundary values would be equal to 1. This function could, for example, be bell-shaped. Then the intersection between two "fuzzy" objects (i.e. two dissimilarity intervals) would no longer be simply a yes-no problem, but could be characterised by a probability distribution. This seducing but rather complex development is for the future.

Partitioning the in-group dissimilarity sub-matrix

The user is also offered another (or supplementary) option, consisting in clustering the $\hat{\Delta}_{ij}^{*}$ values composing the normalised IN-IN* matrix. The philosophy behind this operation is that some of these values, even if not precisely equal to one another, are sufficiently near to one another that they can be considered as *approximately* equal and thence forming a cluster. Normally, this clustering option is avoidable if the user has opted for the "fuzziness" method.

Generally the eye is a good instrument for performing such a clustering when the values are ordered, but if the data set is large this task may become fastidiously long; so Anâtaxis offers two special clustering tools, which can be helpful for this critical operation.

There is a common procedure to both clustering methods, which consists in firstly defining an in-group vector ${}^{\#}\vec{\Delta}_i{}^*$ (or ${}^{\#}\vec{\Delta}_i$, if one has not proceeded with normalisation) in which the components are perfectly ordered.

Let us define the in-group I as containing I members. All I(I-1)/2 components of the dissimilarity in-group semi-matrix are perfectly ordered. The new vector, ${}^{\#}\vec{\Delta}_i{}^*$ (or ${}^{\#}\vec{\Delta}_i$), which may contain less than I(I-1)/2 components since the $\Delta_{ij}{}^*$ or Δ_{ij} ($i,j \in I$) having the same value are collapsed together, is thus formed with all these distinct values (the analogue series with the * symbol is not written) :

$${}^{\#}\Delta_i{}^{(1)} > {}^{\#}\Delta_i{}^{(2)} > {}^{\#}\Delta_i{}^{(3)} > \ldots > {}^{\#}\Delta_i{}^{(m)}.$$

The vector ${}^{\#}\vec{\Delta}_i$ may only contain positive values, but ${}^{\#}\vec{\Delta}_i{}^*$ may also contain negative values.

Then ${}^{\#}\vec{\Delta}_i{}^*$, or ${}^{\#}\vec{\Delta}_i$, must be partitioned in a plausible way. This might be particularly difficult if there are a great number of components within this perfectly ordered vector. As we have said, to help the user in this crucial operation, two specific automated clustering methods are proposed to him. It must be emphasised that these automated clustering methods are simply helpful tools designed to assist the user in his partitioning task, because basically it is his eyes and brain which are the main tools for this work. In a similar way as for sequences alignment, partitioning is a rather complex procedure, difficult to describe in an algorithm, but that an experienced user may do quite well with his biological analogous parallel processing powers... yet preferably with the help of adequate computer programs.

The first clustering algorithm, named "relative differences / distant islands", works in the following way.

Within the vector ${}^{\#}\vec{\Delta}_i$ (or ${}^{\#}\vec{\Delta}_i{}^*$), one looks for a series of successive ${}^{\#}\Delta_i$ components (for the sake of simplicity, we make abstraction of the index i indicating that we are working with vector ${}^{\#}\vec{\Delta}_i$ or ${}^{\#}\vec{\Delta}_i{}^*$, and note the ordering identification numbers of the components of this vector as indexes), such that,

with ${}^{\#}\Delta_M$ being the biggest member of this series, ${}^{\#}\Delta_m$ the smallest, and considering that the smallest probabilistic quantum of dissimilarity is $1/[n(U_0-1)/U_0]$,

$$U_0/[2n(U_0-1)] + ({}^{\#}\Delta_M - {}^{\#}\Delta_m)/({}^{\#}\Delta_M + {}^{\#}\Delta_m) < ({}^{\#}\Delta_{M+1} - {}^{\#}\Delta_M)/({}^{\#}\Delta_{M+1} + {}^{\#}\Delta_M)$$

and

$$U_0/[2n(U_0-1)] + ({}^{\#}\Delta_M - {}^{\#}\Delta_m)/({}^{\#}\Delta_M + {}^{\#}\Delta_m) < ({}^{\#}\Delta_m - {}^{\#}\Delta_{m-1})/({}^{\#}\Delta_m + {}^{\#}\Delta_{m-1}),$$

with $^{\#}\Delta_{m-1}$ and $^{\#}\Delta_{m+1}$ respectively being the $^{\#}\Delta_i$ immediately preceding $^{\#}\Delta_m$, and the $^{\#}\Delta_i$ immediately succeeding $^{\#}\Delta_M$. If these two conditions are satisfied, the series of $^{\#}\Delta_i$ values from $^{\#}\Delta_M$ (inclusive) to $^{\#}\Delta_m$ (inclusive) forms a cluster.

Different partitioning solutions may be obtained that satisfy these conditions (particularly, with whole numbers, the one-value-per-cluster solution) : Anâtaxis proposes the one minimising the number of clusters.

The three following observations must be made on this procedure :

- the search stops when $^{\#}\Delta_m$ is the smallest $^{\#}\Delta_i$, $^{\#}\Delta_{m-1}$ then being an *ad hoc* value (defined by the user) smaller than $^{\#}\Delta_m$, e.g. 0 if there are no negative values (as this might be the case with normalised values);

- when $^{\#}\Delta_M$ is the biggest component of the vector $^{\#}\vec{\Delta}_i$, then $^{\#}\Delta_{M+1}$ is an *ad hoc* value (defined by the user) bigger than $^{\#}\Delta_M$; this upper boundary could be defined as $n.(U_0-1)/U_0$ (i.e. $3n/4$ in the case of a nucleotide sequence of length n, or $4n/5$ if the state of gaps is considered as a "5th" base); or, alternatively, as $^{\#}\Delta_{m+1} = n$ (but the upper limit $n.(U_0-1)/U_0$ is preferred not only because probabilistically more logical, but also because compatible with the general philosophy of clustering adopted here : distinct groups are defined only where quantitative differences between them are clear-cut);

- with this splitting-clustering method, and the following one also, artificial borders between groups, that might be due to the smallness of the sample (I small), are as much as possible avoided : there must be a bigger (relative) difference between the smallest $^{\#}\Delta_i$ of a cluster and the biggest $^{\#}\Delta_i$ of the following cluster where I is small rather than where it is big.

The second clustering method of the components of the vector $^{\#}\vec{\Delta}_i{}^*$ (or $^{\#}\vec{\Delta}_i$), that is proposed to the user as "absolute differences / distant islands", also operates through splitting, but in a quite different manner.

If $^{\#}\Delta_1 - ^{\#}\Delta_2 > ^{\#}\Delta_2 - ^{\#}\Delta_3$, then $^{\#}\Delta_1 >> ^{\#}\Delta_2$ momentarily, otherwise $^{\#}\Delta_1 \approx ^{\#}\Delta_2$ definitely (the partitioning philosophy is still the same : there must be a clear-cut inequality for the partition to be definitely accepted).

In the first case ($^{\#}\Delta_1 >> ^{\#}\Delta_2$), during the next step of this partition algorithm the following question is asked :

is $^{\#}\Delta_2 - ^{\#}\Delta_3 > ^{\#}\Delta_3 - ^{\#}\Delta_4$?

If the answer is no, then the first question is asked again, but now, since $^{\#}\Delta_2 \approx ^{\#}\Delta_3$, it is reformulated as :

is $^{\#}\Delta_1 - ^{\#}\Delta_2 > ^{\#}\Delta_2 - ^{\#}\Delta_4$?

And so on. At the end of the process any two Δ_i values of the vector $^{\#}\vec{\Delta}_i{}^*$ (or $^{\#}\vec{\Delta}_i$) which are approximately equal ('≈') belong to the same cluster. Any pair of two successive values of this perfectly ordered vector which are largely unequal ('>>') defines a boundary between two clusters.

These are the two clustering algorithms proposed by Anâtaxis to the user, who may also decide to adopt a combination of these two methods (either Boolean AND, or

Boolean OR). They have been named "distant islands" methods because they help to detect islands of neighbour dissimilarity values which are numerically strongly distinct from other values. It must be made clear that these clustering methods are only tools, and quite often it is necessary for the user to fine-tune the result he has been able to obtain with these. For example, once a clustering is done, the user may decide to use the "distant archipelagos" option, which consists in "merging" two neighbour clusters containing dissimilarities with common taxa, thus suppressing the partitioning frontier between them. Or, inversely, he may decide to "split" a cluster considered as too heterogeneous. Whatever, clearly, these "distant islands" methods will not cluster the components of a perfectly ordered vector if these form a smooth continuum of values, which is naturally more and more the case the larger and the more diverse the sampling of taxa. In this latter case, the Anâtaxis "fuzziness" method is preferred to the clustering method.

Ideally, one could imagine a combination of clustering and "fuzziness" methods. For example, as already suggested, there could be for each $\hat{\Delta}_{ij}$ dissimilarity a (bell-shaped) probability distribution function, constrained within the uncertainty interval. Then, rather than being defined as a member or not of a given cluster, in a yes-no way, any $\hat{\Delta}_{ij}$ would belong to a given cluster according to the intersection of its probability distribution curve with the numerical interval defined by this cluster. E.g., $\hat{\Delta}_{ij}$ would belong to cluster nb 2 with a probability of 65%, and to neighbour cluster nb 1 with a probability of 35%. Again, this is a seducing possibility, but the difficulty of implementing it reserves it for a future development.

ANÂTAXIS THIRD STEP

If the "fuzziness" method has not been adopted by the user, the partitioning of the vector $^{\#}\vec{\Delta}_1{}^*$ is a crucially necessary aspect of Anâtaxis, because the next and third step for the program is to propose for each triad of in-group taxa, for which all $\hat{\Delta}_{ij}$ values either have been assigned with uncertainty intervals or have been properly clustered, the best possible trichotomous tree (again, for the sake of simplicity, we do not write in the following the caret ^ over the delta Δ).

A prerequisite to this step is to define all possible tetradic cases after normalisation to the median of the OUT-IN vector (o designing the out-group, all $\Delta_{oi}{}^*$ values having been set equal to the median of the Δ_{oi} values, and all the Δ_{ij} of the IN-IN sub-matrix having been modified consequently to their normalised values $\Delta_{oi}{}^*$), a normalisation which implies that $\Delta_{oa}{}^* = \Delta_{ob}{}^* = \Delta_{oc}{}^*$, for any triplet of taxa a, b and c that are members of the in-group. Since one of the four taxa being compared (o) is predefined as out-group to the three others (a, b and c), the problem simplifies to defining the best or most likely trichotomy for the in-group of three taxa, according to a hierarchical set of rules.

For each sextuplet of dissimilarities (Δoa, Δob, Δoc, Δcb, Δca, Δba) a single tree is proposed, the tree corresponding to each sextuplet appearing in the dissimilarities-to-tree correspondence table found in this paper. The general principle of parsimony guides this correspondence table, in the sense that each proposed solution minimises *ad hoc* hypotheses, ordered according to the following hierarchy of evolutionary plausibility : hypotheses of heterogeneity in the rates of evolution among different branches are avoided if possible, and hypotheses of homoplasy are considered only in the last resort.

In all there are four tables, established according to the relationships between Δoa , Δob and Δoc : i.e. Δoc = Δob = Δoa ; Δoc > Δob = Δoa ; Δoc = Δob > Δoa ; and Δoc > Δob > Δoa .

However, as a result of the normalisation, we shall have for each tetradic case the double equality Δoc* = Δob* = Δoa*, hence the sole correspondence table presented in this paper is the one describing, for each triadic case satisfying the conditions Δoc = Δob = Δoa, the trichotomy tree adopted by Anâtaxis. In this table, in which the impossible sets of relationships are set in italic, one finds, at the left of each possible trichotomous evolutionary tree for a triplet of taxa a, b and c, the (in)equality relationships between the three (normalised) dissimilarities (Δcb, Δca, Δba) which imply and are implied by this tree, in a biunivocal (bijective) way. In a future development, the correspondence table algorithm could be expanded so as to take into account the less likely trichotomies for a given (Δoa, Δob, Δoc, Δcb, Δca, Δba) sextuplet or (Δcb*, Δca*, Δba*) triplet, in a probabilistic "fuzziness" way, but again this is not an easy thing to implement.

Since the dissimilarity values are compared according to "fuzziness" rules, the equality sign ("=") found in the correspondence table translates as "approximately (roughly) equal to", and the inequality signs ("<" and ">") translates as "clearly smaller than" and "clearly greater than", i.e. as "<<" and ">>". It cannot be overstated that this means that, if the uncertainty intervals for two compared dissimilarities intersect, or if these two values belong to the same cluster, these two dissimilarity values are considered as roughly equal. Also, note again that since reference is made to an out-group, *de facto* these are tetradic cases, the 4th taxon being o, and the proposed trees are thus rooted tetrachotomies.

The Δoc = Δob = Δoa double equality, with o being the out-group, is the first and foremost condition, that any proposed triadic tree solution for the three dissimilarities Δcb, Δca and Δba of taxa a, b and c, must satisfy. The tree respecting this condition without one having to make any hypothesis of temporal heterogeneity for the rates of evolution among different branches, or of homoplasy, is considered as the most likely. If it is necessary to hypothesise for a branch a relatively rapid (slow) rate of evolution, then it is necessary to hypothesise a relatively slow (rapid) rate in the immediately preceding or following branch; the most likely tree is then the one minimising, among its four (or three) branches, the heterogeneity of branch-specific evolutionary rates, still without homoplasy. Then, if it is necessary to hypothesise for homoplasy in any branch (relatively to o), the most parsimoniously likely tree is the one minimising the level of homoplasy.

When respecting this hierarchy of conditions, there are four possible kinds of trichotomous evolutionary trees. The simplest kind is an unresolved, symmetrical trichotomous tree, (a,b,c) , which is the most adequate in the case where all three dissimilarities are (roughly) identical : $\Delta cb = \Delta ca = \Delta ba$.

Then there is the still simple case with one relatively small dissimilarity, and two relatively large and (roughly) identical dissimilarities. E.g. $\Delta cb = \Delta ca > \Delta ba$, for which tree (c,(b,a)) is the most parsimonious and likely solution (it must be reminded that the word parsimonious is not necessarily used in the sense of cladistic parsimony - Hennig 1950, 1966 -, but in its more general sense of minimisation of a hierarchical set of hypotheses); there are three possible cases of this kind for any given triplet.

Then there is the more complex case with one relatively large dissimilarity, and two relatively small and (roughly) identical dissimilarities, e.g. $\Delta cb = \Delta ca < \Delta ba$. In this case, in the absence of homoplasy (in conformity to the adopted set of rules), there is only one solution compatible with the $\Delta oc = \Delta ob = \Delta oa$ conditions : it is the one grouping together b and a, making them evolve rapidly (symbol '+') since their last common ancestor, with the immediately preceding branch having evolved slowly (symbol '-'), so that, from the root to leaves a, b and c, the a and b lineages have roughly evolved at the same rate as lineage c. Thus, tree (c,(b:+,a:+):-) is the most likely solution (in accordance to the chosen hierarchically-ordered hypotheses), and it must be underlined here that it is the two terminal taxa with highest dissimilarity that together form a clade (again, there are three possible cases of this kind for any given triplet).

This apparently counter-intuitive result needs some comment. Clearly, in the absence of homoplasy and with $\Delta oc = \Delta ob = \Delta oa$, tree (c,(b:+,a:+):-) is the sole possible when $\Delta cb = \Delta ca < \Delta ba$. On the other hand, is an hypothesis of homoplasy really less plausible than a hypothesis of heterogeneity of substitution rates ? In other words, wouldn't either tree (b,(c,a)) with b and c homoplasic, or tree (a,(c,b)) with a and c homoplasic, be more likely answers, with the advantage of these two trees being possibly ultra-metric ?

The answer is not easy, one of the reasons being that the molecular clock is a source of much controversy, both theoretical and empirical (see e.g. GOODMAN, 1981; GOJOBORI *et al.*, 1982; DOVER, 1987; OHTA, 1987; WOLFE *et al.*, 1987; ZUCKERKANDL,1987; CACCONE & POWELL, 1990; EASTEAL, 1990). The choice that has been made here is based on personal experience with real data (NADOT *et al.*, 1995; SOUZA-CHIES *et al.*, 1996; BITTAR *et al.*, 1996; PAWLOWSKI *et al.*, 1996; and BITTAR, unpublished data). Using as a yard-stick the cladistic maximum parsimony method (SWOFFORD, 1993), which consists in minimising the total sum of pairwise dissimilarities between contiguous nodes in a tree, it appeared possible, after analysing more than 10'000 rooted quadruplets, to conclude that the (c,(b:+,a:+):-) scenario, which clearly is not ultra-metric, is five to six times more frequent, thus more probable, than the possibly ultra-metric scenarios with homoplasy. But in no way could the analysed quadruplets be considered as rigorously representative of the general molecular evolution conditions in biology; in fact, the ratio that has been found is only representative of the data that have been treated. So, in the

present state of affairs, the answer to the question of which quadruplet tree is most likely (rate heterogeneity versus homoplasy) is basically a question of opinion.

Clearly, it would be ideal to have an exhaustive dissimilarities-to-tree table, where all possible phyletic scenarios, for any (Δoa, Δob, Δoc, Δcb, Δca, Δba) sextuplet, would appear, with each possible solution having a probability value between 0 and 1 (presently, the most likely scenario is affected with probability 1, the other scenarios with probability 0); then Anâtaxis would be able, in a probabilistic "fuzziness" way, to propose a spectrum of solutions rather than a single global tree... but this is easier to say than to implement.

Whatever, the correspondence table approach is flexible enough for freely allowing alternative evolutionary scenarios if one does not adhere to those presented here.

This important commentary being done, there is finally (we are still analysing the Δoc = Δob = Δoa sub-table) the even more complex case where the three dissimilarities are all clearly different from one another, e.g. Δca > Δcb > Δba . In this case, it is no longer possible to avoid doing a homoplasy hypothesis. In terms of parsimony of the hierarchically-ordered evolutionary hypotheses, the most likely solution, and the one adopted by Anâtaxis, is the tree (c,(b,a:+)) , with a being partially homoplasic with o, and the terminal branch leading to it having evolved relatively rapidly. Another solution could be (a,(b,c:++)) , but, with c being then largely homoplasic with o, and the terminal branch leading to it being affected with a relatively highly rapid rate of divergence (symbol '++'), it is a less parsimonious solution than the preceding one, and thus it is considered as not likely (the two other possible solutions, i.e. the unresolved trichotomy, or a resolved trichotomy with b external to clade (a,c), are even less parsimonious). For any given triplet, there are six possible cases of this kind, all implying partial homoplasy in the same manner.

It is worth emphasizing that, in 6 triads out of 13 in the correspondence table, there is no possible solution without doing a hypothesis of homoplasy : it is a useful feature for a dissimilarity-matrix method to give such warnings of possible or probable homoplasy, which can then be checked more rigorously with a careful analysis of character states, e.g. with the help of program MacClade 3.05 (MADDISON & MADDISON, 1992).

Finally it must be noted that, clearly, Anâtaxis may give, as most distance-matrix methods do, phenetic branch lengths simply based on the level of divergence between any two lineages; but also, and more interestingly, it can offer for each branch of the evolutionary tree, in really phylogenetic terms, a qualitative estimation of the relative speed of divergence, which may then be compared with the phenetic branch length.

Dissimilarities to trichotomies correspondence table

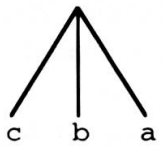
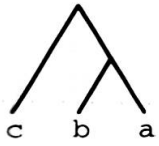
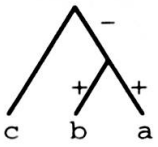
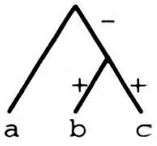
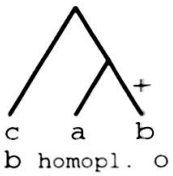
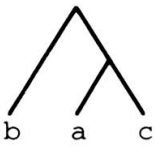
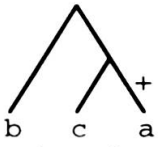
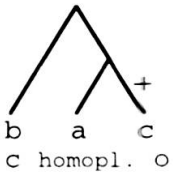
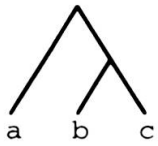
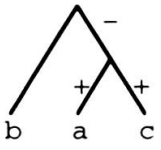
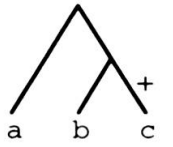
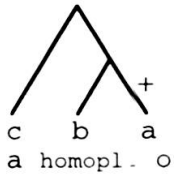
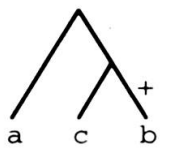
OUT-IN: $\hat{\Delta}_{oc} = \hat{\Delta}_{ob} = \hat{\Delta}_{oa}$ (o = outgroup)

$\hat{\Delta}_{cb}$ ---> $\hat{\Delta}_{ba}$ (direction of inequality)

$\hat{\Delta}_{ca}$

z homopl. o = z partially homoplasic with o
'+' for relatively rapid divergence
'-' for relatively slow divergence

So as to simplify the notation, the symbol Δ is not written in the following table

cb = ba = = ca 	*cb < ba* = = *ca*	*cb > ba* = = *ca*
cb = ba = >* *ca*	*cb < ba* = >* *ca*	cb > ba = >* ca 
cb = ba = <* *ca*	cb < ba = <* ca 	*cb > ba* = <* *ca*
cb = ba > =* *ca*	*cb < ba* > =* *ca*	cb > ba > =* ca 
cb = ba > >* *ca*	*cb < ba* > >* *ca*	cb > ba > >* ca 
cb = ba > <* ca 	cb < ba > <* ca 	cb > ba > <* ca 
cb = ba < =* *ca*	cb < ba < =* ca 	*cb > ba* < =* *ca*
cb = ba < >* ca 	cb < ba < >* ca 	cb > ba < >* ca 
cb = ba < <* *ca*	cb < ba < <* ca 	*cb > ba* < <* *ca*

REFERENCES

BERNARDI, G., MOUCHIROUD, D. & GAUTIER, C. 1993. Silent Substitutions in Mammalian Genomes and Their Evolutionary Implications. *J. of Molecular Evolution* 37, no. 06: 583.

BITTAR, G. 1995. "Phylogénie: histoire, concepts, principes et présentation de nouvelles méthodes d'analyse numérique de l'évolution moléculaire (Phylogeny: history, concepts, principles, and presentation of new methods for numerical analysis of molecular evolution)". Doctorate Thesis in Sciences, interdisciplinary type (Chemistry/Biology), n°2789, University of Geneva, 480 pp.

BITTAR, G. 1996. "Roots and xylem of phylogenetics". Société de Physique et d'Histoire naturelle, Genève, *Archives des Sciences* 49: 137-148 (Fasc. 2, Septembre).

BITTAR, G. & CARTER, L. 1994. "New probabilistic and numerical phenetics methods for inferring natural groupings and phylogenesis". In Proceedings of the International Meeting "Ecology and statistical methods", Niort, France, pp. 145-150.

BITTAR, G., CARTER, L., NADOT, S., SOUZA-CHIES, T., EVRARD, A., BESIN, E. & LEJEUNE, B. 1996. "A phylogenetic analysis of plants, using the chloroplast gene *rps*4 and the Anâtaxis method". Société de Physique et d'Histoire naturelle, Genève, *Archives des Sciences* 49: 149-157 (Fasc. 2, Septembre).

BITTAR, G. & VEUTHEY, A.-L. in preparation. "Phyletically relating Nematode, Insecta and Vertebrata, and positioning Protozoa within Eukaryota, using ribosomal proteins sequences".

BLAISDELL, B. E. 1991. "Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a variety of computer-generated model systems". *J. of Molecular Evolution* 32, 521-528.

BRITTEN, R. J. (1977). "The sources of variation in evolution". In *The encyclopedia of ignorance. Everything you ever wanted to know about the unknown* 1977 edit. (Duncan, R. & Weston-Smith, M., eds.), pp. 209-217. Pergamon Press, Oxford.

CACCONE, A. & POWELL, J. R. 1990. "Extreme rates and heterogeneity in insect DNA Evolution." *J. of Molecular Evolution* 30: 273-280.

CAVALLI-SFORZA, L. L. 1996. *Gènes, peuples et langues,* Odile Jacob, Paris

DOOLITTLE, R. F., ED. 1990. "Molecular evolution : computer analysis of protein and nucleic acid sequences". Vol. 183. Methods in Enzymology. Edited by Abelson, J. N. & Simon, M. I. San Diego (California), London: Academic Press

EASTEAL, S. 1990. "The pattern of mammalian evolution and the relative rate of molecular evolution." *Genetics* 124: 165-173.

FARRIS, J. S. 1972. "Estimating phylogenetic trees from distance matrices". *The American Naturalist* 106: 645-668.

FELSENSTEIN, J. 1993. "PHYLIP (Phylogeny Inference Package)". Version 3.57c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

FITCH, W. M. & MARGOLIASH, E. 1967. "Construction of phylogenetic trees". *Science* 155, 279-284.

GOJOBORI, T., MORIMAYA, E. N. & KIMURA, M. 1990. "Statistical methods for estimating sequence divergence." In *Molecular evolution : computer analysis of protein and nucleic acid sequences,* edited by Russell F. Doolittle, 531-550. San Diego (California), London: Academic Press.

GOJOBORI, T. & YOKOYAMA, S 1987. "Molecular evolutionary rates of oncogenes." *J. of Molecular Evolution* 26: 148-156.

GOODMAN, M. 1981. "Globin evolution was apparently very rapid in early vertebrates: a reasonable case against the rate-constancy hypothesis". *J. of Molecular Evolution* 17, 114-120.

HENDY, M. D. & PENNY, D. 1982. "Branch and bound algorithms to determine minimal evolutionary trees". *Mathematical Biosciences,* 59: 277-290.

HONEYCUTT, R. L., NEDBAL, M. A., ADKINS, R. M. & JANECEK, L. L. 1995. "Mammalian Mitochondrial DNA Evolution: A Comparison of the Cytochrome *b* and Cytochrome *c* Oxidase II Genes." *J. of Molecular Evolution* 40, no. 03: 260-272.

HOLMQUIST, R., CANTOR, CH. R. & JUKES TH. H. 1972. "Improved procedures for comparing homologous sequences in molecules of proteins and nucleic acids". *J. of Molecular Biology* 64: 145-161.

HUELSENBECK, J. P. & RANNALA, B. 1997. "Phylogenetic methods come of age: testing hypotheses in an evolutionary context." *Science* 276: 227-232.

JUKES, TH. H. & CANTOR, CH. R. 1969. "Evolution of protein molecules". In *Mammalian Protein Metabolism,* edited by H. N. Munro, 21-132. New York: Academic Press.

KLOTZ, L. C., KOMAR, N., BLANKEN, R. L. & MITCHELL, R. M. 1979. "Calculation of evolutionary trees from sequence data." *Proc. Natl. Acad. Sci.* USA 76: 4516-4520.

LI, W.-H. 1981. "Simple method for constructing phylogenetic trees from distance matrices." *Proc. Natl. Acad. Sci.* USA 78: 1085-1089.

MADDISON, W. P. & MADDISON, D. R. 1992. "MacClade: Analysis of phylogeny and character evolution." - Version 3.0.5 (1993) - Sunderland, Massachusetts: Sinauer Associates, Inc.

NADOT, S., BITTAR, G., CARTER, L., LACROIX, R. & LEJEUNE, B. 1995. "A phylogenetic analysis of monocotyledons based on the chloroplast gene *rps*4, using parsimony and a new numerical phenetics method". *Molecular Phylogenetics and Evolution* 4: 257-282.

NEI, M. 1991. "Relative efficiencies of different tree-making methods for molecular data". In *Phylogenetic analysis of DNA sequences* (Miyamoto, M. M. & Cracraft, J., eds.), pp. 90-128. Oxford University Press, Oxford.

OLSEN, G. J., MATSUDA, H., HAGSTROM, R. & OVERBEEK, R. 1994. "fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood." *Comput. Appl. Biosci.* 10: 41-48.

OSAWA, S. & JUKES, TH. H. 1995. "On codon reassignment". *J. of Molecular Evolution* 41(02), 247-249.

PALMER, J. D. 1992. "Mitochondrial DNA in plant systematics: applications and limitations". In *Molecular systematics of plants* (Soltis P. S., Soltis D. E., Doyle J. J., eds.), pp. 36-49. Chapman and Hall, New York.

PAWLOWSKI, J., SZADZIEWSKI, R., KMIECIAK, D., FAHRNI, J. & BITTAR, G. 1996. "Phylogeny of the infraorder Culicomorpha (Diptera: Nematocera) based on 28S RNA gene sequences". *Systematic Entomology* 21: 167-178.

ROSE, M. R. & DOOLITTLE, W. F. 1983. "Molecular biological mechanisms of speciation". *Science* 220, no. 8 April: 157-162.

SACCONE, C., LANAVE, C., PESOLE, G. & PREPARAT, A G. 1990. "Influence of base composition on quantitative estimates of gene evolution". In *Molecular evolution : computer analysis of protein and nucleic acid sequences* (Doolittle, R. F., ed.), Vol. 183, pp. 570-583. Academic Press, San Diego (California), London.

SAITOU, N. & NEI, M. 1987. "The neighbor-joining method: a new method for reconstructing phylogenetic trees". *Molecular Biology and Evolution* 4: 406-425.

SOUZA-CHIES, T., BITTAR, G., NADOT, S., CARTER, L., BESIN, E. & LEJEUNE, B. 1997. "Phylogenetic analysis of Iridaceae with parsimony and distances methods using the plastid gene rps4". *Plant Systematics and Evolution* 204: 109-123.

SWOFFORD, D. L. 1993. "PAUP : Phylogenetic Analysis Using Parsimony". Version 3.1.1, Computer programme distributed by the Illinois Natural History Survey, Champaign, Il.; Laboratory of molecular systematics, Smithsonian Institution, Washington D.C.

THOMAS, W. K. & BECKENBACH, A. T. 1989. "Variation in salmonid mitochondrial DNA: evolutionary constraints and mechanisms of substitution". *J. of Molecular Evolution* 29, 233-245.

THOMPSON, J. D., HIGGINS, D. G. & GIBSON, T. J. 1994. "Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". - *Nucleic Acids Research* 22: 4673-4680.

VEUTHEY, A.-L. & BITTAR, G., in preparation, "Phylogenetic relationship of Fungi, Plantae and Animalia, inferred from homologous comparison of ribosomal proteins".

WOLFE, K. H., LI, W.-H. & SHARP, P. M. 1987. "Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs". *Proc. Natl. Acad. Sci.* USA 84, 9054-9058.

WOLFE, K. H., MORDEN, C. W., EMS, S. C. & PALMER, J. D. 1992. "Rapid evolution of the plastid translational apparatus in a non-photosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes". *J. of Molecular Evolution* 35(4), 304-317.

ZUCKERKANDL, E. & PAULING, L. 1965. "Evolutionary divergence and convergence in proteins". In *Evolving genes and proteins,* edited by Bryson and Vogel, 97-166. New York: Academic Press.